Gaze4NLP 2025

Proceedings of the First International Workshop on Gaze Data and Natural Language Processing

associated with

The 15th International Conference on Recent Advances in Natural Language Processing RANLP'2025

Edited by Cengiz Acartürk, Jamal Nasir, Çağrı Çöltekin and Burcu Can Buğlalılar

12 September, 2025
Varna, Bulgaria

The First International Workshop on Gaze Data and Natural Language Processing
Associated with the International Conference
Recent Advances in Natural Language Processing
RANLP'2025

PROCEEDINGS

Varna, Bulgaria
12 September 2025

Online ISBN 978-954-452-104-2

Designed by INCOMA Ltd.
Shoumen, BULGARIA

Preface

It is our great pleasure to present the proceedings of the Gaze4NLP: The First International Workshop on Gaze Data and Natural Language Processing, held on 12th September, 2025 in Varna as part of RANLP (Recent Advances in Natural Language Processing) 2025. The workshop brought together researchers from diverse backgrounds to discuss recent advances and research methodologies in reading, eye tracking, and NLP.

This year, we received 16 submissions, each of which was reviewed by at least 2 members of our program committee. After a rigorous selection process, 9 papers were accepted for presentation and inclusion in these proceedings. The contributions span a wide range of topics, including eye tracking, mouse tracking, aligning language models with reading data, using language model surprisal in predicting reading times.

We would like to express our gratitude to the authors for their high-quality submissions and to the program committee for their contributions in the reviewing process.

We hope that the papers collected here will inspire further research, foster collaboration, and contribute to the advancement of the interdisciplinary area that combines NLP with eye tracking.

Cengiz Acartürk, Jagiellonian University, Poland
Burcu Can, University of Stirling, Scotland, UK
Çağrı Çöltekin, University of Tübingen, Germany
Jamal Nasir, University of Galway, Ireland

Organizing Committee and Volume Editors

• Cengiz Acarturk (Jagiellonian University, Poland)
• Jamal Nasir (University of Galway, Ireland)
• Burcu Can Buglalılar (University of Stirling, Scotland, UK)
• Cagrı Coltekin (University of Tubingen, Germany)

Program Committee

• Ozge Alacam, University of Bielefeld, Germany
• Le An Ha, Ho Chi Minh City University of Foreign Languages and Information Technology, Vietnam
• Fariz Ikhwantri, Simula Research Laboratory, Norway
• Oksana Ivchenko, University of Lille, France
• Pawel Kasprowski, Silesian University of Technology, Poland
• Victor Kuperman, McMaster University, Canada
• Joseph Lemley, University of Galway, UK
• David Reich, University of Potsdam, Germany
• Noam Siegelman, Hebrew University of Jerusalem, Haskins Laboratories, Israel
• Ana Matic Skoric, University of Zagreb, Croatia
• Evis Trandafili, University of New York Tirana, Albania
• Iraklis Varlamis, Harokopio University of Athens, Greece
• Mila Vulchanova, Norwegian University of Science and Technology, Norway

Table of Contents

Conference Program

What Determines Where Readers Fixate Next? Leveraging NLP to Investigate Human Cognition
Adrielli Tina Lopes Rego, Joshua Snell and Martijn Meeter

Benchmarking Language Model Surprisal for Eye-Tracking Predictions in Brazilian Portuguese
Diego Alves

EgoDrive: Egocentric Multimodal Driver Behavior Recognition Using Project Aria
Michael Rice, Lorenz Krause and Waqar Shahid Qureshi

Comparing Eye-gaze and Transformer Attention Mechanisms in Reading Tasks
Maria Mouratidi and Massimo Poesio

A French Eye-Tracking Corpus of Original and Simplified Medical, Clinical, and General Texts - FETA
Oksana Ivchenko and Natalia Grabar

Exploring Mouse Tracking for Reading on Romanian Data
Cristina Maria Popescu and Sergiu Nisioi

Where Patients Slow Down: Surprisal, Uncertainty, and Simplification in French Clinical Reading
Oksana Ivchenko, Alamgir Munir Qazi and Jamal Abdul Nasir

AlEYEgnment: Leveraging Eye-Tracking-While-Reading to Align Language Models with Human Preferences
Anna Bondar, David Robert Reich and Lena Ann Jäger

Predicting Total Reading Time Using Romanian Eye-Tracking Data
Anamaria Hodivoianu, Oleksandra Kuvshynova, Filip Popovici, Adrian Luca and Sergiu Nisioi

What determines where readers fixate next? Leveraging NLP to investigate human cognition

Adrielli Tina Lopes Rego
Department of Education Sciences
Vrije Universiteit Amsterdam
a.t.lopesrego@vu.nl

Joshua Snell
Department of Experimental
and Applied Psychology
Vrije Universiteit Amsterdam
j.j.snell@vu.nl

Martijn Meeter
Department of Education Sciences
Vrije Universiteit Amsterdam
m.meeter@vu.nl

Abstract

During reading, readers perform rapid forward and backward eye movements through text, called saccades. How these saccades are targeted in the text is not yet fully known, particularly regarding the role of higher-order linguistic processes in guiding eye-movement behaviour in naturalistic reading. Current models of eye movement simulation in reading either limit the role of high-order linguistic information or lack explainability and cognitive plausibility. In this study, we investigate the influence of linguistic information on saccade targeting, i.e. determining where to move our eyes next, by predicting which word is fixated next based on a limited processing window that resembles the amount of information humans readers can presumably process in parallel within the visual field at each fixation. Our preliminary results suggest that, while word length and frequency are important factors for determining the target of forward saccades, the contextualized meaning of the previous sequence, as well as whether the context word had been fixated before and the distance of the previous saccade, are important factors for predicting backward saccades.

1 Introduction

The eye movements of readers can reveal aspects of the cognitive mechanism that underlies language processing during reading. Decades of research have explored the explanatory power of eye movements to better understand which factors play a role in text comprehension (Rayner, 1998; Rayner et al., 2006). One well-established phenomenon in reading is the idea that lexical information, such as word length, frequency, and surprisal, influences the durations and locations of fixations in text (Kliegl et al., 2004). However, the influence of higher-level language processing on saccade targeting is less well known (Warren et al., 2011; Vasishth et al., 2013). Cognitive models of eye movements in reading vary in how saccade programming is simulated, and most models leave postlexical information implicit (e.g. Reichle et al., 2009). Furthermore, current machine learning approaches for predicting fixation location in reading are limited in shedding light on human language processing. They do little to explicate what drives saccade decisions and have few parallels to psycholinguistic theories and to behavioural evidence about the human cognitive systems engaged in reading.

Here we investigate to what extent we can successfully leverage deep learning methods to investigate a fundamental question about human language processing: what determines saccade programming during reading? We approach the prediction of the next fixation location as a classification problem at the word level, spanning a window of words that approximates the parallel processing of words in the human visual field (n-3 to n+3) (Snell and Grainger, 2019). In addition, we tailor the input of the model according to what information is likely to be available to the reader at each fixation. To represent the low-level linguistic information available for all words in the processing window, word length, frequency, and surprisal (i.e. negative log-probability) are employed. To represent higher-level linguistic information available on each previous word and the currently fixated word in the input sequence, we employ contextualized word embeddings from GPT-2, a unidirectional large language model (Rad-

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025,
pages 1–6, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-001

ford et al., 2019). Finally, previous fixation information is included to capture some of the dynamics of the sequential nature of eye movements. In sum, we attempt to combine the mapping power of neural networks with a more cognitively plausible set-up to understand what determines the next fixation target in human reading.

2 Related Work

The task of predicting fixation locations in reading has been mainly addressed with one of the two modeling strategies: theory-driven or data-driven. Theory-driven models are cognitive models that simulate eye movements in reading by computationally implementing psycholinguistic theories of reading with the goal of revealing the cognitive mechanisms involved in reading. The next fixation location is explicitly determined, i.e., it is clear how the model arrives at each saccadic decision. However, they are hardly ever evaluated on unseen texts and readers, and are limited in explaining the role of high-order linguistic information in saccade targeting. In E-Z reader (Reichle et al., 2009), for example, saccade targeting is limited to the range of word n-1 to word n+2, and is mainly determined by word length, frequency and predictability. Regressions occur randomly with a certain probability set by the modeler. Perhaps a more elegant mechanism is proposed by SWIFT (Engbert et al., 2005), in which the probability of each word in the model's four-word processing window to be the next saccade target is proportional to its relative word activation in an attention gradient. However, SWIFT is limited in that higher-level language processing is not accounted for. SEAM (Rabe et al., 2024) partially addresses this limitation by having sentence-level dependencies indirectly affect word activations, but this effect only occurs between verbs and subjects.

In contrast, data-driven models of eye movement simulations solely focus on accurately predicting eye movements by harnessing advanced machine learning methods while using previous (and future) fixations and/or linguistic information as input. These models rely on the predictive power of machine learning methods to achieve accurate prediction of fixations on a variety of texts, with different reading goals and reader profiles. They do this without guidance of theories of reading and little to no parallel with human cognition with respect to the model input and/or architecture. The first successful data-driven model (Nilsson and Nivre, 2009) employed logistic regression with manually engineered features extracted from the text stimuli and the previous eye movements of readers to predict the next saccade target location within a five-word window around the currently fixated word; feature importance was not reported. Wang et al. (Wang et al., 2019) combined CNN, LSTM and CRFs to predict next fixation location, based on word length, part-of-speech, and bag-of-words representations, but no regressions nor refixations were produced by the model. Dweng et al. (Deng et al., 2023) proposed Eyettention, which combines the fixation sequence (represented by non-contextualized BERT embeddings, fixation duration and landing position) and the word sequence (represented by contextualized BERT embeddings and word length) using two (bi-)LSTMs and a cross-attention layer. This model was surpassed in performance by ScanDL (Bolliger et al., 2023), a sequence-to-sequence diffusion model that generates synthetic scanpaths by also combining the fixation sequence and the word sequence (both represented by BERT embeddings). While data-driven models have been so far more successful than theory-driven ones in accurately predicting fixation locations, they still lack explanatory power and cognitive plausibility to be useful models to investigate human cognition in reading: Much of the information driving prediction is left implicit (e.g. predicting upcoming fixations based on previous fixations does not explain what underlies saccade targeting), and most information used in the input is not plausibly available to a human reader at each fixation step (e.g. future fixations, and many/all upcoming words).

3 Method

We formulated saccade targeting in reading as a classification problem, where the model has to decide which word to fixate next given a set of candidate words in the input sequence. The classifier was a shallow fully connected neural network, with one hidden layer of 128 nodes, ReLu activation and a drop-out layer [1]. The input sequence consisted of a window of seven words, i.e. the fixated word plus three words before and three words after, to approximate the limited amount of information a human reader can likely take in the visual field at each

[1] Pilot studies were performed with CNNs and LSTMs to preserve the word structure in the input, but, surprisingly, the fully-connected neural network yielded the best results.

fixation. To represent lexical information on each word in the input sequence, we used word length, frequency, and surprisal, which are assumed to be available to the reader to some degree through either past word recognition or current parafoveal processing. To represent higher-order language information, we used the contextualized word embedding of the fixated word from GTP-2, which is assumed to encode the meaning constructed from the text up to the fixated word. Finally, to capture some of the dynamics inherent to the sequential nature of eye movements, we added information on whether each word in the input sequence has been fixated before, the previous fixation duration and the previous saccade length. All features were z-normalized, except for the word embedding and the binary feature encoding whether or not the word had been fixated before.

We trained the classifier on the L1-English part of the MECO corpus (Siegelman et al., 2022), using 5-fold cross-validation with a 80/20 split based on text ids. The material consists of the first 10 texts of the corpus, structured similarly to Wikipedia-style encyclopaedic entries, covering a diverse range of topics. Each text had approximately 200 words and 10 sentences. All participants (n = 46) were native speakers of English and university students. They were instructed to read the texts silently and answer (four) comprehension questions after each text. We used the fixation dataset available in the "fixation report" folder, in the path "release 1.0/version 1.2/primary data/eye tracking data/fixation report", in the OSF directory of the MECO corpus. We only included the fixations on words that had three words to the left and three words to the right, resulting in 66,383 fixations in total. Around 34% of these fixations were to word n+1, followed by 25% to word n+2, 18% to word n, 10% to word n-1, 7% to word n+3, 3% to word n-2, and 1% to word n-3.

Model evaluation consisted of measuring the F1 scores ($2 * (precision * recall)/precision + recall$) for each word position in the input sequence (seven words, including currently fixated word) and the macro-averaged F1 score across word positions. We compare the model performance with three baselines: OB1-reader (Snell et al., 2018), a cognitive model of eye movement control in reading, in which saccade targeting is determined by word recognition and visual attention; the same model trained on random input vectors; and a majority baseline, which always predicts the majority class (word n+1). To evaluate OB1-reader, we ran 10 simulations on the corpus texts and, for each simulation, we selected the fixations that overlapped between the model simulation and the corpus, and checked whether the next fixation target was the same. We then reported the resulting F1 score averaged over simulations.

4 Results

As can be seen in Table 1, our model outperforms the baselines, including the OB1-reader model, although the difference in macro-averages is small. The easiest saccade to predict is to word n+1, which is also the most frequent. Backward saccades are the most difficult to predict, and the farther away from the current fixation, the lower the performance in predicting saccade targeting. OB1-reader performs remarkably well compared to our model, especially at one-word regressions and refixations. Overall, our model improves saccade targeting prediction compared to the baselines, but still performs below chance for word skips and refixations, and poorly for backward saccade targeting.

To determine feature importance, we replaced one feature at a time by its average over the dataset and retrained the model with the ablated feature. Table 2 shows the model performance when removing each feature. When word length is ablated, the model performance especially drops in predicting word skips (word positions 2 and 3). Word frequency also seemed to affect two word-skipping (word position 3). Whether or not the context word has been fixated before is predictive of backward saccades (word positions -1, -2, and -3), as well as refixations and two-word skipping (word positions 0 and 3). Embeddings seems to be informative for backward saccades, but not for word skipping (word positions 2 and 3). Finally, while the previous fixation duration does not seem to be an informative feature in general, the previous saccade distance supports to some extent the prediction of backward saccades (word positions -3 and -2) as well as two-word skipping (word position 3). In sum, word length and frequency were important features for the prediction of forward saccades, while the fixated word's contextualized embedding, whether the word has been fixated before and the previous saccade length were mainly informative of backward sacacades.

	-3	-2	-1	0	1	2	3	macro-avg
Classifier	.002 ±.005	.001 ±.002	.05 ±.017	.24 ±.018	.56 ±.024	.46 ±.018	.12 ±.038	.20 ±.006
OB1-reader	0	0	.11 ± .002	.30 ± .01	.31 ± .01	.32 ± .004	.15 ± .006	.17 ± .002*
Random	0	0	.01 ±.008	.11 ±.006	.44 ±.016	.35 ±.011	.004 ±.006	.11 ±.004 *
Majority	0	0	0	0	.51 ±.017	0	0	.07 ±.002 *

Table 1: F1 scores averaged over cross-validation splits for each true word position target, as well as averaged over positions. * means that the score was significantly different from the classifier model.

	-3	-2	-1	0	1	2	3	macro-avg
Classifier	.002 ±.005	.001 ±.002	.05 ±.017	.24 ±.018	.56 ±.024	.46 ±.018	.12 ±.038	.20 ±.006
w/o word length	.002 ±.005	.001 ±.002	.04 ±.01	.23 ±.03	.54 ±.02	.42 ±.02	.07 ±.02	.19 ±.008 *
w/o word frequency	0	.003 ±.008	.05 ±.02	.25 ±.008	.55 ±.02	.45 ±.01	.07 ±.02	.19 ±.003 *
w/o word surprisal	0	.002 ±.003	.04 ± .01	.24 ± .01	.56 ± .02	.46 ± .01	.10 ± .03	.20 ± .003
w/o has-been-fixated	0	0	.01 ± .01	.21 ± .02	.55 ± .02	.45 ± .01	.06 ± .06	.18 ± .01*
w/o embedding	0	0	.02 ± .01	.24 ± .02	.59 ± .02	.50 ± .01	.17 ± .06	.22 ± .01
w/o previous fixation duration	.004 ± .006	.001 ± .002	.06 ± .02	.25 ± .02	.56 ± .03	.46 ± .01	.10 ± .03	.20 ± .005
w/o previous saccade distance	0	0	.04 ± .01	.26 ± .02	.56 ± .02	.46 ± .01	.09 ± .04	.20 ± .004

Table 2: Feature ablation. This table displays the F1 scores averaged over cross-validation splits for each true word position target, as well as averaged over positions, for each model version in which one feature is ablated. * means that the score was significantly different from the full classifier model.

5 Discussion

In this study, we attempted to investigate the cognitive processes underlying saccade targeting in reading using deep learning. We sought to leverage machine learning while using input whose information content may resemble more closely what is plausibly available to human readers during saccade planning. Importantly, we attempted to fill a gap in understanding the role of high-order language information by investigating to what extent the text meaning, as represented by contextualized embeddings, supports where readers tend to fixate next, beyond lower-level lexical information. Our preliminary results indicated that forward saccades tend to be more driven by automatic, oculomotor cues, as well as low-level linguistic cues, such as word length and frequency, whereas backward saccades are more heterogeneous, with the semantics of the previous context playing a role, but also factors possibly related to oculomotor error, such as skipping a word due to overshooting, as suggested by the features "has-been-fixated" and "previous saccade amplitude". Our results are in line with well-established findings in the literature that support the major role of lower-order linguistic features in forward saccades (Rayner, 1998; Kliegl et al., 2004; Engbert et al., 2005) and the heterogeneous nature of backward saccades (Von Der Malsburg and Vasishth, 2011; Inhoff et al., 2019; Wilcox et al., 2024). Furthermore, refixations seemed to be driven by word length and whether the word had been fixated before, but, surprisingly, not by factors pertaining word meaning, such as frequency, surprisal and its contextualized embedding, suggesting that, at least in this dataset, most refixations were a result of oculomotor and low-level linguistic cues. Ultimately, our goal is to model the complex interplay between the oculomotor system and language processing that drives saccade targeting in reading. Combining the predictive power of machine learning methods with more cognitively plausible and interpretable modeling may shed light on the mechanisms behind this process.

6 Limitations and Future Work

The model proposed here fails to predict backward saccades with an acceptable level of accuracy. Previous correlational research has suggested PMI scores to be predictors of regression targeting in reading (Wilcox et al., 2024). A follow-up study may explore the potential of such measure in informing the prediction of backward saccade targeting in reading. In addition, the dynamics of eye movements is not fully explored in our model,

as only information on the previous fixation is used. It is possible that information on more previous fixations is needed to capture the complex relation between the sequence of eye movements and the sequence of language input.

Finally, we assumed that word length, frequency and surprisal of the words in the upcoming context are fully available to the reader, which is a simplification. As a follow-up, this information will be modulated by OB1-reader's visual attention gradient, based on eccentricity and visual acuity. That is, the closer the words are to the fixation the more accurate the linguistic information available. Future work may investigate whether our neural network model can be merged with a cognitive model, such as OB1-reader, to use word activations generated by the cognitive model as a proxy of low-order visual and linguistic information, together with high-order linguistic information represented by contextualized embeddings, to predict saccade targeting. More of the dynamics of the relation between eye movements and language input might be indirectly captured by the cognitive model's word activations.

7 Acknowledgments

This research was funded by the "Nederlandse Organisatie voor Wetenschappelijk Onderzoek" (NWO) 672 Open Competition-SSH (Social Sciences and Humanities), grant number 406.21.GO.019 to MM. A big thanks to the MultiplEYE Cost Action (CA21131) for funding travel expenses of the first author to attend the Workshop.

References

Lena S Bolliger, David R Reich, Patrick Haller, Deborah N Jakobi, Paul Prasse, and Lena A Jäger. 2023. Scandl: A diffusion model for generating synthetic scanpaths on texts. *arXiv preprint arXiv:2310.15587*.

Shuwen Deng, David R Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena A Jäger. 2023. Eyettention: An attention-based dual-sequence model for predicting human scanpaths during reading. *Proceedings of the ACM on Human-Computer Interaction*, 7(ETRA):1–24.

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

Albrecht W Inhoff, Andrew Kim, and Ralph Radach. 2019. Regressions during reading. *Vision*, 3(3):35.

Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European journal of cognitive psychology*, 16(1-2):262–284.

Mattias Nilsson and Joakim Nivre. 2009. Learning where to look: Modeling eye movements in reading. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 93–101.

Maximilian M Rabe, Dario Paape, Daniela Mertzen, Shravan Vasishth, and Ralf Engbert. 2024. Seam: An integrated activation-coupled model of sentence processing and eye movements in reading. *Journal of Memory and Language*, 135:104496.

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

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

Keith Rayner, Kathryn H Chace, Timothy J Slattery, and Jane Ashby. 2006. Eye movements as reflections of comprehension processes in reading. *Scientific studies of reading*, 10(3):241–255.

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

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

Joshua Snell and Jonathan Grainger. 2019. Readers are parallel processors. *Trends in Cognitive Sciences*, 23(7):537–546.

Joshua Snell, Sam van Leipsig, Jonathan Grainger, and Martijn Meeter. 2018. Ob1-reader: A model of word recognition and eye movements in text reading. *Psychological review*, 125(6):969.

Shravan Vasishth, Titus von der Malsburg, and Felix Engelmann. 2013. What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(2):125–134.

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.

Xiaoming Wang, Xinbo Zhao, and Jinchang Ren. 2019. A new type of eye movement model based on recurrent neural networks for simulating the gaze behavior of human reading. *Complexity*, 2019(1):8641074.

Tessa Warren, Erik D Reichle, and Nikole D Patson. 2011. Lexical and post-lexical complexity effects on eye movements in reading. *Journal of Eye Movement Research*, 4(1):1.

Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2024. An information-theoretic analysis of targeted regressions during reading. *Cognition*, 249:105765.

Benchmarking Language Model Surprisal for Eye-Tracking Predictions in Brazilian Portuguese

Diego Alves
Saarland University
Saarbrücken, Germany
diego.alves@uni-saarland.de

Abstract

This study evaluates the effectiveness of surprisal estimates from six publicly available large language models (LLMs) in predicting reading times in Brazilian Portuguese (BP), using eye-tracking data from the RastrOS corpus. We analyze three key reading time measures: first fixation duration, gaze duration, and total fixation time. Our results demonstrate that surprisal significantly predicts all three measures, with a consistently linear effect observed across all models and the strongest effect for total fixation duration. We also find that larger model size does not necessarily provide better surprisal estimates. Additionally, entropy reduction derived from Cloze norms adds minimal predictive value beyond surprisal, and only for first fixation duration. These findings replicate known surprisal effects in BP and provide novel insights into how different models and linguistic predictors influence reading time predictions.

1 Introduction

In recent years, the use of large language models (LLMs) has emerged as a productive approach in cognitive science and psycholinguistics to better understand human language processing (Hale (2001); Armeni et al. (2017); Wilcox et al. (2020)). These models provide estimates of word predictability, which can be formalised through information-theoretic measures as surprisal.

Surprisal quantifies the unexpectedness of a word given its preceding context. It has been shown that this measure correlates with reading time metrics obtained through eye-tracking experiments (Smith and Levy (2013); Hofmann et al. (2022); Demberg and Keller (2008)). These findings support theories claiming that human language comprehension is governed, at least to some extent, by the efficient processing of probabilistic information.

Despite extensive research on surprisal linked to cognitive processing, the focus has largely been on English, leaving cross-linguistic applicability underexplored. Wilcox et al. (2023a) showed how well language model surprisal can predict reading times in eleven languages, providing important information regarding cross-linguistic variability in the cognitive processing of language. However, Brazilian Portuguese (BP) was notably absent from this analysis, leaving a gap in our understanding of the role of surprisal in the processing of this language.

To address this gap, the present study focuses on BP, employing the RastrOS corpus, a large-scale eye-tracking dataset collected from students in higher education in Brazil, which also includes carefully constructed norms of predictability of words (Leal et al., 2022).

The aim of this study is to evaluate how surprisal values derived from a variety of publicly available LLMs predict three key eye-tracking reading time measures: first fixation duration, gaze duration, and total fixation time in Brazilian Portuguese. Furthermore, we investigate the role of entropy reduction as a contributing factor in modelling reading times. We also assess the linearity of the relationship between surprisal and reading times, determining whether linear models sufficiently capture this mapping or whether more complex patterns are present.

Our work not only provides missing data for Brazilian Portuguese but also identifies the most effective publicly available LLMs for modelling human reading behaviour in this language. Moreover, it offers a baseline for researchers aiming to use surprisal to analyse linguistic phenomena in Brazilian Portuguese following information-theoretic principles such as the Uniform Information Density (UID) hypothesis (Jaeger and Levy, 2006).

The remainder of the paper is organised as fol-

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025,
pages 7–17, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-002

lows. Section 2 reviews related work on using LLMs to model reading times. Section 3 presents the dataset, describes the large language models tested, and explains our evaluation methods. Section 4 then presents the results. We conclude with a summary of our findings and directions for future work in Section 5, followed by a discussion of the study's limitations in Section 6.

2 Related Work

Wilcox et al. (2023a) examined surprisal's relationship to reading times in eleven languages across five language families. Using monolingual and multilingual transformer-based language models (trained on the Wiki40B dataset, Guo et al. (2020), and mGPT, Shliazhko et al. (2024)), they showed that both surprisal and contextual entropy predict reading times, and that the relationship between surprisal and reading time is linear.

This linear relationship was also supported by Xu et al. (2023), who analysed seven languages and found evidence of superlinear effects in some cases, with results highly dependent on the language model used to estimate surprisal.

Additionally, Wilcox et al. (2023b) tested the quality–power (QP) hypothesis, which posits that higher-quality language models (LMs) better predict human reading behaviour. By training LMs on 13 languages with varying amounts of training data (from 1 million to 1 billion tokens), they found that, in most cases, models trained on more tokens showed stronger predictive power for eye-tracking data, supporting the QP hypothesis within the tested range.

Lin and Schuler (2025) proposed a neural study to complement these observations regarding reading time. By evaluating surprisal estimates from 17 Transformer models across three language families using fMRI data, they showed that the positive relationship between model perplexity and predictive power also generalizes to neural measures.

However, regarding LLMs, Oh and Schuler (2023) demonstrated that despite having better perplexity, larger models predict human reading times less accurately. Specifically, they tend to underpredict reading times for named entities and overpredict for function words, suggesting that memorization in these models reduces their alignment with human processing.

This tendency is also observed by Liu et al. (2023) who examined the effect of temperature scaling on large language model surprisal estimates and their fit to English reading time data, showing that calibration improves with model size, and temperature scaling significantly enhances prediction.

Moreover, Nair and Resnik (2023) demonstrated that while surprisal theory explains how a word's predictability influences processing difficulty via probabilistic updating, it does not fully capture all aspects of incremental processing, such as effects from low-frequency words and garden-path disambiguation. To address these limitations, Wang et al. (2025) developed a model that integrates syntactic information with statistical surprisal estimated from LLMs, resulting in significantly higher correlations with human reading times than surprisal alone.

Therefore, although surprisal alone cannot fully account for cognitive language processing, it has a significant impact across many languages. Additionally, both the size of the language model and the amount and quality of training data affect the relationship between reading time and word predictability. Consequently, it is crucial to identify the best language model for each language (and language variety) before applying surprisal estimates in various research fields.

3 Methodology

3.1 Eye-Tracking Data

The RastrOS corpus was developed to support psycholinguistic research on Brazilian Portuguese (BP), particularly focusing on lexical predictability and sentence processing. It comprises two main components: predictability norms collected via a Cloze test and eye-tracking data gathered from reading tasks.

A total of 393 native BP speakers from six Brazilian universities participated in the Cloze test, primarily undergraduate students. Each participant completed Cloze tasks on five randomly selected paragraphs, balanced across three genres: journalistic (40%), literary (20%), and popular science (40%).

The Cloze corpus includes 50 paragraphs, comprising 120 sentences and 2,494 words (2,831 tokens). Source texts were drawn from the Lácio-Web corpus (Aluísio et al., 2004), public domain literature, and contemporary online texts.

Participant responses were compared to target words based on orthographic match, morphosyntactic class (PoS), and inflection, with semantic simi-

larity assessed via word embeddings. The dataset is annotated with PoS tags (using the Palavras parser; Bick 2000), word frequency (from Corpus Brasileiro (Sardinha, 2010) and BrWaC (Wagner Filho et al., 2018)), and includes surprisal and entropy reduction values derived from the Cloze test results.

The eye-tracking data of the RastrOS were collected from 37 undergraduate students and were recorded using the EyeLink 1000 eye-tracker at a sampling rate of 1000 Hz.

Participants read 120 sentences taken from the same 50-paragraph Cloze corpus, a total of 2,494 words total (2,831 tokens including punctuation). Each sentence is annotated with 36 eye-tracking metrics (e.g., first fixation duration, gaze duration, and total fixation time).

3.2 Large Language Models

For our analysis, we selected six publicly available large language models that vary in the number of parameters and the training data used:

1. Bloom-560m[1] (Workshop, 2022) - Multilingual model trained on 1.5 TB of pre-processed text, of which 11.1% is Portuguese. 559 million parameters distributed over 24 layers with 16 attention heads and 1024-dimensional hidden states.

2. Bloomz-7b1[2] (Muennighoff et al., 2022) - Same training corpus as bloom-560m but with 7 billion parameters over 30 layers with 32 attention heads, and 4096-dimensional hidden states. bloomz is a fine-tuned version of bloom, trained with multitask instructions to improve zero-shot performance.

3. Llama-2-7B-hf[3] – Pretrained on 2 trillion tokens from public sources, then fine-tuned with public instruction datasets and over one million human-annotated examples. It has 1024 hidden dimensions with 32 attention heads over 32 layers (Wang et al., 2023). Evaluation tests were performed only in English.

4. Llama-3-2-1B[4] – 1 billion parameter model, pretrained on up to 9 trillion tokens of data in 8 different languages (including Portuguese) from publicly available sources.

5. Llama-3-2-3B[5] – Same training data as llama-3-2-1B but with 3 billion parameters,

6. Mistral-7b[6] – Trained on a mix of web data and code, with 7 billion parameters. It has 32 layers, 32 attention heads, and a hidden size of 4096 dimensions. The model evaluation was conducted exclusively on English (Jiang et al., 2023).

With this selection, our aim is to provide a meaningful comparison between language models of different sizes and training objectives, including models fine-tuned for specific tasks (e.g., bloomz-7b1), and models primarily focused on English (e.g., llama-2-7B-hf and mistral-7B), despite being trained on multilingual data. Unfortunately, only the BLOOM models provide sufficient information about the proportion of Portuguese in their training data, although they do not specify which variety of Portuguese was used.

3.3 Evaluation Methods

To evaluate the performance of the different language models in predicting reading times, we adopted the methodology proposed by Wilcox et al. (2023a).

Thus, although the RastrOS corpus provides 36 different word-based measures of reading time, we focus on three commonly used metrics (Rayner, 1998):

1. First fixation duration - the duration of the first fixation on a word during its first pass. Annotated as IA_FIRST_FIXATION_DURATION in RastrOS.

2. Gaze duration - the sum of all first-pass fixations on a word. IA_FIRST_RUN_DWELL_TIME in RastrOS.

3. Total fixation duration - the sum of all fixations on a word during the trial. IA_DWELL_TIME in RastrOS.

First fixation reflects the initial processing of a word and is associated with early stages of lexical

[1] https://huggingface.co/bigscience/bloom-560m

[2] bigscience/bloomz-7b1

[3] https://huggingface.co/meta-llama/Llama-2-7b-hf

[4] https://huggingface.co/meta-llama/Llama-3.2-1B

[5] https://huggingface.co/meta-llama/Llama-3.2-3B

[6] https://huggingface.co/mistralai/Mistral-7B-v0.1

access. Gaze duration captures the time spent on a word during first-pass reading and is sensitive to lexical and syntactic processing. Total fixation time includes any regressions back to the word and reflects later stages of comprehension, such as reanalysis or integration difficulties (Rayner, 1998).

3.3.1 Surprisal

To compute word-level surprisal values, we used the surprisal Python library[7]. Sentences from RastrOS were first recomposed and loaded in their original order. Using the AutoHuggingFaceModel interface provided by the library, we instantiated each selected model and computed token-level surprisal values for each recomposed sentence. As a post-processing step, we merged the subword tokens produced by the language models to reconstruct the original words for analysis.

The regression models follow the structure proposed by Wilcox et al. (2023a), aiming to predict the reading time $y(w_t, w_{<t})$ of a word w_t given its preceding context $w_{<t}$. The predictor vector x_t includes not only information about the target word w_t, but also features from the two preceding words w_{t-1} and w_{t-2}, in order to account for potential spillover effects on reading time.

Each model includes baseline predictors such as word length and log unigram frequency (corresponding to Word_Length and Freq_brWaC_log in RastrOS) for the target word and the two preceding words. These features form the baseline structure of the predictor vector x_t at position t.

We used linear mixed-effects regression models, implemented via the lmer() function from the lme4 R package (Bates et al., 2015).

To measure how much surprisal improves model performance, we compare the baseline model (Appendix A, equation 1) to models that include surprisal values, specifically the surprisal of the target word and its two preceding words as estimated by the LLM (Appendix A, equation 2). The delta is defined as the difference in per-word log-likelihood between the surprisal-enhanced model and the baseline: a positive delta indicates that surprisal helps the model better explain that word's reading time. By aggregating these deltas across all words, we assess whether incorporating surprisal significantly improves prediction accuracy.

Additionally, all regression models are trained and evaluated using 10-fold cross-validation. To assess the significance of the observed differences (Δ) between target and baseline models, we use a paired permutation test. This non-parametric test evaluates whether Δ significantly differs from zero and whether different models differ from each other, without assuming any specific distribution of the test statistic. p-values are computed based on the empirical distribution of likelihood differences, estimated by averaging over permutations of the likelihood values.

For each reading time measure, we compared the Δ values obtained using surprisal estimates from the LLMs listed in Subsection 3.2.

3.3.2 Entropy Reduction

Wilcox et al. (2023a) tested the influence of contextual entropy as a predictor, comparing it to a baseline model that included the features from the baseline structure combined with surprisal values.

Rather than using contextual entropy, we employed entropy reduction values from the RastrOS corpus (Entropy_Reduction), derived from Cloze test results. Lowder et al. (2018) demonstrated that entropy reduction significantly predicts reading time. This is limitation of this approach when compared to entropy estimates generated by a language model trained on a large corpus. Nevertheless, we decided to use the available entropy reduction values provided by the corpus to have at least an estimation of the effect.

Thus, using baselines that include surprisal values, we compared models for each reading measure and LLM by adding entropy reduction as a predictor, considering the target token and the two preceding tokens for both surprisal and entropy reduction, with Δ and the statistical tests as described in 3.3.1. The model including entropy reduction is described in Appendix B.

3.3.3 Linearity

For the analysis of surprisal and entropy reduction as predictors, we used regression models that assume a linear relationship between surprisal and reading time as supported by previous studies (e.g., Smith and Levy (2013); Wilcox et al. (2020); and Shain et al. (2024)). However, as recent work has challenged this assumption, proposing superlinear (e.g., Meister et al. (2021) and Hoover et al. (2022)) or sublinear (Hoover et al., 2023) links, we decided, following Wilcox et al. (2023a) to test this by comparing the performance Δ of our linear models with models capable of capturing non-linear rela-

[7] https://pypi.org/project/surprisal/

tionships.

To analyse linearity, we used generalized additive models (GAMs), which flexibly capture potential non-linear effects.

If the GAM fits a visually linear pattern, this supports the hypothesis of a linear link. We modelled reading times based on Freq_brWaC_log, Word_Length, and surprisal from sentence-level. Our GAMs included smooth terms for current and previous word surprisal and tensor product terms to model non-linear interactions between log-frequency and word length, following the method applied by Wilcox et al. (2023a).

Thus, we compared generalized additive models (GAMs) that model surprisal effects on reading time either as linear terms or as flexible non-linear smooth functions, alongside a baseline model without surprisal, as described in Appendix C. Using 10-fold cross-validation, we calculated the prediction error (RMSE) for each model on held-out data and computed the improvement Δ over the baseline (without surprisal) for both linear and non-linear surprisal models. We then tested the significance of these improvements and differences between linear and non-linear models using paired permutation tests.

4 Results

4.1 Suprisal Models Compared to Baseline

Figure 1 presents the mean Δ log-likelihood per word for each LLM across all three reading time measures, shown as separate panels.

Regarding the different reading time measures, surprisal shows the highest predictive power for total fixation time duration, followed by gaze duration, and finally the lowest Δ values for first fixation duration. These results align with those reported by Wilcox et al. (2023a), showing similar magnitudes of Δ across reading measure conditions.

The analysis of the models concerning first fixation duration shows that the Δ values are approximately 0.0025. All Δ values are significantly different from 0 ($p < 0.001$). The pairwise comparison of the different models shows that there are no statistically significant differences among them.

Regarding the gaze duration, there is a higher variation in Δ values, with larger standard error bars. Statistical tests show that, for all models, Δ differs significantly from 0, except for bloomz-7b1 ($p = 0.0014$). When comparing the different language models, the statistical tests indicate that the models have significantly different Δ values, except for:

- llama-2-7b similar to bloom-560, llama-3-2-1B, and llama-3-2-3B
- llama-3-2-3B similar to llama-3-2-1B
- mistral_7B similar to llama-3-2-3B and llama-2-7b

Finally, when considering total fixation time, except for bloomz-7b1, Δ values are around 0.05 and are all significantly different from 0. Also, all models differ in the pairwise comparison, except for:

- mistral_7B which is similar to llama-2-7b, llama-3-2-1B, and llama-3-2-3B
- llama-2-7b, similar to llama-3-2-1B

These results show that the best models are not necessarily those with the highest number of parameters, as for gaze duration, statistically similar results were obtained for models with 560 million, 1, 3, and 7 billion parameters. This effect is even more pronounced when considering total fixation time, with some statistically similar results observed for models with 1, 3, and 7 billion parameters.

The overall analysis of Figure 1 indicates that the best model—considering both gaze and total fixation durations—is llama-3-2-3B. Moreover, it is notable that the fine-tuned model bloomz-7b1, despite having 7 billion parameters, performs the worst in predicting reading times. Additionally, although not evaluated in languages other than English, mistral_7B shows statistically similar Δ values when compared to llama-3-2-3B.

The estimated effects of surprisal, including coefficients and standard errors for each model, are presented in Table 1.

We observe some consistency among the models with the best Δ values. The most discrepant model is bloom-7B1, reflecting its poor performance. Other predictors also show significant effects, except for the log frequency of the second word preceding the target, which was statistically significant only for bloom-7B1.

4.2 Entropy Reduction Models Compared to Surprisal Baseline

The Δ results comparing baseline models (with surprisal) to models that include both entropy reduction and surprisal are presented in Figure 2.

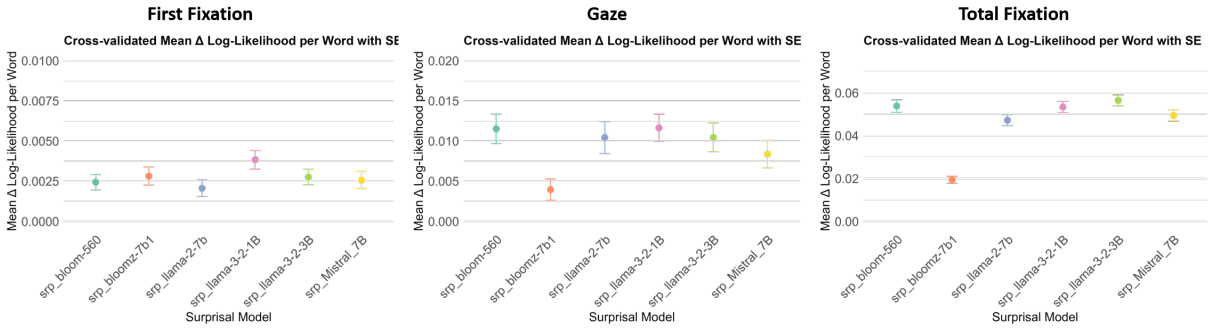

Figure 1: Predictive power of surprisal across reading time measures and LLMs. Dots indicate mean Δ log-likelihood per word; error bars show ±1 standard error of the mean. Note that each panel uses a different y-axis scale.

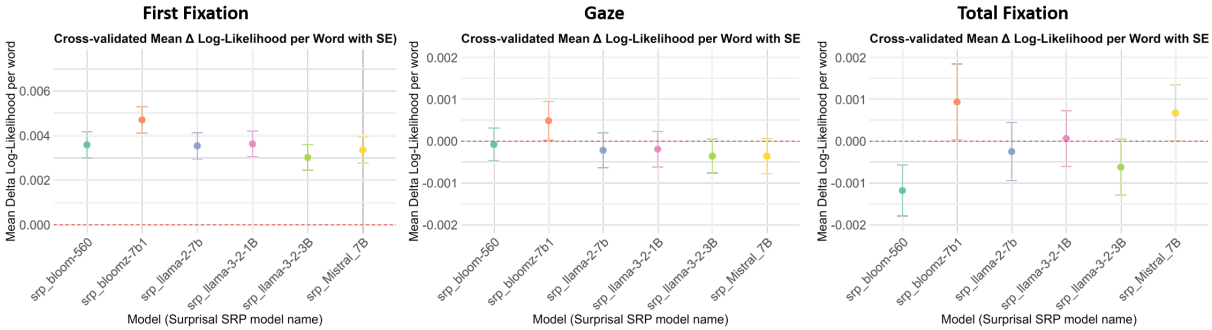

Figure 2: Predictive power of entropy reduction across reading time measures and LLMs. Dots indicate mean Δ log-likelihood per word; error bars show ±1 standard error of the mean. Note that each panel uses a different y-axis scale.

Model	Effect	Std. Error
bloom-560	9.52	0.15
bloom-7B1	5.01	0.13
llama-2-7B	8.16	0.14
llama-3-2-1B	8.57	0.15
llama-3-2-3B	8.61	0.15
mistral-7B	8.37	0.15

Table 1: Surprisal coefficient for the target word in a linear model including surprisal, frequency, and word length as predictors (considering target word and the two previous ones).

The statistical analyses show that, for first fixation duration, all Δ values are significantly different from 0, although the models perform similarly. In contrast, for both gaze duration and total fixation duration, all Δ values are close to 0, and no significant differences between models were observed.

Wilcox et al. (2023a) observed an improvement in the prediction of gaze duration when adding contextual entropy as a predictor, with a weak—albeit consistent—effect across languages. In our study, however, we do not observe the same effect. Including entropy reduction appears to have a positive impact (independent of the language model) only for first fixation duration, and even then, the Δ values are low (around 0.004).

4.3 Linearity analysis

Figure 3 presents the results of comparing the Δ obtained from a linear GAM model with surprisal to a baseline model without surprisal, as well as the corresponding Δ values from a non-linear model, for the prediction of total fixation duration.

As expected from the results presented in Section 3.3.1, the statistical tests showed that all Δ are significantly different from 0. Moreover, when comparing the linear Δ with the non-linear one for each language model, we observe that the results are not significantly different.

Thus, these results corroborate the claim that the effect of surprisal on reading time is linear, consistent with the findings of Wilcox et al. (2020). This linear effect is observed across all LLMs considered, regardless of parameter size, training data, or supported languages.

Figure 4 shows the differences in surprisal ef-

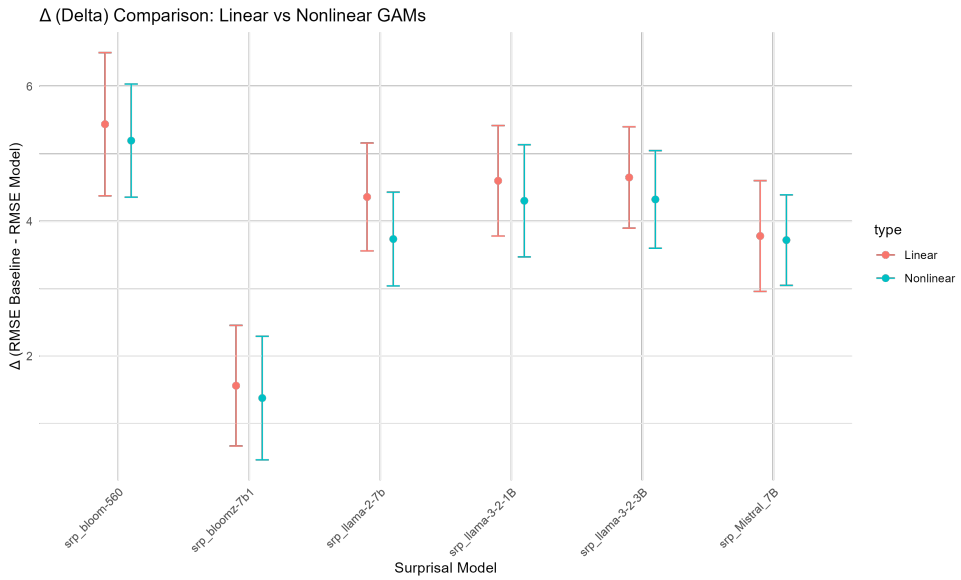

Figure 3: Comparison Between Linear and Non-linear Models for the prediction of total fixation duration. Dots indicate mean Δ log-likelihood per word; error bars represent the standard error of the mean delta RMSE across cross-validation folds..

fects between linear and non-linear models for each LLM. A linear fit can be observed, especially in the denser regions of surprisal values. Notably, the model bloom-7b-1, which showed the poorest Δ values when the effect of surprisal was analysed, also exhibits the greatest visual deviation from linearity in the non-linear model.

Similar results are observed for both first fixation and gaze durations, although the non-linear models exhibit substantially larger error bars for the first fixation measures.

4.4 Part-of-Speech Analysis

As a complementary analysis, we investigated the linearity of the relationship between surprisal estimates and eye-tracking measures across different parts of speech (PoS) in the RastrOS corpus.

To do this, we conducted ordinary least squares (OLS) linear regression analyses on data aggregated by PoS. Entries with erroneous PoS tags (i.e., "ERR", which appeared twice in the corpus) were excluded. For each PoS category, we computed the mean values of surprisal estimates from six language models, as well as mean fixation durations.

For each surprisal model, we then performed an OLS regression using SciPy's linregress function, obtaining the slope, intercept, coefficient of determination (R^2), p-value, and standard error.

Table 2 presents the slope, R^2, and p-value from the OLS regression for each language model across parts of speech.

Model	slope	R^2	p-value
bloom-560	71.88	0.86	<0.001
bloom-7B1	45.86	0.62	<0.001
llama-2-7B	38.82	0.53	<0.001
llama-3-2-1B	35.95	0.51	<0.001
llama-3-2-3B	47.91	0.57	<0.001
mistral-7B	65.05	0.71	<0.001

Table 2: Slope, R^2, and p-values from OLS regressions of surprisal estimates (per LLM) on total fixation duration aggregated by part of speech..

The LLM with the highest R^2 value is bloom-560, followed by mistral-7B. Indicating that for this aggregated analysis in terms of PoS, the smallest model gave the best results. However, in this case, we considered a simpler regression analysis, considering only the fixed effect of surprisal.

Figure 5 presents the linear regression plot obtained using bloom-560, showing the estimated mean total fixation time (i.e., IA_DWELL_TIME) as a function of the mean surprisal value for each part of speech (PoS) in RastrOS.

As expected, parts of speech typically associated with longer word forms and higher information load (e.g., verbs, nouns, and adjectives) exhibit higher values of both reading time and surprisal. In contrast, conjunctions, determiners, and pronouns show lower values, while auxiliary verbs are the least surprising and associated with the shortest reading times. The same tendency was observed

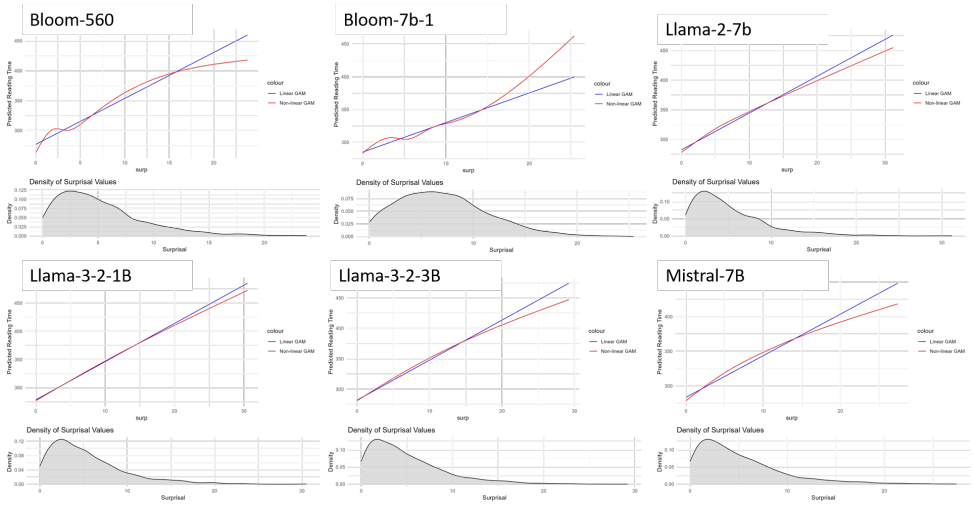

Figure 4: Surprisal versus reading time relationship: Non-linear GAMs are in red and linear control GAMs are in blue. Grey subplots indicate the distribution of surprisal values.

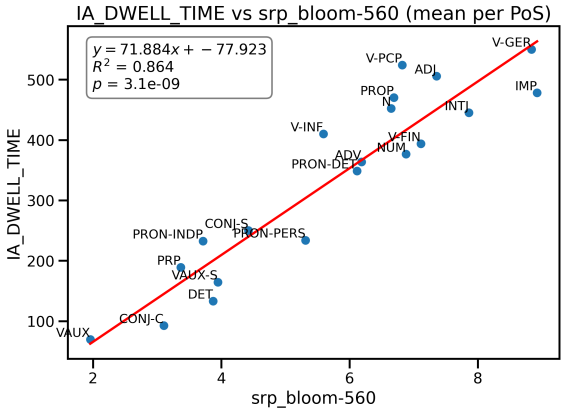

Figure 5: Linear regression plot of mean surprisal estimates against total fixation duration for each part-of-speech (PoS) category for bloom-560. Each point represents a PoS tag, labeled accordingly. Red lines indicate the best-fit regression line.

for all LLMs.

5 Conclusion

In this study, we examined the ability of surprisal estimates from six publicly available large language models (LLMs) to predict reading times in Brazilian Portuguese (BP), using eye-tracking data from the RastrOS corpus. Our findings confirm that surprisal significantly correlates with three key reading time measures (i.e., first fixation duration, gaze duration, and total fixation time) supporting the role of probabilistic predictability in BP processing.

The best-performing model, Llama-3-2-3B, appears to outperform others, including larger or fine-tuned models such as Bloomz-7b1, suggesting that model architecture and training data quality may be more important than sheer size. Moreover, the relationship between surprisal and reading times was consistently linear, aligning with previous findings. However, entropy reduction, calculated from Cloze norms, provided minimal additional predictive power.

These results extend surprisal-based research to BP and offer a baseline for model selection in future studies.

6 Limitations

Several limitations should be noted, first, the RastrOS corpus, though carefully constructed, is relatively small and genre-biased (e.g., dominated by journalistic texts), which may limit the generalizability of our findings.

Second, the language models tested were primarily trained on multilingual data with unclear proportions of Portuguese, and none were specifically optimized for BP. This raises questions about whether models trained exclusively on BP data might provide better fits.

Third, our entropy reduction analysis relied on Cloze norms rather than model-derived entropy, potentially underestimating its predictive power.

Finally, while we focused on surprisal as a key predictor, other linguistic factors, such as syntactic complexity, were not considered and may have an impact on reading time variance.

Acknowledgments

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Sandra M Aluísio, Gisele Montilha Pinheiro, Aline MP Manfrin, Leandro HM de Oliveira, Luiz C Genoves Jr, and Stella EO Tagnin. 2004. The lácio-web: Corpora and tools to advance brazilian portuguese language investigations and computational linguistic tools. In *LREC*.

Kristijan Armeni, Roel M Willems, and Stefan L Frank. 2017. Probabilistic language models in cognitive neuroscience: Promises and pitfalls. *Neuroscience & Biobehavioral Reviews*, 83:579–588.

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Gabor Grothendieck, Peter Green, and Maintainer Ben Bolker. 2015. Package 'lme4'. *convergence*, 12(1):2.

Eckhard Bick. 2000. *The parsing system palavras: Automatic grammatical analysis of Portuguese in a constraint grammar famework*. Aarhus Universitetsforlag.

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

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2440–2452, Marseille, France. European Language Resources Association.

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

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.

Jacob Louis Hoover, Morgan Sonderegger, Steven T Piantadosi, and Timothy J O'Donnell. 2023. The plausibility of sampling as an algorithmic theory of sentence processing. *Open Mind*, 7:350–391.

JL Hoover, M Sonderegger, and TJ O'Donnell. 2022. With better language models, processing time is superlinear in surprisal (poster). york, england.

T. Jaeger and Roger Levy. 2006. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Sidney Evaldo Leal, Katerina Lukasova, Maria Teresa Carthery-Goulart, and Sandra Maria Aluísio. 2022. Rastros project: Natural language processing contributions to the development of an eye-tracking corpus with predictability norms for brazilian portuguese. *Language Resources and Evaluation*, 56(4):1333–1372.

Yi-Chien Lin and William Schuler. 2025. Surprisal from larger transformer-based language models predicts fmri data more poorly. *arXiv preprint arXiv:2506.11338*.

Tong Liu, Iza Škrjanec, and Vera Demberg. 2023. Temperature-scaling surprisal estimates improve fit to human reading times–but does it do so for the" right reasons"? *arXiv preprint arXiv:2311.09325*.

Matthew W Lowder, Wonil Choi, Fernanda Ferreira, and John M Henderson. 2018. Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive science*, 42:1166–1183.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the uniform information density hypothesis. *arXiv preprint arXiv:2109.11635*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Sathvik Nair and Philip Resnik. 2023. Words, subwords, and morphemes: what really matters in the surprisal-reading time relationship? *arXiv preprint arXiv:2310.17774*.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

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

Tony Berber Sardinha. 2010. Corpus brasileiro. *Informática*, 708:0–1.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mgpt: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

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

Jorge A Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brwac corpus: A new open resource for brazilian portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Daphne P Wang, Mehrnoosh Sadrzadeh, Miloš Stanojević, Wing-Yee Chow, and Richard Breheny. 2025. Extracting structure from an llm-how to improve on surprisal-based models of human language processing. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4938–4944.

Tony T Wang, Miles Wang, Kaivalya Hariharan, and Nir Shavit. 2023. Forbidden facts: An investigation of competing objectives in llama-2. *arXiv preprint arXiv:2312.08793*.

Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023a. Testing the predictions of surprisal theory in 11 languages. *arXiv preprint arXiv:2301.12345*.

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. *arXiv preprint arXiv:2006.01912*.

Ethan Gotlieb Wilcox, Clara Isabel Meister, Ryan Cotterell, and Tiago Pimentel. 2023b. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511. Association for Computational Linguistics.

BigScience Workshop. 2022. Bloom: Bigscience language open-science open-access multilingual language model. https://huggingface.co/bigscience/bloom. International collaboration, May 2021–May 2022.

Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721.

A Appendix A

For the tests assessing the effect of surprisal, we use the following models: (1) Baseline and (2) with surprisal.

$$\begin{aligned}\text{reading\_time} \sim\ & \text{Freq\_brWaC\_log} \\ & + \text{Word\_Length} \\ & + \text{prev\_freq} + \text{prev\_len} \\ & + \text{prev2\_freq} + \text{prev2\_len} \\ & + (1 \mid \text{SESSION\_LABEL})\end{aligned} \tag{1}$$

$$\begin{aligned}\text{reading\_time} \sim\ & \text{prev\_surp} \\ & + \text{prev2\_surp} \\ & + \text{Freq\_brWaC\_log} \\ & + \text{Word\_Length} \\ & + \text{prev\_freq} + \text{prev\_len} \\ & + \text{prev2\_freq} + \text{prev2\_len} \\ & + (1 \mid \text{SESSION\_LABEL})\end{aligned} \tag{2}$$

B Appendix B

For the tests assessing the effect of entropy reduction, we use the model 2 in Appendix A as baseline and (3) with entropy.

$$\begin{aligned}\text{reading\_time} \sim\ & \text{prev\_surp} \\ & + \text{prev2\_surp} \\ & + \text{entropy\_Reduction} \\ & + \text{prev\_entropy} \\ & + \text{prev2\_entropy} \\ & + \text{Freq\_brWaC\_log} \\ & + \text{Word\_Length} \\ & + \text{prev\_freq} + \text{prev\_len} \\ & + \text{prev2\_freq} + \text{prev2\_len} \\ & + (1 \mid \text{SESSION\_LABEL})\end{aligned} \tag{3}$$

C Appendix C

The GAM formula used for non-linear models we use is:

$$\begin{aligned}\text{reading\_time} \sim\ & s(\text{surp},\ bs = \text{"cr"},\ k = 6) \\ & + s(\text{prev\_surp},\ bs = \text{"cr"}, \\ & k = 6) \\ & + te(\text{Freq\_brWaC\_log}, \\ & \text{Word\_Length},\ bs = \text{"cr"}) \\ & + te(\text{prev\_freq},\ \text{prev\_len}, \\ & bs = \text{"cr"})\end{aligned} \tag{4}$$

And for linear models:

$$\begin{aligned}\text{reading\_time} \sim\ & \text{surp} + \text{prev\_surp} \\ & + te(\text{Freq\_brWaC\_log}, \\ & \text{Word\_Length},\ bs = \text{"cr"}) \\ & + te(\text{prev\_freq},\ \text{prev\_len}, \\ & bs = \text{"cr"})\end{aligned} \tag{5}$$

EgoDrive: Egocentric Multimodal Driver Behavior Recognition using Project Aria

Michael Rice
University of Galway
School of Computer Science
Galway, Ireland
m.rice11@universityofgalway.ie

Lorenz Krause
University of Galway
School of Computer Science
Galway, Ireland

Waqar Shahid Qureshi
University of Galway
School of Computer Science
Galway, Ireland

Abstract

Egocentric sensing using wearable devices offers a unique first-person perspective for driver behavior analysis and monitoring, with the potential to accurately capture rich multimodal cues such as eye gaze, head motion, and hand activity directly from the driver's viewpoint. In this paper, we introduce a multimodal driver behavior recognition framework utilizing Meta's Project Aria smart glasses, along with a novel, synchronized egocentric driving dataset comprising high-resolution Red Green Blue (RGB) video, gaze-tracking data, Inertial Measurement Unit (IMU) signals, hand pose landmarks, and YOLO-based semantic object detections. All sensor data streams are temporally aligned and segmented into fixed-length clips, each manually annotated with one of six distinct driver behavior classes: *Driving*, *Left Mirror Check*, *Right Wing Mirror Check*, *Rear-view Mirror Check*, *Mobile Phone Usage*, and *Idle*. We design a Transformer-based recognition framework in which each modality is processed by a specialized encoder and then fused via Temporal Transformer layers to capture cross-modal temporal dependencies. To investigate the trade-off between accuracy and efficiency for real-time deployment, we introduce two model variants: EgoDriveMax, optimized for maximum accuracy, and EgoDriveRT, designed for real-time performance. These models achieve classification accuracies of 98.6% and 97.4% respectively. Notably, EgoDriveRT delivers strong performance despite operating with only 104K parameters and requiring just 2.65 ms per inference without the use of a specialized graphical processing unit—highlighting its potential for efficient, real-time in-cabin driver monitoring.

Figure 1: Project Aria Glasses. (Engel et al., 2023)

1 Introduction

Egocentric sensing offers powerful capabilities for capturing and interpreting human behavior in complex, real-world scenarios. In particular, the fusion of diverse sensor modalities can provide a rich, temporally aligned representation of user actions. However, integrating these heterogeneous data streams in a unified framework while ensuring real-time performance poses substantial technical challenges. Driver behavior analysis and action recognition provide a compelling and high-stakes application domain to explore and evaluate such systems.

In this work, we investigate the technical feasibility of such an approach to driver behavior recognition through a proof-of-concept system using Meta's Project Aria glasses (Engel et al., 2023). Our approach integrates high-resolution RGB video, eye gaze tracking, hand pose landmarks, IMU data, and semantic object detections to recognize six driver behaviors. We demonstrate

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025,
pages 18–25, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-003

that effective multimodal fusion can be achieved while maintaining real-time performance, introducing two Transformer-based architectures that explore the accuracy-efficiency trade-off: EgoDriveMax and EgoDriveRT.

Our contributions are threefold: (1) we demonstrate the technical feasibility of real-time driver behaviour recognition using multimodal egocentric sensing, (2) we propose two efficient Transformer-based architectures that achieve high accuracy under strict latency and resource constraints, and (3) we introduce a proof-of-concept style, egocentric driving dataset comprising the aligned aforementioned data streams. Although our evaluation is conducted in a controlled setting with a singular participant and vehicle, the consistently strong performance across diverse driver actions indicates the potential for scalable deployment in real-world driver monitoring systems.

2 Related Work

Egocentric Vision. A rapidly growing area within computer vision, primarily driven by advances in wearable and augmented reality technologies. Meta are an established force in this domain, particularly in the open-source ecosystem, due to major contributions such as the Ego4D dataset (Grauman et al., 2021), the Project Aria initiative itself (Engel et al., 2023) and the HOT3D dataset (Banerjee et al., 2024), among others. Interest has also begun to permeate through into the automotive research space with implementations such as EgoFormer (Qazi et al., 2024), EgoSpeed-Net (Ding et al., 2022) and others paving the way for egocentric driver behavior modeling and in-cabin understanding.

Beyond Meta and the automotive sector, the academic egocentric vision landscape includes several influential datasets and methodologies. Epic-Kitchens-100 (Damen et al., 2020) provides fine-grained action recognition in kitchen environments, while EGTEA Gaze+ (Li et al., 2018) combines egocentric video with gaze data for activity understanding. Recent advances in egocentric representation learning include EgoVLM (Vinod et al., 2025) for vision-language understanding and EgoNCE (Lin et al., 2022) for self-supervised learning from temporal relationships.

Multimodal Learning. Effective multimodal learning requires architectures capable of aligning and fusing heterogeneous data streams with varying sampling rates and representational characteristics. Recent work has explored various fusion strategies, from early concatenation to attention-based approaches. Meta also possess a strong foothold in this research community, with implementations such as 'Reading in the Wild' (Yang et al., 2025) demonstrating transformer-based multimodal fusion using RGB, head pose, and eye-tracking data, for the recognition of the reading action in a variety of scenarios. Also created by Meta's researchers, Moon et al. (2023)'s IMU2CLIP work represents a significant advance in aligning IMU sensor data with textual representations through contrastive learning, displaying how motion sensors can be integrated into multimodal frameworks and providing a potential avenue for resource efficient human action recognition via motion-to-text conversion.

Many of the recent advancements in this area have been driven by either transformer-based architectures or contrastive learning focused approaches. Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) and its variants demonstrate effective cross-modal alignment through contrastive objectives, while works like VLMo (Bao et al., 2022) explore 'Mixture-of-Modality-Experts' based approaches for vision-language tasks.

Automotive Action Recognition. Action Recognition implementations for automotive applications in the academic world have traditionally relied on exocentric cameras and/or single-modality approaches. Martin et al. (2019)'s Drive&Act dataset represents the most comprehensive effort in this space, utilizing multiple cameras types alongside pose estimation for robust driver behavior recognition in numerous lighting conditions from the third-person perspective. Furthermore, several other works such as those from Lin et al. (2021) and Li et al. (2024) explore alternative methodologies such ss RGB-D cameras and mmWave radars for driver-centric behavior identification.

Hoskeri (2023)'s proof-of-concept work comes closest to our approach, demonstrating the feasibility of using smart glasses with forward-facing cameras and IMU sensors for basic driver monitoring. Their controlled lab-based study achieved strong performance (93-99% F1) on limited steering and head movement patterns, establishing initial feasibility but leaving open questions about multimodal integration and real-world deployment.

However, the landscape of driver behavior recognition also includes a wide range of non-academic implementations. In the commercial sector, Tesla's

cabin-facing camera system and Seeing Machines' Driver Monitoring Systems (DMS) represent current industry standards, typically achieving 95%+ accuracy for basic attention detection but with limited behavioral granularity. Smart Eye's AI-powered systems demonstrate real-time gaze tracking capabilities, though primarily for attention monitoring rather than detailed action recognition.

Finally, the challenge of achieving real-time performance with such systems is an extremely pertinent one and has driven research into numerous efficiency focused architectures, with those from the commercial domain subject to much more stringent regulations than those from academia.

3 EgoDrive Dataset

To investigate the technical feasibility of multimodal egocentric driver behavior recognition, we developed a proof-of-concept style dataset. Our dataset design prioritizes technical requirements over scale and scope, with the resulting dataset potentially serving as a template for future multimodal egocentric behavioral analysis studies captured using Project Aria (Engel et al., 2023).

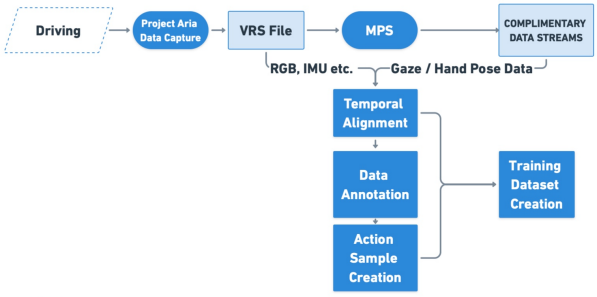

Figure 2: Dataset Creation Flowchart.

3.1 Data Gathering

As previously stated, all data was captured using Meta's Project Aria glasses as a single, integrated sensing platform. A major advantage of this approach is its inherent temporal synchronization across modalities, simplifying the reliable alignment of modality-specific timestamps by ensuring all data is referenced to a common device-time. We selected RGB camera (15fps), eye-tracking cameras (30fps), Simultaneous Localization and Mapping (SLAM) cameras (15fps), and IMUs (800Hz and 1KHz) based on their complementary roles in behavioral analysis: visual context, attention tracking, spatial awareness, and motion dynamics respectively. All recording adhered to General Data Protection Regulations (GDPR) , including informed consent where feasible, anonymization in post-processing, and secure data handling (GDPR, 2016).

A controlled, single-participant approach allows for the isolation of technical challenges in multimodal processing for a proof-of-concept based study, without confounding factors from inter-participant variability. Following the culmination of the data capture process, hand tracking and gaze estimates were obtained through Meta's Machine Perception Services (MPS).

3.2 Dataset Creation

Creating a temporally aligned multimodal dataset from asynchronous sensor streams poses several technical challenges, which our methodology had to overcome.

RGB timestamps are designated as the primary temporal reference. IMU data—comprising 6D input from both the accelerometer(3D) and gyroscope(3D) —is linearly interpolated to match the RGB frame rate, resulting in a standardized sample shape of *(Sequence Length, IMU Hz / RGB FPS, 6)*. Gaze data, sampled at twice the RGB rate, is temporally aligned using a simplified, mean nearest-neighbor matching strategy. This produces a single (x, y) pixel-coordinate gaze point per RGB frame and results in a sample shape of *(Sequence Length, 2)*. Hand landmark data, returned by Meta's Machine Perception Services (MPS), at the same sampling rate as the RGB stream, required no resampling. Each sample is thus represented with a shape of *(Sequence Length, 8)*, where each 8-dimensional vector corresponds to the x-y positions of the left and right wrists and palms.

This alignment pipeline maintains temporal coherence across modalities, while preserving each sensor's native sampling behavior. In addition, semantic context was incorporated through object detections generated by a custom-trained YOLOv11 model tailored for in-cabin environments (see Section 4 for training details). Each frame's detections are encoded as a fixed-length feature vector, with details limited to four key objects. Each of the four objects was represented using five dimensions: a binary presence indicator (0 or 1), the x and y coordinates of the top-left corner of the bounding box, followed by the bounding box's height and width.

Figure 3: Annotated Dataset Samples

Following on from the dataset's compilation, manual frame-by-frame annotation mapped frame indices to six behaviorally relevant classes: *Driving*, *Left Wing Mirror Check*, *Right Wing Mirror Check*, *Rear-view Mirror Check*, *Mobile Phone Usage*, and *Idle*. These classes were selected to represent distinct attention patterns and physical actions that create differentiable multimodal signatures. Each training sample spans **32** frames (2.13s), with longer actions segmented into multiple samples and shorter actions padded to maintain consistent temporal context. The final dataset, processed for training, consisted of 2,448 samples, with a real-world consistent bias towards the *'Driving'* action, with the exact class distribution visible in Figure 4.

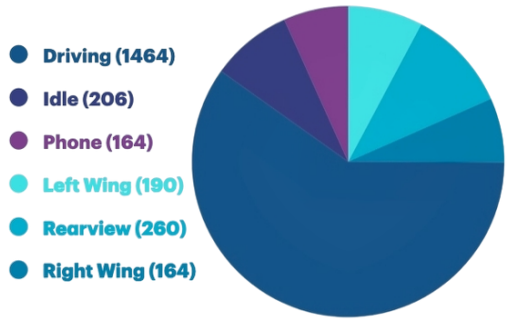

Figure 4: Dataset Class Distribution

4 Methodology

To address the challenge of fusing asynchronous and heterogeneous sensor streams, we adopt a modular processing pipeline centred around modality-specific encoders and transformer blocks. Full training methodologies for both the final models as well as the in-cabin object detection model are detailed below.

4.1 Object Detection Model

As previously stated, object detections for this implementation resulted from the training of a custom object detection model. Training frames were randomly sampled from the RGB streams of the main dataset and manually annotated with eight object classes: *Right Wing Mirror*, *Left Wing Mirror*, *Rearview Mirror*, *Gear Stick*, *Infotainment Unit*, *Speedometer*, *Steering Wheel*, and *Mobile Phone*. This process resulted in a dataset of over 4,000 annotated images. Once annotation was complete, the dataset was divided into an 80/10/10 train/validation/test split. A YOLOv11 backbone for fine-tuning was selected for its balance of efficiency and performance, achieving a precision of *96.5%* and a mAP50–95 of *88.1%* after training.

4.2 Model Architectures

We designed a Transformer-based architecture to address the core technical challenge of fusing heterogeneous sensor streams with different sampling rates and representational characteristics. Our approach processes each modality through specialized, unimodal encoders that extract meaningful features, which are then projected into a shared embedding space and passed through Temporal Transformer blocks for cross-modal reasoning.

Each sensor stream requires tailored processing to handle its unique characteristics. The RGB encoder processes visual sequences *(B, T, C, H, W)* using a pretrained *Swin-Tiny* Video Transformer (Liu et al., 2021) for spatial features, complemented by a *ResNet-18* (He et al., 2015) motion stream computing frame differences. Both streams are fused through projection networks and temporal 1D convolutions.

The gaze encoder projects normalized (0-1) x,y coordinates through linear layers, also followed by 1D convolutions, while the hand landmark encoder handles missing landmarks through learnable replacement vectors and validity masks, processing three parallel streams (coordinates, masks, missingness patterns) through feedforward networks and temporal attention.

Finally, object detection features undergo linear projection and 1D convolution for temporal modeling, while the IMU encoder processes signal through stacked 1D Convolutional Neural Networks (CNN) with pooling, followed by Gated Re-

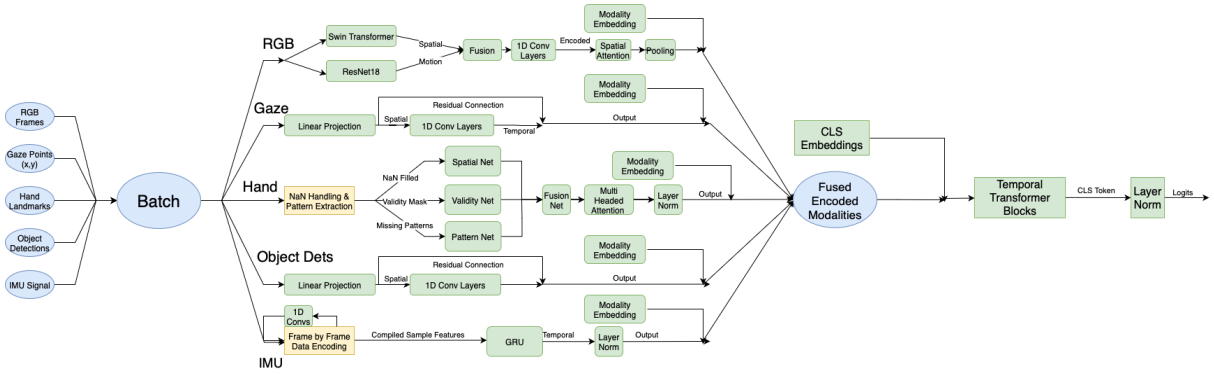

Figure 5: EgoDriveMax Architecture

current Unit (GRU) (Chung et al., 2014) layers for long-term modeling.

Projected features from all available modalities are concatenated and processed through one or more (stacked) Temporal Transformer blocks with multi-head attention, enabling the model to learn complex dependencies between behavioral cues across different sensor streams. A subsequent LayerNorm module is used to stabilize outputs.

4.3 Architectural Variants

To explore the accuracy-efficiency trade-off critical for real-time deployment, we developed two architectural variants that demonstrate different approaches to multimodal processing:

As shown in Table 1, the ***EgoDriveMax*** variant prioritizes absolute accuracy with 2 Transformer blocks, 4 attention heads, 256-dimensional features, and full RGB processing, totaling 42M parameters, while *EgoDriveRT* instead targets real-time performance with 1 block, 2 heads, 32-dimensional features, and RGB encoder removal, resulting in just 104K parameters - a 400x parameter reduction. Dropout was standardized across both models at a value of 0.1.

Model	Blocks	Heads	Feature Dim.	RGB
Max	2	4	256	✓
RT	1	2	32	✗

Table 1: Configuration details for the Max and RT model variants.

4.4 Training

All models were trained using a 60/20/20 train/validation/test split for a maximum of 20 epochs, with early stopping (patience = 5) to prevent overfitting. Optimization was performed using the Adam optimizer (Kingma and Ba, 2017) with a learning rate of 1×10^{-4}. Categorical Cross-Entropy loss was applied due to the multi-class nature of the task. To address class imbalance, loss weighting was used to increase the penalty for misclassifying underrepresented classes.

EgoDriveMax was trained on a single NVIDIA A100 GPU using Google Colab, while EgoDriveRT was trained locally on an Apple M4 chip. All training metrics and experiment logs were tracked using Weights & Biases (W&B).

Figure 6 shows the validation accuracy curve for EgoDriveRT, illustrating stable and smooth convergence. The EgoDriveMax model exhibited comparable convergence behavior during training.

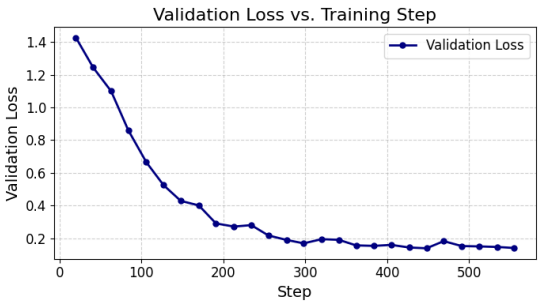

Figure 6: EgoDriveRT Validation Curve

5 Results

5.1 Proof-of-Concept Validation

We evaluate our approach to assess the technical feasibility of real-time multimodal driver behavior recognition. Our results demonstrate that effective multimodal fusion can achieve strong performance while maintaining practical inference constraints.

Table 2 shows our primary finding: the lightweight EgoDriveRT model achieves *97.4%*

accuracy with just *2.65ms* inference time on Apple's M4 chip using the Metal Performance Shaders framework, compared to EgoDriveMax's *98.6%* accuracy at *1595ms*.

Model	Acc	F1	Params	Inf Time
EgoDriveMax	**98.6%**	**98.0%**	42M	1595ms
EgoDriveRT	97.4%	96.6%	**104K**	**2.65ms**

Table 2: Model variant test results.

This *400x* parameter reduction (104K vs 42M) with minimal accuracy loss demonstrates that efficient multimodal architectures can capture essential behavioral patterns without requiring computationally expensive visual processing. Inference results displayed via annotated results from the EgoDriveMax model can be viewed below in Figure 7.

Figure 7: Example Action Detections

5.2 Per-Action Analysis

Table 3 displays the individual per-action results using both model variants, illustrating some interesting findings. The RT model's superior performance on Left Mirror Check (100% vs 96.9%) suggests that for certain actions, the simplified architecture may avoid overfitting to visual features while better leveraging complementary modalities like gaze and head motion.

The consistently strong performance across all actions using both models supports the effectiveness of the core technical approach for distinguishing behaviorally relevant driver actions. However, this performance may also be partially influenced by the controlled scope of the study; in broader, more diverse scenarios, a decline in performance would be a reasonable expectation.

Action	Acc	Prec	Rec	Model
Left Wing Mirror	96.9%	100%	96.9%	Max
Left Wing Mirror	**100%**	**100%**	**100%**	RT
Right Wing Mirror	**97.4%**	***100%***	**97.44%**	Max
Right Wing Mirror	94.9%	***100%***	94.9%	RT
Rearview Mirror	**97.9%**	97.9%	**97.9%**	Max
Rearview Mirror	91.2%	**100%**	91.2%	RT
Mobile Phone	94.1%	**94.1%**	94.1%	Max
Mobile Phone	**96.3%**	89.7%	**96.3%**	RT
Driving	**99.3%**	**98.7%**	**99.3%**	Max
Driving	98.7%	97.3%	98.7%	RT
Idle	**100%**	**100%**	**100%**	Max
Idle	97.1%	97.1%	97.1%	RT

Table 3: Model variant test results.

5.3 Ablation Study

To evaluate the contribution of each modality to overall performance, we conducted five additional training runs of the EgoDriveMax model, each time removing a singular modality. Due the complexity and robustness introduced by the multimodal setup, the model maintained strong performance across all ablations. Nonetheless, several meaningful trends emerged.

Modalities	Acc	Prec	Rec	F1
All	**98.6%**	**98.5%**	**97.6%**	**98.02%**
w/o Obj Dets	97.6%	97.4%	96.5%	96.9%
w/o Gaze	98.0%	97.6%	97.2%	97.4%
w/o RGB	97.4%	96.6%	97.3%	96.9%
w/o Hands	98.2%	98.0%	97.2%	97.6%
w/o IMU	97.4%	96.3%	97.3%	96.7%

Table 4: Ablation test results across different modality combinations.

As expected, the configuration using all available modalities achieved the highest scores across all evaluation metrics. Analyzing the F1 scores from the ablation runs, the IMU stream was found to be the most influential, providing the most discriminative features to the model. This was followed by object detections and RGB video frames, both of which contributed significantly. In contrast, the removal of gaze features and hand landmarks led to only minor drops in performance. This suggests that these modalities may be partially redundant, with their information content potentially approximated by other inputs—e.g., gaze direction could be inferred from a combination of object detection

bounding box locations and IMU-based motion patterns, reducing the utility of explicit gaze data.

6 Conclusions and Limitations

This work demonstrates the technical feasibility of real-time multimodal egocentric driver behavior recognition using wearable sensors. The most significant finding is that our lightweight *EgoDriveRT* model achieves near-optimal performance (97.4% accuracy) with 400x fewer parameters than the *EgoDriveMax* model and sub-3ms inference times. This efficiency suggests that the rich behavioral information captured through gaze tracking, hand pose, IMU data, and semantic object detection may be sufficient for accurate driver action recognition without computationally expensive visual processing.

Our single-participant controlled study validates the core technical approach, though inherently limits the generalizability of findings to broader populations and while the results certainly establish technical feasibility, real-world deployment would require validation across diverse drivers, vehicles, and environmental conditions, as well as an expanded action set, to ensure robust performance.

7 Future Work

Future research should prioritize multi-participant validation to capture inter-individual variability and explore on-device deployment strategies to preserve user privacy. The modular design also opens opportunities for personalization and continuous learning in long-term deployments. Furthermore, should future generations of the Project Aria device include onboard compute, this work presents the foundations for the development of a fully self-contained driver monitoring system. Finally, the architectural insights and dataset methodology presented here offer a strong foundation for building scalable, efficient, and context-aware egocentric driver monitoring systems.

References

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. 2024. Introducing HOT3d: An egocentric dataset for 3d hand and object tracking.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. 2022. VLMo: Unified vision-language pre-training with mixture-of-modality-experts.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2020. The EPIC-KITCHENS dataset: Collection, challenges and baselines.

Yichen Ding, Ziming Zhang, Yanhua Li, and Xun Zhou. 2022. EgoSpeed-net: forecasting speed-control in driver behavior from egocentric video data. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '22, pages 1–10. Association for Computing Machinery.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. 2023. Project aria: A new tool for egocentric multi-modal AI research.

GDPR. 2016. Regulation - 2016/679 - EN - gdpr - EUR-lex.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar,

Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2021. Ego4d: Around the world in 3,000 hours of egocentric video.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Rahul Hoskeri. 2023. Poster abstract: Driving behavior monitoring with unobtrusive smart-glasses.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Guan-Hua Li, Hsin-Che Chiang, Yi-Chen Li, Shervin Shirmohammadi, and Cheng-Hsin Hsu. 2024. A driver activity dataset with multiple RGB-d cameras and mmWave radars. In *Proceedings of the 15th ACM Multimedia Systems Conference*, MMSys '24, pages 360–366. Association for Computing Machinery.

Yin Li, Miao Liu, and James M. Rehg. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Computer Vision – ECCV 2018*, volume 11207 of *Lecture Notes in Computer Science*, pages 639–655. Springer International Publishing.

Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. 2022. Egocentric video-language pretraining.

Zeyang Lin, Yinchuan Liu, and Xuetao Zhang. 2021. Driver-skeleton: A dataset for driver action recognition. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1509–1514.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows.

Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon ReiB, Michael Voit, and Rainer Stiefelhagen. 2019. Drive&act: A multimodal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2801–2810. IEEE.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2023. IMU2clip: Language-grounded motion sensor translation with multimodal contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13246–13253. Association for Computational Linguistics.

Tayeba Qazi, M. Rupesh Kumar, Prerana Mukherjee, and Brejesh Lall. 2024. EgoFormer: Ego-gesture classification in context of autonomous driving. 24(11):18133–18140.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Ashwin Vinod, Shrey Pandit, Aditya Vavre, and Linshen Liu. 2025. EgoVLM: Policy optimization for egocentric video understanding.

Charig Yang, Samiul Alam, Shakhrul Iman Siam, Michael J. Proulx, Lambert Mathias, Kiran Somasundaram, Luis Pesqueira, James Fort, Sheroze Sheriffdeen, Omkar Parkhi, Carl Ren, Mi Zhang, Yuning Chai, Richard Newcombe, and Hyo Jin Kim. 2025. Reading recognition in the wild.

Comparing Eye-gaze and Transformer Attention Mechanisms in Reading Tasks

Maria Mouratidi
Utrecht University
m.mouratidi@students.uu.nl

Massimo Poesio
Utrecht University
Queen Mary University of London
m.poesio@uu.nl

Abstract

As transformers become increasingly prevalent in NLP research, evaluating their cognitive alignment with human language processing has become essential for validating them as models of human language. This study compares eye-gaze patterns in human reading with transformer attention using different attention representations (raw attention, attention flow, gradient-based saliency). We employ both statistical correlation analysis and predictive modeling using PCA-reduced representations of eye-tracking features across two reading tasks. The findings reveal lower correlations and predictive capacity for the decoder model compared to the encoder model, with implications for the gap between behavioral performance and cognitive plausibility of different transformer designs.

1 Introduction

The impressive capabilities of Transformer models in linguistic tasks have revolutionized Language Models in Natural Language Processing (NLP) research. A key difference in their architecture from previous models is the incorporation of an attention mechanism, which assigns a degree of relevance between words in the input. Previous work has shown that transformer models show signs of processing steps similar to humans (Clark et al., 2019; Voita et al., 2019), and tend to mirror the structure of the classic NLP pipeline (Tenney et al., 2019).

Attention during reading has also been extensively studied in human eye-movement research. Eye-movements track much of linguistic processing, including both lower-level word processing (Just and Carpenter, 1980; Clifton Jr. et al., 2007) and higher-level comprehension (Reichle et al., 2010; Southwell et al., 2020).

While transformer attention is not explicitly modeled after human attention in text processing, both mechanisms seem to process text by allocating resources on relevant linguistic targets. This similarity, combined with the broader effort of explainability research to explain artificial models in human terms, has driven comparisons of model attention to human eye-movement patterns. From a cognitive science perspective, the goal of this comparison is to determine the cognitive plausibility of computational models like transformers. This involves understanding whether they merely achieve high, human-like performance due to genuinely assimilating the human cognitive process, or due to other artificial processes learned independently.

Previous work (Kozlova et al., 2024; Bensemann et al., 2022; Eberle et al., 2022; Morger et al., 2022; Wu et al., 2024; Hollenstein and Beinborn, 2021; Brandl and Hollenstein, 2022) has investigated this parallel using various techniques to extract attention scores from transformers and compare them to eye-movements from established eye-tracking datasets. However, several knowledge gaps exist. Most existing literature has focused on encoder transformer models, leaving open the question of whether more advanced and recent decoder models can equally align with eye-movements. Additionally, since many studies neglect the impact of low-level text properties on eye-movements, any correlation driven primarily by these surface-level features would be insufficient evidence of deeper cognitive alignment. Finally, eye-tracking datasets consist of multiple eye-tracking features that provide informative signals regarding reading patterns. However, these features are often intercorrelated, so they may capture redundant aspects of the same underlying attention mechanism. The literature has not been able to consolidate overlapping information from multiple features into a single analysis, where previous studies most often focus arbitrarily on a single metric.

To address the identified knowledge gaps, this

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025,
pages 26–36, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-004

study compares the attention mechanism of a decoder-only model with human attention during reading. It employs both correlation analysis and predictive modeling using PCA-reduced representations of eye-tracking features and accounting for surface-level properties of the text. Moreover, the effect of different attention representation methods (raw attention, attention flow, and gradient-based saliency) is investigated on the results. The results provide insights into how the model architecture, attention method, and reading task collectively influence the similarity of model attention patterns to eye-movement behavior.

2 Background

2.1 Human attention and Eye-Movements

Eye-movement research has a long and successful history in studying human cognitive tasks. Eye-movements in reading are shown to provide information about cognitive language processing, like syntactic parsing and semantic integration (Frazier and Rayner, 1982), expectations about the text (Ehrlich and Rayner, 1981), and reading goals (Rayner, 2009).

Eye-movements consist of fixations and saccades (Rayner et al., 2006). Saccades are short, rapid movements to other parts of the text, while fixations occur when eyes remain stationary in between saccades (Reichle et al., 2003) and are considered the key point of information processing. Eye-tracking measures focus on different aggregations of these movements, such as the total fixated time on each word. Both low-level bottom-up processing and higher-level comprehension are reflected in eye-tracking measures. Familiar (Clifton Jr. et al., 2007), or frequently occurring words (Inhoff and Rayner, 1986) are subject to faster processing, whereas rare occurring words such as novel proper nouns tend to have longer fixations (Barrett and Hollenstein, 2020). Longer words also receive longer fixations, while shorter words are more likely to be skipped. Word length additionally interacts with the functional role of a word, where function words are fixated less than content words (Rayner, 2009). More top-down processes like assessing the predictability of the text given the preceding context draws shorter gaze durations and the reverse holds for unpredictable, surprising words (Ehrlich and Rayner, 1981).

2.2 The Rise of Transformer Models and the Quest for Interpretability

Transformer models were introduced when Vaswani et al. (2017) proposed attention as a novel method for handling contextual relations in language models. The attention mechanism compares different embedding representations of the sequence to determine the degree of relevance between each pair of words. The model then attends to important parts of the sequence depending on these relevance scores. The exceptional performance of these models and the intuition and transparency of the attention mechanism drew much attention from both deep learning and interpretability research. As transformer models are advancing, there is a growing demand to interpret not only their outputs, but also the internal mechanisms that lead to those outputs. The first advancements in interpretability emerged from the Computer Vision field (Simonyan et al., 2014; Zeiler and Fergus, 2014), where saliency maps were used to trace model decisions back to input pixels. This technique estimates the contributions of input raw data or intermediate activations to model predictions (Li et al., 2022) and is also commonly applied to NLP research. For transformer architectures specifically, attention-based and gradient-based methods have gained popularity for representing importance allocated to the input sequence.

2.2.1 The interpretability debate

Raw attention scores extracted directly from the model provide an easily understandable weighting of the input sequence. For example, in the original paper introducing attention, Vaswani et al. (2017) showed that examining the raw attention towards ambiguous pronouns like ”its” could reveal how anaphora resolution is represented in the model.

Previous work correlating raw attention and human eye-gaze shows mixed findings. While Sood et al. (2020) reported non-significant correlations for later layers of models like XLNet, Bensemann et al. (2022), Eberle et al. (2022) and Morger et al. (2022) found strong correlations in early Transformer layers. Bensemann et al. (2022) noted that correlation strength is generally higher in early layers and not dependent on the model’s size, though it can be influenced by the training process. Kozlova et al. (2024) similarly found strong early-layer correlations in the context of anaphora resolution. Eye-gaze features like First Fixation Duration (FFD) are

often found to align better with single-pass model behavior than cumulative measures like Fixation Count (F) or Total Reading Time (TRT), as FFD reflects initial processing (Ikhwantri et al., 2023). Furthermore, research has explored the ability of language models to predict human eye-movements as an indicator of their cognitive plausibility (Hollenstein et al., 2022).

However, the increased focus on faithful explanations opened a debate about the effectiveness of the attention mechanism as an explanation method. Some critics (Jain and Wallace, 2019) argue that raw attention weights do not always strongly correlate with gradient-based measures of feature importance, and that different attention distributions can lead to effectively identical model predictions (Jain and Wallace, 2019; Serrano and Smith, 2019), questioning whether attention provides a unique explanation for the model's behavior. However, Wiegreffe and Pinter (2019) argued that producing identical explanations to gradient-based methods is not necessary for plausible model explanations, especially when the goal shifts from explaining the model's predictions to broadly understanding the model's internal behavior (Bastings and Filippova, 2020). Nevertheless, attention flow and saliency-based methods have been proposed as more suitable for quantifying word importance in sentence processing.

Attention flow (Abnar and Zuidema, 2020) is an interpretation method based on flow networks from graph theory. This method tackles the problem of uniform raw attention in higher layers and models a global view of attention, as it captures the entire information propagation through the network layers. In attention flow, the raw attention graph is treated as a flow network that consists of nodes connected by directed edges. A flow function assigns values to edges such that the maximum total flow from a source node reaches a target node, under some capacity and conservation constraints.

Gradient-based saliency differs from attention-based methods as it does not utilize the transformer's attention mechanism[1]. Instead, it measures how sensitive the model output is to changes in each input token's embeddings. For each target token in the sequence, gradients are computed with respect to all input tokens and are normalized to produce a saliency score for each token.

Hollenstein and Beinborn (2021) found that fixation durations correlate better with saliency-based than with attention-based importance, suggesting saliency as a more cognitively plausible metric for interpretation. Morger et al. (2022) supported this finding for gradient-based saliency and for attention flow, across multiple languages. Similarly, Eberle et al. (2022) observed strong alignment of attention flow with human fixation times in natural reading, competitive with a specialized cognitive model of human reading (E-Z reader).

2.2.2 Alignment in task-specific contexts

Human reading strategies are task-dependent and influence how attention is allocated to different parts of the sequence. Task specificity thus plays a crucial role in the alignment between human and model attention. While Wu et al. (2024) found that finetuning models on task-specific objectives can enhance correlations with human gaze when using saliency methods, Eberle et al. (2022) showed that task-specific finetuning did not significantly increase correlation, and models aligned better with natural reading patterns than with task-specific ones. Brandl and Hollenstein (2022) further demonstrated that more in-depth reading (characterized by longer total reading times and lower skipping rates) generally correlates better with model attention compared to faster, shallow reading.

2.3 Our contribution

Despite evidence that transformer attention patterns align with human reading behavior, most existing work has focused on encoder-only or encoder-decoder architectures, leaving questions about newer decoder-only models that process text left-to-right (Hollenstein and Beinborn, 2021). Additionally, many studies overlook text properties' influence on eye-movements and lack methods for integrating multiple eye-tracking features in the analysis. While Wu et al. (2024) investigates an early decoder-only model (GPT-2), their analysis focused on gradient-based saliency in a task-specific setting. This study compares decoder-only model attention with human attention using raw attention, attention flow, and gradient-based saliency across both natural and task-specific reading. We address three research questions: **(RQ1)** To what extent do human eye-movements correlate with decoder model attention? **(RQ2)** Can decoder models predict eye-movements independently of text fea-

[1]Even though gradient-based saliency relies on input-output gradients rather than attention scores, it is loosely referred to as an *attention method* in this study for brevity, in the sense of attributing importance to the input.

tures like word frequency, length, and surprisal? **(RQ3)** How does Principal Component Analysis of eye-tracking features and task-specificity affect these correlations and predictions?

3 Methods

3.1 Eye-tracking Data

This study uses the Zurich Cognitive Language Processing Corpus (ZuCo) (Hollenstein et al., 2018). ZuCo combines EEG and eye-tracking recordings from English native speakers reading natural sentences. 12 participants read sentences under different conditions (tasks). In Task 2 (”Normal Reading”) the participants were asked to read 300 sentences containing certain relations and answer a comprehension question after each sentence. In Task 3 (”Task-specific Reading”), the participants were instructed to focus on a specific relation type before reading the sentence. 407 sentences were presented in blocks of the same relation so the subjects knew what relation to look for. For each sentence, the participants had to indicate whether the specified relation was present in the sentence or not.

The raw eye-tracking data consists of the following eye-tracking features on the word-level: gaze duration (GD), total reading time (TRT), first fixation duration (FFD), single first duration (SFD), go-past time (GPT), fixation count (F) and mean pupil size (mPS). These features are normalized to their relative value in each sentence and then are averaged across participants to ensure robustness across different sentences and reading behaviors.

This study compares human and model data in two ways: (1) by analyzing each gaze feature individually, and (2) by combining the most informative aspects of these features using Principal Component Analysis (PCA). A PCA representation is derived separately for each task across all sentences within that task, with each word represented by its normalized eye-tracking values. We experimented with different numbers of components to identify the optimal balance between compact representation and captured variance. Ideally, a single component is preferred, as it encodes the most compact and efficient representation of the multidimensional eye-tracking data.

3.2 Model Attention

This study uses Llama 3.1-8B for investigating transformer attention, which is a latest generation decoder model, and BERT-base-uncased for the encoder comparison. Both BERT and Llama models remain in their pretrained states without task-specific fine-tuning because the goal is to investigate the fundamental model alignment with human attention, rather than deliberately optimizing it. For both models, explicit instructions similar to those the participants received are prepended to the input sentences to better resemble the original experiment and guide model attention closer to the human cognitive task. Attention patterns are extracted using standard forward passes without masking to reveal how each model’s attention mechanism responds to the same instructional context during inference.

3.2.1 Raw attention

The input is tokenized and passed through the model to obtain the raw attention scores for each layer, averaged across attention heads. Because tokenization can split original words into smaller subtokens, the attention scores are aligned with the human data by assigning each original word the maximum attention score among its subtokens, following the approach of Sood et al. (2020). The final score for each word in each layer is computed as the average attention it receives from all other words (including instruction words). A normalization by the sum of the total attention is applied to obtain relative attention scores per word and to account for variability across sentences.

3.2.2 Attention flow

Attention flow is calculated using Edmonds-Karp’s maximum-flow algorithm (Edmonds and Karp, 1972). The last token is considered the target ”sink” token in each sentence, where reading presumably ”stops”. For the Llama model, attention flow is implemented under a reduced number of paths to respect its causal attention structure. Finally, a decay is applied to account for the inherent bias to early input tokens in decoder models, using the position-based weighting proposed by Metzger et al. (2022).

3.2.3 Gradient-based saliency

Saliency is calculated by taking the L1 normalized gradient of the model’s output logit with respect to each input token embedding. This process is repeated with each token serving as the prediction target, and the resulting saliency scores are averaged across all targets to obtain a global saliency score for each token. As with previous methods, token-

level saliency scores are combined at the word level and normalized to reflect relative saliency within each sentence.

3.3 Analysis

The eye-tracking data (both PCA-reduced and individual features) are compared word by word with the transformer scores using Spearman's correlation. In addition to correlation analysis, we use linear regression models (ordinary least squares) to assess whether there is a predictive relationship between model attention and eye-gaze, and to measure how additional text features linked to eye-movements may influence this relationship. The predictive relationship is assessed through adjusted R^2 on unseen data (20% of the dataset). We incorporate 5 text-related features as additional predictors alongside model attention: word frequency (across many corpora), word length, functional category (function vs. content words) and surprisal derived from the respective transformer model.

Four regression models are fitted for each combination of transformer model (BERT, Llama), attention method (raw attention, attention flow, gradient-based saliency) and reading task (Task 2, Task 3). To determine whether model attention improves predictive capacity, we compare performance to a baseline model that uses only text features as independent variables. We similarly compare the PCA model to the average performance of models predicting individual eye-gaze features, with PCA predictions transformed back to the original feature space. Both comparisons use the Wilcoxon signed-rank test of mean squared errors between regression models. To further analyze the contribution of attention and text features to the regression models, we visualize feature importance using absolute t-values[2].

4 Results

4.1 Experiment 1: Replication

The correlation analysis between BERT raw attention and human eye-movements successfully replicates previous findings. BERT's first layer shows the highest correlations with Task 2 eye-movements ($\mu = 0.69, \sigma = 0.02$), as indicated by the blue line in Figure 1, which is consistent with results from Morger et al. (2022). Correlation

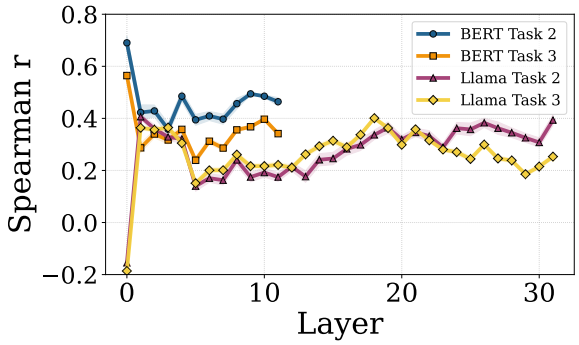

Figure 1: Average layer-wise correlations of each transformer's raw attention across 6 eye-tracking features.

strength generally decreases across subsequent layers. A similar trend is observed in Task 3, where the first layer again exhibits the strongest correlation ($\mu = 0.56, \sigma = 0.008$), as shown by the orange line.

The results with the alternative attention methods also replicate previous findings. Attention flow shows the strongest correlations with human eye-movements. For Task 2, the correlation is $\mu = 0.74, \sigma = 0.007$ (blue plain bar, Figure 2), while for Task 3 $\mu = 0.62, \sigma = 0.007$ (orange plain bar). These results align with Eberle et al. (2022), who found that attention flow produces better alignment with human eye-movements than the strongest correlating layer of raw attention.

Gradient-based saliency performs less strongly across both tasks. Task 2 correlations reach $\mu = 0.68, \sigma = 0.02$ (blue hatched bar, Figure 2), matching the score reported by Hollenstein and Beinborn (2021). Task 3 correlations are $\mu = 0.53, \sigma = 0.005$ (orange hatched bar), similar to results from Wu et al. (2024). This makes saliency the least correlating attention method with human eye-movements for the BERT model. For all attention methods, Task 3 correlations mirror Task 2 patterns but at reduced magnitudes, consistent with previous work (Eberle et al., 2022). All reported correlations are statistically significant ($\alpha < 0.05$).

4.2 Experiment 2: Extension

4.2.1 What about Llama?

Correlations of Llama's raw attention with human eye-movements fall visibly lower than BERT correlations. Llama's first layer shows almost negative correlations with Task 2 eye-movements, while the second layer is the one with the highest correlations ($\mu = 0.4, \sigma = 0.009$), as shown by the purple

[2]Code is available in Github: https://github.com/maria-mouratidi/thesis

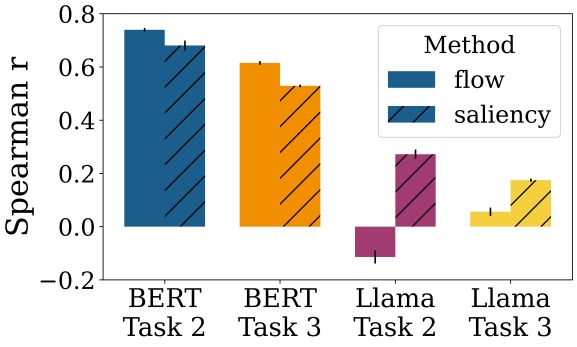

Figure 2: Average correlations across 6 eye-tracking features with each transformer's *attention flow* and *gradient-based saliency*.

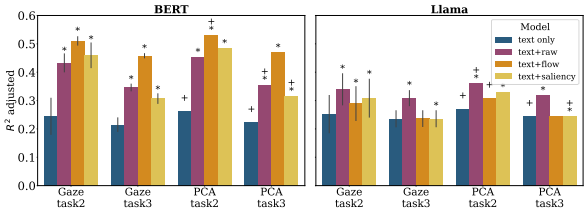

Figure 3: R^2 adjusted scores of the regression models, with or without attention as a feature. On the x-axis are the prediction targets and task conditions of each model. For the Gaze models, performance is averaged over each eye-tracking target. Asterisks indicate significant improvement over baseline, while crosses indicate significant improvement of PCA over the Gaze variant.

line in Figure 1. Like BERT, correlations decrease in subsequent layers, though with some upward trend toward the final layers. Task 3 correlations in Llama remain close to Task 2 correlations, and even exceed them in some layers (yellow line, Figure 1) showing that task-specific patterns are not affecting Llama as much as BERT.

Similarly to raw attention, attention flow and gradient-based saliency produce more moderate correlations with Task 2 and Task 3 eye-movements compared to BERT, with saliency ($r = 0.27, r = 0.18$) showing a clear advantage over flow ($r = -0.11, r = 0.05$) (purple and yellow bars, Figure 2). All correlations for the Llama model are statistically significant ($\alpha < 0.05$).

4.2.2 Predicting eye-movements

The regression models demonstrate clear benefits from incorporating transformer attention as a feature. For BERT (left panel, Figure 3), all attention methods significantly outperform the text-only baseline (blue bars). Moreover, attention flow contributes to the highest model fit for all tasks and DV conditions, reaching an overall $R^2 \approx 0.5$ (orange bars). This result is predictable from the higher correlations of attention flow with the eye-tracking features in Figure 2.

Llama (right panel, Figure 3) shows more modest performance than BERT. The models incorporating raw attention achieve R^2 values between 0.3 and 0.4 and both raw attention and saliency outperform the baseline across all conditions. As expected, attention flow does not significantly improve predictive capacity, except for the Task 2 Gaze condition. A clear pattern that emerges for both models is that Task 2 eye-movements are more predictable than Task 3 from all attention methods.

4.2.3 Reducing eye-tracking features

During the PCA exploration phase, we noticed that the SFD and GPT features accounted for most of the variance in a potential second PCA component. This is likely due to the nature of these features: SFD becomes zero when words receive multiple fixations, so it is often sparse, and GPT is likely more noisy as an intermediate feature between immediate processing (like FFD) and overall processing measures (like TRT). To maintain model simplicity, we removed SFD and GPT from the analysis entirely[3]. This reduction resulted in 94% explained variance in the first PCA component for Task 2 and 97% for Task 3. The simplified approach allows us to retain only one PCA component per task while preserving the most informative gaze patterns.

The cross annotations in Figure 3 demonstrate that all text-only PCA baselines show statistically significant improvement over their corresponding Gaze variants. This pattern extends to models using attention as well, where more than half outperform their Gaze counterparts in both BERT and Llama and both tasks. For the remaining PCA models that do not reach significance, the performance differences are minimal. This finding indicates that a single-PCA representation of the most important eye-tracking features can successfully replace the training procedure of multiple Gaze models.

4.2.4 Attention's role in prediction

To gain some perspective of the contribution of attention methods to predicting eye-movements, we examine the average feature importances over all

[3]This makes a total of 5 eye-tracking features included in the analysis. Whenever 6 features are mentioned, it means that the PCA component is also considered as a feature.

linear regression models. As seen in the upper panels of Figure 4, all attention methods for the BERT model receive the highest significance compared to the other text features. Attention flow demonstrates the greatest difference from other features, with only length and surprisal showing significant contributions. Llama models present a different

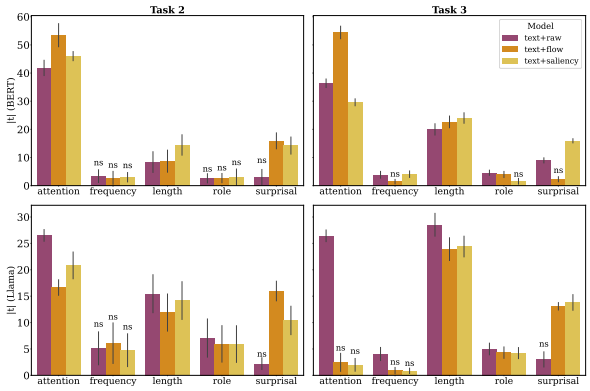

Figure 4: Feature importances based on mean absolute t-value across models predicting 6 eye-tracking features. "Ns" signifies non significant t-values ($p > 0.05$)

pattern, in the lower panels of Figure 4. Raw attention shows greater contribution than other attention methods but competes closely with word length in Task 3. Other attention methods maintain high contributions in Task 2 but become insignificant in Task 3. For non-raw attention methods in Llama, word length becomes a large contributor, followed by surprisal.

To further explore relationships between features that may influence their respective relative importance in the predictions, we examine correlations between attention methods and text features in Figure 5. BERT shows attention patterns that align with established eye-movement research. Higher frequency words receive smaller BERT attention values, while longer, content words draw more attention than short, function words. Notably, surprisal appears only slightly represented in BERT's attention mechanism. When it comes to Llama, similar correlation directions appear for raw attention, but in smaller magnitudes. Here, model attention using any method is more correlated to surprisal than any other text feature.

5 Discussion

We return to our research questions to briefly reiterate the key findings: Regarding correlation strength **(RQ1)**, decoder-only models like Llama show

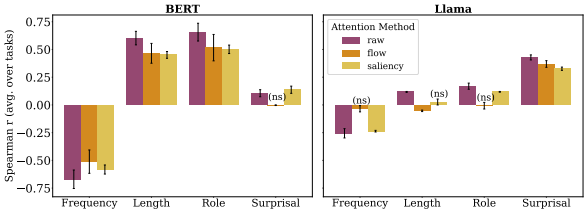

Figure 5: Correlations of text features with each attention method on a word-level.

medium-strength correlations with human eye-movements when using raw attention or gradient-based saliency. For **RQ2**, regression models combining Llama raw attention or gradient-based saliency with other text features achieve moderate performance in predicting human reading behavior. The regression models' predictive success relies heavily on attention, followed by word length and surprisal features. Concerning feature reduction and task effects **(RQ3)**, a single PCA component successfully replaces individual gaze targets while maintaining equivalent, or more favorable alignment results. Across all attention methods, both BERT and Llama show stronger alignment with Task 2 eye-movements than Task 3, suggesting that task-specific departures from normal reading are not equally well encoded in pretrained model attention mechanisms. The following sections provide deeper discussion of these findings and their implications.

5.1 What do layer-wise correlations imply?

The layer-wise raw attention correlations may hint at the type of linguistic processing that is most similar between humans and models. Early layers typically process surface-level features before higher-level semantic integration occurs, in both encoder (Tenney et al., 2019) and decoder models (Vig and Belinkov, 2019). Thus, the fact that higher correlations occur in early layers (first layer for BERT and second for Llama) suggests that better alignment can be found in bottom-up processes. This finding also has theoretical grounding in eye-movement research. The oculomotor control system can guide saccades before full lexical identification occurs (Rayner et al., 2011), and other early processing features like word length and frequency (Inhoff and Rayner, 1986) have robust independent effects on word skipping and fixation durations.

However, the correlation patterns are not monotonic across layers. Later layers show stronger correlations than middle layers. This suggests that

final layers may be more similar to the higher-level processing that also influences gaze behavior, for example when expectations about the text are formed using contextual information (Ehrlich and Rayner, 1981), or when the reader is more engaged in next-word prediction (Goldstein et al., 2022).

5.2 Why is Llama falling behind?

We consider two primary explanations for the alignment gap between the encoder and decoder model.

5.2.1 Pre-training objectives

First, Llama is a generative model optimized for next-token prediction, while BERT is trained in masked language modeling to capture bidirectional contextual representations. This fundamental difference in training objectives may explain why BERT aligns more with human reading behavior, which primarily involves comprehension rather than generation. Although the brain engages in next-word prediction during reading (Goldstein et al., 2022), the autoregressive nature of decoder models may not fully capture the integrative parts of human language comprehension that involve both forward and backward contextual dependencies. This difference is empirically supported by our feature correlation analysis, where Llama attention correlates most strongly with surprisal (a prediction-based feature) while BERT attention correlates mostly with the other text features (Figure 5).

5.2.2 The role of model size

This study's comparison between BERT-base (110M parameters) and Llama 3.1-8B (8B parameters) confounds architecture type with model scale. The substantial size difference may contribute to the observed alignment gaps, as larger models can distribute attention-relevant information across more parameters and layers. The specific model variants were chosen because of 1) BERT-base-uncased for replication and validation of previous work and 2) Llama 3.1-8B for the best performance-efficiency tradeoff among available decoder models. Nevertheless, future work should compare models of similar sizes to isolate architectural effects from scale effects.

5.3 Why is alignment with task-specific reading more difficult?

Task 3 eye-movements occur under fundamentally different reading conditions than Task 2, leading to smaller alignment with pretrained transformers. In the task-specific condition, readers form expectations about which words or syntactic structures might signal the specified relation, leading to more selective attention distribution among words (Hollenstein et al., 2020). When searching for specific words, reading resembles visual search, and certain text-level influences on eye-movements like word frequency disappear. (Rayner, 2009). Even when receiving explicit instructions, the models cannot replicate this type of selective attention. This limitation appears to be fundamental to current pretraining techniques rather than architecture-specific, as both encoder (BERT) and decoder (Llama) models show consistently better alignment with natural reading than task-specific reading. This suggests that neither masked language modeling nor autoregressive prediction objectives adequately prepare models for goal-directed attention strategies.

5.4 To each their own attention method

The results of different attention extraction methods vary significantly between the two models. Attention flow aligns best with eye-movements for BERT, while raw attention performs better for Llama. This may relate to the original motivation for attention flow, which was proposed as a way to represent attention in encoder models (Abnar and Zuidema, 2020). In decoder models, attention is restricted to preceding tokens, leading to an early token bias. When this effect is normalized as recommended by Abnar and Zuidema (2020), the attention signal may become more diluted. This highlights the need to carefully match the attention explanation method to both the model architecture and the explanation task.

5.5 One dimension for all eye-tracking features

Models using the PCA representation match or outperform Gaze models in correlation and regression analyses. This approach serves primarily as a methodological efficiency tool rather than aiming to increase predictive power. By capturing the shared variance across multiple eye-tracking features in a single component, PCA removes redundancy inherent in correlated features while preserving the essential reading patterns. So we were able to remove this practical challenge of determining which of the many available features are suitable for the comparison, without distorting them using averaging techniques. When PCA models

show similar alignment patterns to individual feature models, this suggests that much of the variance relevant to the comparison is captured by a common underlying dimension of reading behavior. This has implications for future eye-tracking studies, where researchers may be able to focus their analysis on this common dimension rather than examining all traditional eye-tracking measures individually, when the goal is understanding attention alignment with computational models.

6 Conclusion

This study examined the alignment between decoder-only models and human attention during reading. Overall, eye-movement data correlated with and was predictable from transformer attention, suggesting partial model alignment with human language processing. Early layers showed stronger alignment with eye-movements, hinting that bottom-up processes are more consistent with human reading behavior. However, lower alignment with task-specific reading suggests these pretrained models lack human-like flexibility to adapt attention based on task goals. Despite these shared patterns across architectures, the decoder model underperformed compared to the encoder model, showing lower correlations, weaker predictive power, and different patterns of feature prioritization, likely due to architectural differences. Finally, different methods of representing transformer attention significantly impact alignment comparisons, which emphasizes the importance of well-motivated, model and task-specific choices in explaining transformer mechanisms.

6.1 Limitations

This work assumes eye-movements provide sufficient information about cognitive language processing, though eye-tracking misses covert cognitive mechanisms and information processed outside the fixation region (Rayner, 2009; Reingold et al., 2016). Additionally, different attention explanation methods produce variable results, creating uncertainty about their faithfulness in explaining transformer attention mechanisms. Ultimately, any attention method is only a proxy to the true model representation.

Acknowledgments

We gratefully acknowledge the John S. Latsis Public Benefit Foundation for their generous support through the Postgraduate Scholarship Program.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197. Association for Computational Linguistics.

Maria Barrett and Nora Hollenstein. 2020. Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for natural language processing. *Language and Linguistics Compass*, 14(11):1–16.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155. Association for Computational Linguistics.

Joshua Bensemann, Alex Peng, Diana Benavides-Prado, Yang Chen, Neset Tan, Paul Michael Corballis, Patricia Riddle, and Michael Witbrock. 2022. Eye gaze and self-attention: How humans and transformers attend words in sentences. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–87. Association for Computational Linguistics.

Stephanie Brandl and Nora Hollenstein. 2022. Every word counts: A multilingual analysis of individual human alignment with model attention. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, volume 2, pages 72–77. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286. Association for Computational Linguistics.

Charles Clifton Jr., Adrian Staub, and Keith Rayner. 2007. Eye movements in reading words and sentences. In Roger P.G. Van Gompel, Martin H. Fischer, Wayne S. Murray, and Robin L. Hill, editors, *Eye Movements: A Window on Mind and Brain*, pages 341–371. Elsevier.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 4295–4309. Association for Computational Linguistics.

Jack Edmonds and Richard M. Karp. 1972. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of Association for Computing Machinery*, 19(2):248–264.

Susan F. Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655.

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.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A Norman, Orrin Devinsky, and Uri Hasson. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

Nora Hollenstein and Lisa Beinborn. 2021. Relative importance in sentence processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 2, pages 141–150. Association for Computational Linguistics.

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, Jonathan Rotsztejn, Marius Tröndle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 138–146, Marseille, France. European Language Resources Association.

Fariz Ikhwantri, Jan Wira Gotama Putra, Hiroaki Yamada, and Takenobu Tokunaga. 2023. Looking deep in the eyes: Investigating interpretation methods for neural models on reading tasks using human eye-movement behaviour. *Information Processing & Management*, 60(2):103195.

Albrecht W. Inhoff and Keith Rayner. 1986. Parafoveal word processing during eye fixations in reading: effects of word frequency. *Perception & Psychophysics*, 40(6):431–439.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 3543–3556. Association for Computational Linguistics.

Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Anastasia Kozlova, Albina Akhmetgareeva, Aigul Khanova, Semen Kudriavtsev, and Alena Fenogenova. 2024. Transformer attention vs human attention in anaphora resolution. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 109–122. Association for Computational Linguistics.

Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. 2022. Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64(12):3197–3234.

Niklas Metzger, Christopher Hahn, Julian Siber, Frederik Schmitt, and Bernd Finkbeiner. 2022. Attention flows for general transformers.

Felix Morger, Stephanie Brandl, Lisa Beinborn, and Nora Hollenstein. 2022. A cross-lingual comparison of human and model relative word importance. In *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, pages 11–23, Gothenburg, Sweden. Association for Computational Linguistics.

Keith Rayner. 2009. The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62(8):1457–1506.

Keith Rayner, Kathryn H. Chace, Timothy J. Slattery, and Jane Ashby. 2006. Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, 10(3):241–255.

Keith Rayner, Timothy J. Slattery, Denis Drieghe, and Simon P. Liversedge. 2011. Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of experimental psychology. Human perception and performance*, 37(2):514–528.

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.

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

Eyal M. Reingold, Heather Sheridan, K. L. Meadmore, Denis Drieghe, and S. P. Liversedge. 2016. Attention and eye-movement control in reading: The selective reading paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, 42(12):2003–2020.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 12–25. Association for Computational Linguistics.

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

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Kaiser. 2017. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30 of *NIPS'17*, pages 6000—6010. Curran Associates, Inc.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics.

Guojun Wu, Lena Bolliger, David Reich, and Lena Jäger. 2024. An eye opener regarding task-based text gradient saliency. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 255–263. Association for Computational Linguistics.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833. Springer International Publishing.

A French Eye-Tracking Corpus of Original and Simplified Medical, Clinical, and General Texts - FETA

Oksana Ivchenko **Natalia Grabar**
CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
oksana.ivchenko.etu@univ-lille.fr, natalia.grabar@univ-lille.fr

Abstract

Eye tracking offers an objective window on real-time cognitive processing of information being read: longer fixations, more regressions, and wider pupil dilation reliably index linguistic difficulty. Yet, there is a paucity of the available corpora annotated with eye-tracking features. We introduce in this paper the FETA corpus – a French Eye-TrAcking corpus [1]. It combines three types of texts (general, medical and clinical) in two versions (original and manually simplified). These texts are read by 46 participants, from which we collect eye-tracking data through dozens of eye-tracking features.

1 Introduction

Literacy, when reading general purpose and health-related information, depends critically on a reader's ability to understand such information (Eklics and Fekete, 2024; Brown, 2008). For instance, patients and the general public consult health-related sources – diagnosis leaflets, drug leaflets, web portals – on a daily basis (Fox, 2014), yet these materials are often written at a level well above the average reading proficiency (McCray, 2005). Text simplification (lexical, syntactic, or semantic) has therefore become a central strategy for improving accessibility (Saggion, 2017), but robust *evaluation* of simplification quality remains challenging (Grabar and Saggion, 2022). Eye tracking offers an objective window on real-time cognitive processing: longer fixations, more regressions, and wider pupil dilation reliably index linguistic difficulty (Singh et al., 2016). Employing gaze data to detect complex fragments can guide automatic or human adaptation of text, ultimately facilitating patient-oriented communication. Despite the maturity of eye-movement research in English (Hollenstein et al., 2022; Kuperman et al., 2020; Cop et al., 2017), French still lacks an openly available, large-scale corpus that combines (i) general-language and technical texts, (ii) technical medical and clinical texts, (iii) parallel simplified versions of texts, and (iv) fine-grained eye-tracking annotations.

To fill in this gap we introduce the FETA (French Eye-TrAcking) corpus designed through an eye-tracking experiment that captures reading behaviour across three text types (medical, clinical, general), each paired with manually produced lexical, syntactic, and semantic simplifications. Thus, our work makes several key contributions: it combines three types of texts (general, medical and clinical) in two versions (original and simplified), and gathers eye-tracking data from 46 participants.

In what follows, we describe the corpus texts (original documents and creation of their simplified versions) in Section 2. In Section 3, we describe the experimental protocol and participants. Section 4 is dedicated to the pre-processing of the eye-tracking data and extraction of eye-tracling features. Section 5 introduces the description of the eye-tracking-annotated corpus: metrics for the texts and eye-tracking features. Finally, we conclude in Section 6 and drow up some limitations in Section 7.

2 Corpus Construction

Our study employs a balanced, French-language corpus consisting of 16 texts sourced from two publicly available resources: the CLEAR corpus (Grabar and Cardon, 2018), corpus of clinical cases (Grabar et al., 2020), and general texts from Wikipedia. The set of 14 texts processed spans three text types: general-language articles from Wikipedia present common topics like *Weekend* or *Camelot*, medical-language articles from Wikipedia describe some specialized topics like *Vascular Cerebral Accidents* or *Obstetrics*, and clin-

[1] https://hdl.handle.net/11403/feta

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025,
pages 37–43, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-005

Original	Simplified
Les *hémocultures* ont permis d'isoler un *Staphylococcus aureus*. (Blood cultures made it possible to isolate a *Staphylococcus aureus*.)	Les *hémocultures* (analyses des bactéries éventuelles dans le sang) ont montré la présence de la bactérie *Staphylococcus aureus*. (Blood cultures (tests for possible bacteria in the blood) showed the presence of the *Staphylococcus aureus* bacterium.)
Un cathéter a été posé. (A catheter was inserted.)	Un cathéter a été posé *pour évacuer l'urine.* (A catheter was inserted *to drain the urine.*)

Table 1: Examples of manual simplification presented in the inline original format, with translations.

ical cases from toxicology and gastrology. Clinical cases describe symptoms, diagnoses, treatments, and follow-ups for individual patients or small cohorts. Their narrative structure resembles hospital discharge summaries and is densely packed with specialised terminology and reasoning about therapeutic choices. Such texts impose a high cognitive load on lay readers who must comprehend health information relevant to themselves or their relatives. Original clinical texts contain 653 words, general texts contain 1,684 words, and medical texts 2,906 words. A detailed breakdown by screen and sentence is provided in Table 4 (Appendix).

To facilitate controlled eye-tracking experiments, we partitioned the 14 texts into two equally balanced *sets*, *Set 1* and *Set 2*, each containing a uniform mix of medical articles, clinical cases, and general texts, thereby equalising topic distribution and baseline difficulty across sets. Each text has been manually simplified as explained in Section 2.1 and examplified in Table 1. Then, we compose two presentation *versions*:

- *Version A*: half the texts appear in their original form, the remainder in simplified form.
- *Version B*: the original/simplified assignment is reversed, creating a mirror of Version A.

Random assignment to Set 1 or 2 gave each participant one version per text, preserving counterbalancing and single exposure

The primary aim of the experiment is to record eye-tracking indicators during natural text reading. To complement these gaze data with a behavioural measure of comprehension, we administer short multiple-choice questions after selected text segments. Each question pertains to the segment that has just been read, and participants respond by choosing *True*, *False*, or *I don't know*. To keep the reading experience as natural as possible and to minimise task interruption, comprehension questions are presented for only a random subset of segments.

2.1 Simplification Pipeline

All texts were manually simplified in respect with the plain-language recommendations (OCDE, 2015) at syntactic, lexical and semantic levels, as examplified in Table 1.

Syntactic level. Syntactic simplification aimed to reduce structural complexity by transforming embedded and multi-clause constructions into shorter, clause-minimal units. Where possible, passive constructions were rewritten in the active voice to make sentence roles (agent, action, patient) more explicit.

In addition, we prioritized the use of direct (subject–verb–object) word order to avoid ambiguity, clarified negations, and systematically avoided gerunds and past participial forms. As a result, simplified versions contain more sentences than their originals.

Lexical level. Lexical simplification involved replacing domain-specific or low-frequency terms with more accessible alternatives to improve comprehensibility. Different strategies were applied: **i)** high-frequency synonyms were used when semantic precision could be maintained, **ii)** hypernyms, or more general terms, were substituted for complex medical terminology, and **iii)** in-text definitions were inserted in parentheses directly after specialised terms to support interpretation. These strategies aimed to retain the intended meaning while lowering lexical complexity for readers without specialised knowledge. In many cases, the original term was kept alongside its explanation to aid familiarity and consistency.

Semantic level. Semantic simplification focused on enriching the text with contextual information to make implicit knowledge more explicit. This was especially important in clinical texts, where technical discourse often assumes prior medical knowledge that general readers may not possess. The goal was to reduce inferential effort by clarifying relationships, causes, effects, and by defining specialized concepts in context. Several semantic strategies were applied: **i)** causal or descriptive links were added to explain the function or consequence of a condition, like in the third example in Table 1, in which the role of catheter is explained; **ii)** integrated paraphrases combined description and terminology to bridge gaps in understanding. These modifications clarify the meaning of complex medical expressions and also anchor them in relatable concepts.

Overall, the simplified corpus exhibits (i) more sentences through syntactic segmentation, (ii) sometimes more lexemes through lexical substitutions, and (iii) richer contextual clues through semantic elaboration. Hence, the material remains fair to the original meaning but is cognitively easier for non-specialist readers.

3 Experimental Protocol and Participants

3.1 Experimental Protocol

The protocol is composed of several steps:

Pre-screening (online). Prior to scheduling, each participant completed a form collecting demographic data (age, gender, highest education), ocular health information (e.g. myopia, astigmatism, corrective lenses), reading habits, and informed-consent details about the study.

Day-of self-evaluation. On arrival, participants filled out a two-pages, self-assessment questionnaire on the perceived difficulty of understanding medical information in daily life, using a four-point Likert scale: *very easy/easy/difficult/very difficult*.

Set-up and calibration. Gaze was recorded with a Tobii Pro Spectrum eye tracker sampling at 600 Hz. Text stimuli were presented on a 24-inch monitor at a native resolution of 2880 × 1620 px; Participants were seated 60 cm from the display (adjusted by $\pm$ 5 cm to accommodate height and optimise calibration), as on Figure 1. A random five-point calibration was accepted when accuracy and precision thresholds of $0.5°$ and $0.2°$, respectively, were met. Calibration quality was manually inspected, and participants who marginally exceeded these thresholds were retained if visual inspection confirmed stable gaze traces.

Figure 1: Experiment set-up and calibration.

Familiarisation block contained several slides: *Slide 1*: introductory instructions; *Slide 2*: a short, easy text common to all participants; *Slides 3–4*: two comprehension questions on this text, answered aloud (*True/False/I don't know*).

Main reading block. Each participant was randomly assigned to *Set 1* or *Set 2* and to *Version A* or *Version B*. The block comprised in average 59 slides: original or simplified texts according to the set–version counter-balancing scheme. Slides advanced via a mouse click. All comprehension questions were to be answered loudly.

Mid-session break. After the first timeline (half of the slides), participants took a short pause. A second five-point recalibration followed the break.

Second timeline and debrief. The remaining slides were presented, after which participants answered some oral questions on perceived text difficulty and comprehension ease. The entire experiment lasted about 50 to 70 minutes, depending on how the participant read.

3.2 Participant Demographics

Forty-six native French speakers (32 women, 14 men; age range 18–43 years, $M = 23.3$, $SD = 6.7$) took part in the study. Their educational backgrounds were diverse but none held a medical or healthcare qualification. All reported normal or corrected-to-normal vision. Participants received an honorarium of €12. To balance exposure to text conditions, they were randomly assigned to four counter-balanced groups: Set 1-A (n=12), Set 1-B (n=11), Set 2-A (n=11), and Set 2-B (n=12). More detailed information is provided in the Table 3.

4 Pre-processing and Annotation

Text–Gaze Alignment. Text presentation and word-level AOIs (one per word) were handled automatically in Tobii Pro Lab. Fixations were matched to those AOIs within the software, eliminating custom tokenization or manual ID assignment.

Feature Extraction and Normalisation. Eye-movement events were classified in Tobii Pro Lab using the I-VT (Velocity-Threshold Identification) algorithm with the following settings:
Eye selection: average of both eyes.
Noise reduction: moving median (window = 3 samples).
Velocity calculator: window length = 20 ms.
I-VT threshold: 30 deg/s.
Fixation merging: max. gap = 75 ms; max. angle = 0.5°.
Discard short fixations: min. fix. duration = 60 ms.

5 Dataset Statistics

5.1 Text Metrics

Table 4 in the Appendix compares original and simplified versions for each text (number of screens, sentences, and words). We can see that, across the corpus, simplification primarily increased the number of sentences due to syntactic simplification, with more modest changes in word counts.

By domain, clinical texts rose from 32 to 42 sentences (+31.3%) and from 653 to 805 words (+23.3%). General texts showed a strong sentence increase (73 → 107, +46.6%) but virtually no change in word count (1 684 → 1 691, +0.4%), reflecting many short sentence splits without added explanations. Medical texts increased from 144 to 179 sentences (+24.3%) and from 2,906 to 3,081 words (+6.0%).

5.2 Gaze Metrics

For every word–participant pair we release ten eye-movement features. The full list of features is in Appendix 8.1.

Figure 2 shows the difference in reading original and simplified clinical text for the *Total_duration_of_fixations* feature.

Table 2 reports the median and inter-quartile range (IQR) of each feature, aggregated by domain (clinical, medical, general) and by version (original, simplified). The headline metric, *Total Fixation Duration* (TFD), shows a clear reduction in reading effort (Figure 3): in clinical texts, the median TFD drops –17%; in medical texts –6%; in general texts –3%.

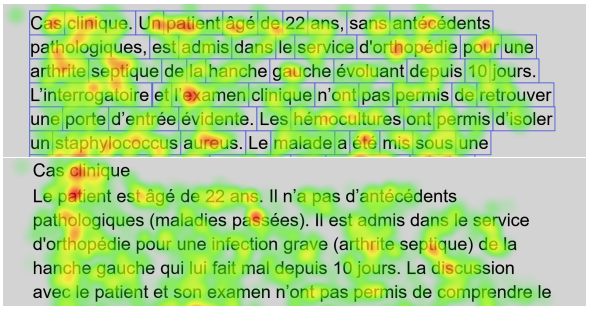

Figure 2: Original clinical case (top) and Simplified clinical case (bottom)

Table 2: Median and inter-quartile range (IQR) of word-level total fixation duration (in milliseconds).

Domain	Version	Median (ms)	IQR (ms)
Clinical	Original	225	435
	Simplified	187	342
Medical	Original	203	375
	Simplified	193	357
General	Original	183	333
	Simplified	182	323

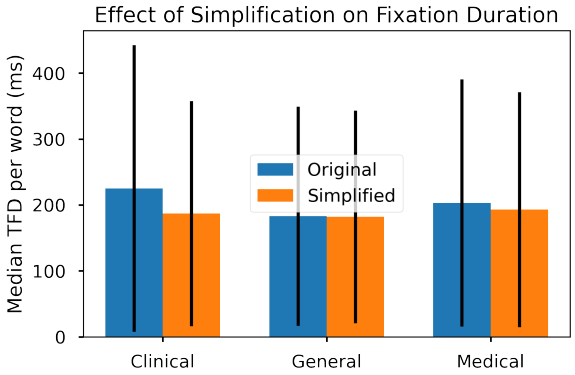

Figure 3: Median word-level fixation duration by domain and version.

6 Conclusion and Future Work

We introduced the FETA (French Eye-TrAcking) corpus, built with general-language and health documents in two versions (technical and manually simplified), thus covering several topics and genres. This corpus was read by 46 participants, through a precise experimental protocol. This permitted to collect several eye-tracking features, of which 10 are provided as part of the FETA corpus.

Besides, eye-tracking data are also being collected from speech-language pathology students,

which will permit to compare the reading from non-specialised and specised participants.

7 Limitations

Although all recordings met our calibration criteria, occasional attentional shifts or transient tracker losses may have gone undetected. Consequently, some fixations – especially at line breaks – could be mis-assigned, lowering word-level accuracy.

To preserve natural reading, only eleven multiple-choice questions were randomly inserted across the 50 slides. This design prevents us from verifying comprehension on every individual slide, which means that local misunderstandings might therefore remain unnoticed.

Due to Tobii Pro Lab's limitations in processing large datasets, raw data export proved challenging. We will include additional eye-tracking features and raw data as data processing continues.

8 Ethical Considerations

Participation in this study is voluntary, with informed consent obtained from all participants, ensuring compliance with the European General Data Protection Regulation (EU) 2016/679 and the modified French Data Protection Act of January 6, 1978. All personal data collected in the course of this research are anonymized to protect participant privacy and are accessible only by the designated project manager. This study has been registered in the University of Lille registry under reference 2022-075, affirming our commitment to upholding the highest standards of data protection and participant rights.

Acknowledgement

This work was partially funded by the French National Agency for Research (ANR) as part of the CLEAR project (Communication, Literacy, Education, Accessibility, Readability), ANR-17-CE19-0016-01.

References

Jo Brown. 2008. How clinical communication has become a core part of medical education in the uk. *Medical Education*, 42(3):271–278.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

Kata Eklics and Judit Fekete. 2024. From a simulated patient interview to a case presentation.

Susannah Fox. 2014. The social life of health information. Technical report, Pew Internet & American Life Project, Washington DC.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Natalia Grabar, Clément Dalloux, and Vincent Claveau. 2020. CAS: corpus of clinical cases in French. *Journal of BioMedical Semantics*, 11(1):1–7.

Natalia Grabar and Horacio Saggion. 2022. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *TALN-RECITAL 2022*, pages 453–463, Avignon, France.

Nora Hollenstein, Maria Barrett, and Marina Björnsdóttir. 2022. The copenhagen corpus of eye tracking recordings from natural reading of Danish texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1712–1720, Marseille, France. European Language Resources Association.

Victor Kuperman, Noam Siegelman, Sascha Schroeder, Alina Alexeeva, Cengiz Acarturk, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2020. Text reading in English as a second language: Evidence from the multilingual eye-movements corpus (MECO). *Studies in Second Language Acquisition*.

A McCray. 2005. Promoting health literacy. *J of Am Med Infor Ass*, 12:152–163.

OCDE. 2015. *Guide de style de l'OCDE Troisième édition: Troisième édition*. OECD Publishing.

Horacio Saggion. 2017. *Automatic Text Simplification*, volume 32 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool, University of Toronto.

Abhinav Deep Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. 2016. Quantifying sentence complexity based on eye-tracking measures. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 202–212, Osaka, Japan. The COLING 2016 Organizing Committee.

8.1 Appendices

Table 3: Participant overview by experimental group.

Set	*n*	F/M	Age	BA	MA	PhD	Emp
1A	12	9 / 3	19–42	9	2	2	5
1B	11	8 / 3	18–35	8	2	1	4
2A	11	7 / 4	18–43	9	1	1	4
2B	12	8 / 4	18–36	8	3	1	2
Total	**46**	**32 / 14**	18–43	34	8	5	15

BA = Bachelor's (Licence); MA = Master's; PhD = Doctorate; Emp = "Employed" counts anyone working (including student+worker). Levels are those held at the time of the experiment.

Features provided.
Duration_of_first_fixation: time (ms) of the first fixation on a word.
First-pass_duration: cumulative fixation time from first entering the word until leaving it to the right.
First-pass_first_fixation_duration: first-fixation duration restricted to the first-pass window.
First-pass_regression: binary flag (1 = gaze exits the word to the left during first pass).
Maximum_duration_of_fixations / *Minimum_duration_of_fixations*: longest and shortest single fixations on the word.
Number_of_fixations: count of fixations on the word.
Re-reading_duration: fixation time accumulated after the first pass.
Regression-path_duration: time from first entering the word until leaving it to the right *after* any regressions.
Total_duration_of_fixations: sum of all fixation durations on the word (early + late).

8.2 Appendices

Table 4: Comparison of Original and Simplified Texts

Text type	Text name	Version	Screens	Sentences	Tokens
Clinical	gastro	original	3	17	285
		simplified	3	19	323
	obGyn	original	3	11	249
		simplified	3	24	307
	toxico	original	4	19	398
		simplified	5	29	469
	gastro	original	3	13	255
		simplified	3	25	336
General	camelot	original	8	42	840
		simplified	8	58	880
	quince	original	7	44	751
		simplified	7	54	785
	popcorn	original	9	44	865
		simplified	7	51	752
	weekend	original	9	31	843
		simplified	9	49	811
Medical	autopsy	original	10	39	943
		simplified	9	65	925
	stroke	original	3	10	276
		simplified	3	22	328
	chikungunya	original	21	102	1983
		simplified	21	138	1975
	erytheme	original	7	34	653
		simplified	9	58	960
	obstetrics	original	12	57	1104
		simplified	12	65	1202
	ulcer	original	15	77	1526
		simplified	15	92	1551

Exploring Mouse Tracking for Reading on Romanian Data

Maria Cristina Popescu, Sergiu Nisioi
Human Language Technologies Research Center
Faculty of Mathematics and Computer Science
University of Bucharest
popescu.cris.cristina@gmail.com
sergiu.nisioi@unibuc.ro

Abstract

In this paper, we investigate the use of the Mouse Tracking for Reading (MoTR) method for a sample of Romanian texts. MoTR is a novel measurement tool that is meant to collect word-by-word reading times. In a typical MoTR trial, the text is blurred, except for a small area around the mouse pointer and the participants must move the mouse to reveal and read the text. In the current experiment, participants read such texts and afterwords answered comprehension questions, aiming to evaluate reading behavior and cognitive engagement. Mouse movement is recorded and analyzed to evaluate attention distribution across a sentence, providing insights into incremental language processing. Based on all the information gathered, the study confirms the feasibility of this method in a controlled setting and emphasizes MoTR's potential as an accessible and naturalistic approach for studying text comprehension.

1 Introduction

Language understanding is one of the most complex human cognitive activities. Whether reading or listening, the human brain processes linguistic input incrementally, integrating each word as it is encountered. This is known as incremental language processing and is characterized by both sequentiality and variability: some words are processed quickly, others require more cognitive effort due to low predictability, frequency, or syntactic complexity (Smith and Levy, 2013).

One of the main goals of psycholinguistics is to measure this incremental effort in real time. Early work in the 1970s introduced the *gaze-contingent moving window paradigm*, which involved making display changes in the text based on eye position as participants were reading, and then examining how these changes influenced eye movement behavior (McConkie and Rayner, 1975). Early studies demonstrated how parafoveal and foveal vision interact during reading and established the methodological foundation for modern incremental processing research. Eye-tracking provides the most precise measurement but is expensive and requires specialized equipment.

To address these limitations, alternative paradigms have been proposed. Self-paced reading (SPR) (Just et al., 1982) and the Maze task (Boyce et al., 2020; Forster et al., 2009) can be deployed online at low cost, but they involve linear reading and artificial constraints, affecting the interaction with the text.

In this paper we investigate the usage of Mouse Tracking for Reading (MoTR) method (Wilcox et al., 2024) on Romanian texts. The method was designed to balance the high accuracy and naturalness of eye-tracking with the low cost and online availability of other "self-paced" incremental measurements, such as SPR and Maze (Wilcox et al., 2024). In our MoTR experiment, participants are presented with several texts that are blurred, except for a small area around the mouse, which is clear. They have to move the mouse to reveal and read the text, and its position is recorded for post-processing (see Figure 1). After completing the reading of each sentence, the participants answer a comprehension question related to the text read, to validate the quality of the data and confirm the cognitive engagement of the candidate. Until now, published studies using MoTR have been conducted on English texts, often relying on corpora such as the Provo Corpus (Luke and Christianson, 2018; Wilcox et al., 2024), with relatively small documented applications in other languages (Schneider et al., 2021; Haveriku et al., 2025; Oğuz et al., 2025).

In this context, the present work addresses an important gap, being the first to test MoTR on Romanian texts.

The texts used in our experiment are Romanian

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025,
pages 44–51, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-006

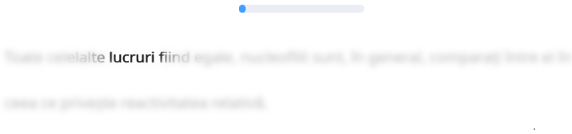

Figure 1: Example of the blurred interface used in the MoTR experiment. The only word visible in this image is “lucruri” (English *things*)

.

	Sentence len		Complexity	
	Eng.	Rom.	Eng.	Rom.
mean	22.82	24.46	0.24	0.13
std	7.74	8.62	0.19	0.25
min	7	9	0.02	0
max	45	49	0.93	1
samples	569		569	
sentences	158		158	

Table 1: Statistics comparing the English and Romanian Human Translation sentences (and 569 lexical complexity annotated samples). Romanian complexity annotations have a higher variance and a lower average complexity than the original English counterparts.

versions of the Multilingual Lexical Simplification Pipeline (MLSP) Shared Task 2024 competition dataset (Ghaddar et al., 2024a) The Romanian part has been created by manually translating the original English data such that each sentence contains similar lexical complexity annotations to the originals (Anghel et al., 2025). The data set is made so that it can be easily integrated into MultiLS (Ghaddar et al., 2024b), a recently developed framework for lexical analysis in multiple languages, providing a standardized context for comparative studies of text complexity and comprehension. Statistics regarding sentence length and annotated complexity scores are visible in Table 1.

An important advantage of this corpus is that certain words in each sentence are annotated with explicit human judgments of complexity scores assigned by five young adults. The complexity scores reflect the estimated difficulty of each word in its context. With this information, we analyze the relationship between the linguistic complexity of words and reading time, capturing the relation between perceived lexical difficulty and linguistic processing.

At the same time, we extend the existing experimental infrastructure by developing a complete pipeline in Romanian: from corpus preparation, to their integration into Magpie framework and the generation of comprehension questions. Overall, the contribution of the paper consists both in the methodological adaptation of MoTR for the Romanian language, and in demonstrating its applicability in the analysis of lexical complexity and reading times in an experimental setting.

2 Methodology

2.1 A MoTR Trial

In each trial of this experiment, participants are exposed to a web interface containing blurred text (see Figure 1), except for a small clear area around the tip of the mouse cursor. Each participant is instructed to move the mouse to reveal the text word by word, thus allowing sequential reading of the text. After the participant confirms the completion of reading a text by pressing a button, they are presented with a question with a “yes” or “no” answer, regarding the sentence read and intended to assess comprehension of the read content. At that moment, the entire text is blurred, no longer visible. Participants can move to the next screen only after answering the question.

Cursor movements are recorded throughout the reading, except for the moment when the participant answers the comprehension question. The cursor coordinates are subsequently analyzed as a proxy for gaze direction, effectively simulating the behavior of an eye-tracking system.

The experiment is implemented in Magpie[1], a web platform designed to conduct behavioral experiments directly in the browser. It allows for real-time transmission of cursor coordinates and task flow management.

2.2 Participants and Data Collection

Five native Romanian speakers (3F, 2M), 22-30 years old, agreed to participate in the experiment voluntarily. All participants are native Romanian speakers and have at least completed high school. None of them have diagnosed visual impairments. The study was conducted in a restricted and controlled environment, each session (approximately 2 hours per session) was directly monitored, to ensure that the rules and instructions were respected.

The data collected included:

- Mouse coordinates and timestamps

[1]Magpie is framework for building psychological online experiments that run in the participants’ browser: https://magpie-experiments.org/

- Word indices and reveal times
- Comprehension question responses
- Total reading duration per trial

To ensure that the MoTR method closely approximates the reader's visual attention, several parameters are calibrated:

- Spotlight size: 102 pixels - large enough to disambiguate word focus, but small enough to prevent excessive or fatiguing mouse movements;
- Gradual blur transition - simulating the shift from foveal to peripheral vision;
- Line spacing: 55px to avoid vertical interference;
- Cursor sampling rate: 20Hz -balancing temporal precision with transmission stability.

We make a small adaptation in terms of line spacing from the configurations proposed by Wilcox et al. (2024) so that users are less prone to accidentally move the mouse on the lines below or above the current reading areas.

3 Results

Our first objective is to provide an overview of how participants use the MoTR method and the variability that arises between individuals, items, and trials. At this stage, we focus on analyzing data from a single participant, selected due to their representative behavior. This case serves as an illustrative example of typical MoTR usage and provides a clear foundation for interpreting results in the broader analysis.

Total Reading Time (TRT) is used as our main measure of processing effort. It captures the full time spent on a word, including all refixations, and is widely used as an indicator of deep syntactic and semantic processing (Just and Carpenter, 1980; Rayner, 1998).

To ensure cognitive engagement and data quality, each sentence in the experiment is followed by a yes/no comprehension question. Participants show high accuracy, with individual scores ranging between 81% and 92%, and a group mean of approximately 88.5%.

This high level of accuracy confirms that participants have a high degree of comprehension, making the reading-time data more reliable. These results suggest that the MoTR interface supports natural reading and allows for meaningful variation in comprehension to be captured.

3.1 Correlation Analysis

We compute Pearson correlation coefficients to assess the linear relationship between total reading time and two basic lexical predictors:

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}} \quad (1)$$

where:

- x, y are numerical vectors of equal length n,
- m_x, m_y are the means of vectors x and y respectively.

Variable	Pearson *r*
Frequency Score	-0.53
Word Length	+0.58

Table 2: Pearson coefficients between *totalReadingTime* and lexical predictors. Frequency shows a moderate negative correlation with reading time, while word length shows a positive correlation. Frequency scores are obtained using the `wordfreq` library (Speer, 2022), which compiles word frequencies from diverse sources including Wikipedia, news, subtitles, and web data.

Pearson correlation analysis, as shown in Table 2, reveals a negative relationship between word frequency and reading time, and a positive one between word length and reading time. These findings align with well-established psycholinguistic assumptions: frequent words are processed more quickly, while longer words require more cognitive effort (Smith and Levy, 2013).

3.2 Regression Analysis

To estimate the influence of lexical and orthographic features on reading time, we use a regression model using *Support Vector Regression* (SVR), predicting continuous values by fitting a function within a margin of tolerance (Awad and Khanna, 2015). Although we employ a linear kernel, we opt for SVR instead of classic linear or Ridge regression due to its robustness in handling outliers and its ability to ignore small errors via the ϵ-insensitive loss function.

The model is implemented using the `SVR` module from the `scikit-learn` library (Pedregosa

et al., 2011). The penalty parameter C is chosen via cross-validation and set to 100. Although this is a relatively large value, it consistently yields optimal predictive performance at cross-validation.

The SVR model predicts reading time as a linear combination of four features: frequency score, word length, syllable count, and the presence of diacritics. All features included in the model capture linguistic properties that influence processing difficulty. Frequency, word length, and syllable count are widely recognized as key factors influencing reading time (Rayner, 1998). We also include diacritics because omitting an accent can momentarily break the visual rhythm of a sentence and add a small cognitive load during speed reading (Marcet and Perea, 2022). Including diacritics in the model helps capture a subtle but systematic aspect of Romanian orthography that can influence reading behavior.

Each predictor is weighted by a learned coefficient β_i, and the model includes an intercept term β_0. Formally, the model takes the form: *totalReadingTime* $\approx \beta_1 \cdot$ FrequencyScore + $\beta_2 \cdot$ WordLength + $\beta_3 \cdot$ Syllables + $\beta_4 \cdot$ HasDiacritics + β_0.

Coefficient	Value
β_0	397.06
β_1	-34.39
β_2	186.38
β_3	-21.93
β_4	-9.58

Table 3: Estimated coefficients of the SVR model.

The intercept β_0 is the baseline reading time. β_1 shows that frequent words are read faster, while β_2 indicates that longer words take more time. β_3 and β_4 reflect smaller negative effects from syllable count and diacritics.

Model Performance

The SVR model is evaluated using 10-fold cross-validation, with the data split so that no sentences appears in both training and test sets. The model achieves a root mean square error (RMSE) of approximately **238.92 ms**. The coefficient of determination (R^2) is **0.37**, indicating that around 37% of the variance in reading times is explained by the model. The Pearson correlation between predicted and actual values is $r = 0.635$ ($p < 0.001$), suggesting a moderate and statistically significant fit.

Metric	Value
Coefficients	$[-34.39, 186.38, -21.93, -9.58]$
Intercept	397.06
RMSE (mean, CV)	238.92 ms
R^2 Score	0.37
Pearson r	0.635 ($p < 0.001$)
Accuracy	87.24%

Table 4: Performance of the SVR model with a linear kernel and four predictors.

In addition to classic error metrics (RMSE, R^2), we evaluate model performance using the accuracy metric defined in (Hollenstein et al., 2022), where real and predicted values are scaled to [0, 100], and accuracy is defined as:

$$\text{Accuracy} = 100 - \text{MAE}$$

where MAE (Mean Absolute Error) represents the average absolute difference between predicted and actual values.

Our model achieves an accuracy score of **87.24%**, confirming a very good match between predicted and normalized real reading times.

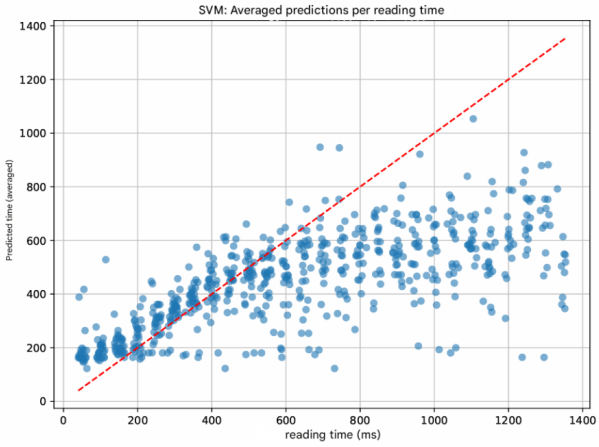

Figure 2: SVR predictions versus reading time. We observe a good alignment between predictions and actual values, with reasonable dispersion around the line $y = x$. We can observe a logarithmic tendency of reading times.

Figure 2 indicates a reasonable alignment between the predicted and actual values, with no obvious systematic deviations. The scatter around the identity line suggests natural variation in reading behavior.

The SVR model provides a flexible estimation of the relationship between word features and reading

time, showing strong predictive performance in a cognitive-linguistic context (Li and Rudzicz, 2021).

3.3 Language Model Log-Probability

Linguistic surprisal is a computational measure of how unpredictable a word is in its context, and implicitly, how cognitively demanding it is to process (Smith and Levy, 2013; Hale, 2001). According to information theory, surprisal is defined as the negative logarithm of the conditional probability:

$$\text{Surprisal}(w_i) = -\log_2 P(w_i \mid w_1, w_2, ..., w_{i-1})$$

This equation reflects the idea that a highly expected word (high probability) requires less cognitive effort to process. In contrast, an unexpected word with low probability lead to higher surprisal values that typically requires longer reading time (Levy, 2008; Smith and Levy, 2013).

In masked language models such as BERT, surprisal is not based solely on the preceding context but instead uses the entire sentence. As such, we use the log-probability from the model as a proxy for the surprisal of a target word w_i.

Log-probability is calculated using the model *dumitrescustefan/bert-base-romanian-cased-v1*, a BERT-base model pre-trained on diverse Romanian data sources (Wikipedia, OSCAR, etc.) and adapted for masked language modeling tasks (Dumitrescu et al., 2020). The score is computed for each word in the sentences by masking it and retrieving the model's conditional probability. When a word is split into multiple subtokens during tokenization, we mask all subtokens simultaneously and compute the model's joint probability for the full word.

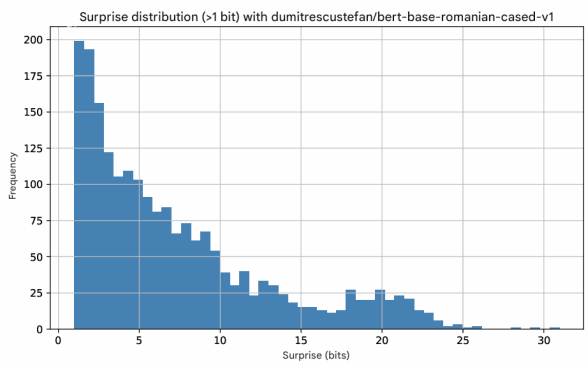

Figure 3: Distribution of log-probability values (> 1 bit) estimated using *bert-base-romanian-cased*. The distribution is right-skewed, with relatively few words showing high logprob.

For better visualization, we exclude very low surprisal values (below 1 bit), which dominated the frequency range and obscured the structure of more informative intervals.

To investigate the influence of surprisal on real-time processing, we analyze the relationship between estimated surprisal and total reading time per word (*totalReadingTime*). The computed Pearson correlation coefficients shows the following significant relationships:

Variable	**Pearson r**
Surprisal vs. Reading Time	+0.361
Frequency vs. Reading Time	−0.540
Word Length vs. Reading Time	+0.581

Table 5: Pearson correlation coefficients between predictors and reading time. Surprisal and word length correlate positively with reading time, while frequency correlates negatively.

These findings confirm that:

- Surprising words are associated with longer reading times;
- Frequent words are processed more quickly;
- Longer words tend to require more time to read.

To evaluate these predictors together, we fit a multiple linear regression model with surprisal, frequency, and word length as features. The estimated model predicts reading time as a linear combination of these three predictors. Each feature is multiplied by a learned coefficient (β_1, β_2, β_3), and the model includes a constant term β_0. Formally, the model takes the form:

ReadingTime $\approx \beta_1 \cdot$ Surprisal + $\beta_2 \cdot$ Frequency + $\beta_3 \cdot$ WordLength + β_0.

The model is statistically significant ($F(3, 3890) = 853.5, p < 0.001$) and explains approximately 39.7% of the variance in reading times ($R^2 = 0.397$). The estimated coefficients are:

- $\beta_1 = +16.35$ (each additional bit of surprisal increases reading time by 16 ms),
- $\beta_2 = -45.71$ (higher word frequency reduces reading time),
- $\beta_3 = +57.30$ (each additional character increases reading time).

These results support the hypothesis that surprisal, frequency, and length contribute systematically to the cognitive effort involved in lexical processing (Levy, 2008).

3.4 Lexical Complexity and Reading Times

In addition to computationally derived predictors (surprisal, frequency, and word length), we also evaluate the relationship between a manually annotated measure of lexical complexity (*ht_complexity*) and total reading time. The *ht_complexity* values reflects human judgments of how difficult each word is to understand in its context, with higher values indicating greater perceived difficulty. The analysis is implemented by aligning annotated tokens with reading times from the dataset derived through manual translation and revision of the MLSP Shared Task 2024 corpus, as detailed in (Cristea and Nisioi, 2024).

The results indicate a significant positive correlation between lexical complexity and reading time (**r = 0.402**, $p < 0.0001$), suggesting that more complex units tend to require longer processing times.

This finding supports the hypothesis that lexical complexity directly impacts cognitive effort during reading, consistent with earlier work on linguistic processing and comprehension (Just and Carpenter, 1980).

3.5 Interindividual Variation in Reading Behavior

To evaluate the consistency of relationships between linguistic features and reading time, we extend our analysis to all five participants. This broader view provides more detailed insight into lexical effects and allows us to observe interindividual variability in language processing.

The average reading time (*totalReadingTime*) varies considerably across participants, with means ranging from approximately 435 to 604 milliseconds (Table 6).

In addition to the average reading times, the observed variability within each participant reflects clear differences in central tendency and dispersion. These results point to individual differences in reading styles and the stability of reading behavior (Just and Carpenter, 1980; Rayner, 1998).

We run separate regressions for each participant using word frequency and length as predictors. All show the same direction of effects—frequent words are read faster, longer words slower—despite variation in strength. This confirms that core lexical effects remain consistent across readers.

Participant	Mean	S.D.	Min	Max
P1	596.27	647.72	36.0	5751.0
P2	435.85	492.61	39.0	9011.0
P3	488.18	488.49	34.0	6349.0
P4	532.81	461.76	40.0	5400.0
P5	603.83	529.04	38.0	8005.0

Table 6: Descriptive statistics of reading times for each participant, including mean, standard deviation, minimum, and maximum values (all in milliseconds). Substantial differences can be observed across participants, both in mean and dispersion, suggesting variable reading styles.

4 BERT-based Predictor

We use the *bert-base-romanian-cased-v1* model (Dumitrescu et al., 2020), a pretrained version on large Romanian corpora that preserves the standard BERT architecture. The contextual embeddings generated by this encoder are integrated into a regression model, in order to predict the total reading time of a word based on the full sentence in which it appears.

Applying a logarithmic transformation (as suggested by the results in Figure 2) to the target value significantly improves model performance. This pre-processing step stabilizes the reading time distribution, reduces the influence of outliers, and allows the model to learn more robust relationships between contextual embeddings and cognitive reading difficulty.

We evaluate the final model on a test set of 773 examples, yielding the following metrics:

- Pearson correlation coefficient: **0.76**
- Spearman correlation coefficient: **0.78**
- Mean Absolute Error (MAE): **0.41** (in log space)
- Coefficient of determination R^2: **0.56**

These results confirm that large langauge models encode strong features for predicting reading times. The contextual embedding of the target token, combined with additional linguistic features and a log-transformed target, leads to accurate reading time predictions. Figure 4 shows a clear alignment between predicted and actual reading times, reflecting

the model's strong predictive performance and robustness.

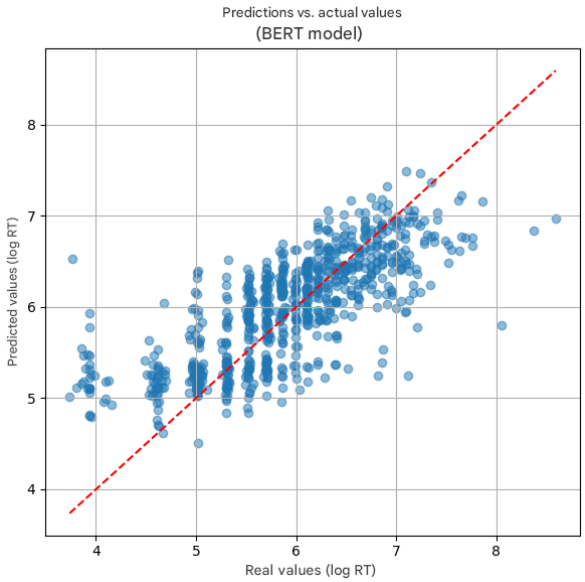

Figure 4: Predicted vs. actual reading times. The strong alignment along the diagonal suggests that the BERT-based model accurately predicts reading times from contextual embeddings.

5 Conclusions

This study shows that the *Mouse Tracking for Reading* (MoTR) method can be a practical and effective way to study how people read and process Romanian. Even though the number of participants was small, the results suggest that MoTR works well in controlled experiments.

One of MoTR's main advantages is its simplicity and accessibility. Because it runs in a web browser, it can be used both online and in physical locations, without the need of expensive equipment. While it doesn't offer the accuracy of eye-tracking, the blurred context outside the spotlight eliminates unwanted parafoveal effects, offering control over the text segments being read.

The statistical models confirm that reading times are strongly influenced by word length, frequency, and surprisal, findings that are in line with previous psycholinguistic research.

This research makes a new contribution by applying the MoTR paradigm in an experimental setting using Romanian, using a corpus adapted and validated for this task.

Future work involves expanding the experiment to a larger sample to increase confidence in results, a comparison between MoTR and traditional eye-tracking data, and the impact of time-guided lexical complexity predictions.

In conclusion, MoTR's ability to capture subtle aspects of cognitive processing during reading, along with its technical accessibility, makes it a strong alternative to traditional methods in experimental psycholinguistics.

Acknowledgments

This work was partially funded by the Romanian National Research Council (CNCS) through the Executive Agency for Higher Education, Research, Development and Innovation Funding (UEFISCDI) under grant PN-IV-P2-2.1-TE-2023-2007 (InstRead), and is supported by COST Action MultiplEYE, CA21131.

References

Fabian Anghel, Petru-Theodor Cristea, Claudiu Creanga, and Sergiu Nisioi. 2025. RALS: Resources and Baselines for Romanian Automatic Lexical Simplification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mariette Awad and Rahul Khanna. 2015. *Support Vector Regression*, pages 67–80. Apress, Berkeley, CA.

Veronica Boyce, Richard Futrell, and Roger P. Levy. 2020. Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111:104082.

Petru Cristea and Sergiu Nisioi. 2024. Archaeology at MLSP 2024: Machine translation for lexical complexity prediction and lexical simplification. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA2024)*, pages 610–617, Mexico City, Mexico. Association for Computational Linguistics.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online. Association for Computational Linguistics.

Kenneth I Forster, Christine Guerrera, and Lisa Elliot. 2009. The maze task: Measuring forced incremental sentence processing time. *Behavior research methods*, 41(1):163–171.

Abbas Ghaddar et al. 2024a. MLSP 2024 shared task on multilingual linguistic and semantic proficiency. In *MLSP Shared Task*.

Abbas Ghaddar et al. 2024b. MultiLS: A framework for measuring linguistic and semantic complexity across languages. *arXiv preprint arXiv:2402.14972*.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Alba Haveriku, Sara Bedulla, Nelda Kote, and Elinda Kajo Meçe. 2025. Understanding reading patterns of Albanian native readers through mouse tracking analysis. In *International Conference on Advanced Information Networking and Applications*, pages 433–443. Springer.

Nora Hollenstein, Itziar Gonzalez-Dios, Lisa Beinborn, and Lena Jäger. 2022. Patterns of text readability in human and predicted eye movements. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 1–15, Taipei, Taiwan. Association for Computational Linguistics.

Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Marcel A Just, Patricia A Carpenter, and Jacqueline D Woolley. 1982. Paradigms and processes in reading comprehension. *Journal of experimental psychology: General*, 111(2):228.

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

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.

Steven G. Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Ana Marcet and Manuel Perea. 2022. Does omitting the accent mark in a word affect sentence reading? evidence from spanish. *The Quarterly Journal of Experimental Psychology*, 75(1):148–155.

George W McConkie and Keith Rayner. 1975. The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, 17(6):578–586.

Metehan Oğuz, Cui Ding, Ethan Gotlieb Wilcox, and Zuzanna Fuchs. 2025. Using MoTR to probe gender agreement in russian. *PsyArXiv*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

Cosima Schneider, Nadine Bade, Michael Franke, and Markus Janczyk. 2021. Presuppositions of determiners are immediately used to disambiguate utterance meaning: A mouse-tracking study on the german language. *Psychological research*, 85(3):1348–1366.

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

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

Ethan Gotlieb Wilcox, Cui Ding, Mrinmaya Sachan, and Lena Ann Jäger. 2024. Mouse tracking for reading (MoTR): A new naturalistic incremental processing measurement tool. *Journal of Memory and Language*, 138:104534.

Where Patients Slow Down: Surprisal, Uncertainty, and Simplification in French Clinical Reading

Oksana Ivchenko[1]* **Alamgir Munir Qazi**[2]* **Jamal Abdul Nasir**[2]
[1]Univ. Lille, CNRS, UMR 8163 - STL, F-59000 Lille, France
[2]School of Computer Science, University of Galway, Ireland
{oksana.ivchenko.etu@univ-lille.fr}
{a.qazi1, jamal.nasir}@universityofgalway.ie

Abstract

This eye-tracking study links language-model surprisal and contextual entropy to how 23 non-expert adults read French health texts. Participants read seven texts (clinical case, medical, general), each available in an Original and Simplified version. Surprisal and entropy were computed with eight autoregressive models (82M–8B parameters), and four complementary eye-tracking measures were analyzed. Surprisal correlates positively with early reading measures, peaking in the smallest GPT-2 models ($r \approx 0.26$) and weakening with model size. Entropy shows the opposite pattern, with negative correlations strongest in the 7B-8B models ($r \approx -0.13$), consistent with a skim-when-uncertain strategy. Surprisal effects are largest in *Clinical Original* passages and drop by $\sim$20% after simplification, whereas entropy effects are stable across domain and version. These findings expose a scaling paradox – where different model sizes are optimal for different cognitive signals – and suggest that French plain-language editing should focus on rewriting high-surprisal passages to reduce processing difficulty, and on avoiding high-entropy contexts for critical information.

1 Introduction

Developing efficient methods to detect reading difficulty in healthcare materials is crucial for text simplification efforts (Fox, 2014). However, standard readability metrics provide limited insight into where and why readers struggle. Healthcare materials are frequently difficult for patients to understand (Rey et al., 2023), yet traditional measures fail to capture the localized nature of reading difficulty. Eye-tracking shows that effort is highly localized: readers invest extra time where their expectations are violated or where contextual uncertainty is high, then skim easier stretches (Ehrlich and Rayner, 1981; Rayner, 1998). Probabilistic language models (LMs) quantify these two information states that drive reading difficulty. Surprisal captures the unexpectedness of the word that actually appears and robustly predicts reading time (Smith and Levy, 2013; Goodkind and Bicknell, 2018). Contextual entropy captures an anticipatory state: high entropy can induce skipping or shorter gazes, whereas low entropy makes prediction errors costlier (Linzen and Jaeger, 2016; Pimentel et al., 2023). Early eye movement measures reflect immediate processing difficulty, while late measures indicate integration and comprehension costs (Camblin et al., 2007). Recent work reveals a scaling paradox: surprisals from very large transformers ($>$ 2B) can diverge from human reading times, whereas mid-sized GPT-2 models sometimes align better (Oh and Schuler, 2023). This suggests that model size alone does not guarantee better psycholinguistic validity. Moreover, nearly all evidence comes from English newspapers or novels, with minimal work on health genres or French.

Using French clinical and general texts in original and simplified versions, we investigate which LM-based predictors best track reading difficulty for automated simplification systems. Specifically:
RQ1: Do effects vary by Domain (Clinical vs General) and Version (Original vs Simplified)?
RQ2: Which LMs align best with human data for each predictor?

In what follows, we describe the corpus texts (original documents and the creation of their simplified versions) in Section 2. In Section 3, we present the methodology. Section 4 is dedicated to the results. Finally, we conclude in Section 5.

*These authors contributed equally to this work.

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025, pages 52–57, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-007

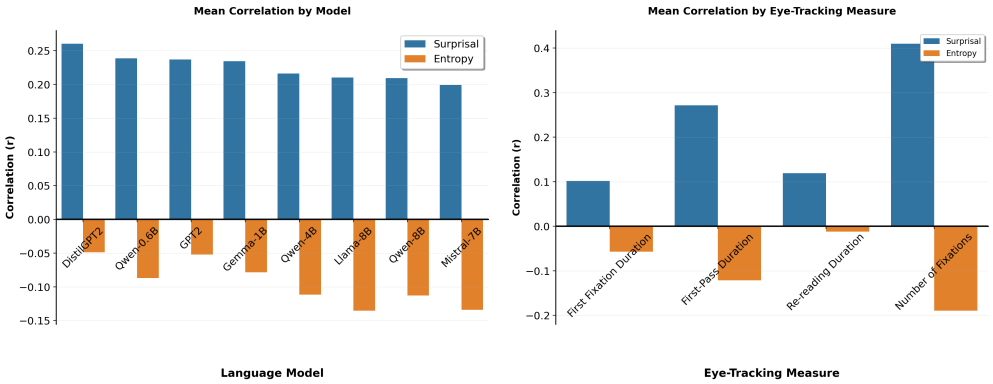

Figure 1: Mean Correlation

2 Data

2.1 Texts

We constructed a dataset of French-language texts from general and medical domains, based on excerpts from two corpora: CLEAR (Grabar and Cardon, 2018) and CAS (Grabar et al., 2020).

Each text was manually simplified following (OCDE, 2015) guidelines through syntactic, lexical, and semantic modifications, typically resulting in longer, clearer versions. Counterbalancing eliminated familiarity bias by exposing each participant to only one version of each text. Table 2 in the Appendix contains the full breakdown by words, sentences and screens.

2.2 Participants & Procedure

Gaze data were recorded using a Tobii Pro Spectrum eye tracker sampling at 600 Hz.

Texts were presented slide-by-slide, with some slides including comprehension questions for engagement. Tobii Pro Lab managed text presentation and automatically defined word-level Areas of Interest (AOIs).

The sample comprised 23 French participants aged 18–42 years (M = 22.8, SD = 6.2). Participants come from various social backgrounds - including students, doctoral students, and working professionals - but none have medical training.

3 Modeling

3.1 Language Models

We evaluated eight pre-trained autoregressive LMs spanning nearly three orders of magnitude in size (Table 1). Selection criteria were (i) good French coverage and (ii) architectural variety: four Byte-Pair Encoding (BPE) tokenisers (DistilGPT-2, GPT-2, two Qwen variants, Llama-3.1) and four SentencePiece models (Gemma-1B, Mistral-7B, Qwen-4B, Llama-8B). All models were run via `HuggingFace Transformers` with identical inference settings (`temperature` = 0, no sampling).

3.2 Eye-Movement Measures

We focus on four eye–movement measures, each indexing a distinct stage of processing:

Duration of first fixation (DFF) – immediate lexical access (time of the very first fixation);
First-pass duration (FPD) – initial comprehension (total dwell time during the first encounter);
Number of fixations (NFix) – overall processing effort (count of all fixations on the word);
Re-reading duration (RRD) – later integration/repair (time spent re-visiting the word).

These measures collectively span the complete timeline from initial word recognition to final comprehension, allowing us to assess how psycholinguistic predictions manifest across different aspects of the reading process.

3.2.1 Surprisal

We computed word-level surprisal as the negative log probability of each word given its left context:

$$\text{Surprisal}(w_i) = -\log_2 P(w_i \mid w_1, \ldots, w_{i-1}) \quad (1)$$

For each sentence, we obtained the model's probability distribution over the vocabulary at each position using a forward pass, extracted the probability assigned to the observed word, and converted to bits using base-2 logarithms. Surprisal values were aggregated from subword tokens to word level by

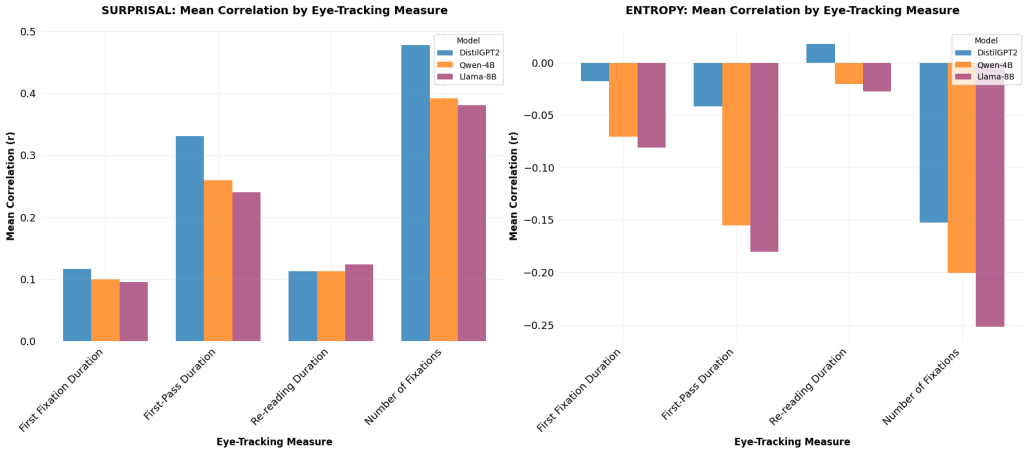

Figure 2: Mean correlations between model-generated surprisal (left panel) and entropy (right panel) with four eye-tracking measures. DistilGPT2 (82M parameters) consistently outperforms medium-sized Qwen-4B (4B parameters) and large Llama-8B (8B parameters) across the reading measures.

Model	Parameters
DistilGPT2	82M
GPT2	124M
Qwen3-0.6B	600M
Gemma-3-1B-IT	1B
Qwen3-4B	4B
Mistral-7B-Instruct-v0.3	7B
Qwen3-8B	8B
Llama-3.1-8B-Instruct	8B

Table 1: Overview of language models evaluated in this study, ranging from 82M to 8B parameters.

summing surprisal across all tokens comprising each word.

3.2.2 Contextual Entropy

We calculated the entropy of the model's predictive distribution at each word position:

$$H(i) = -\sum_{w} P(w \mid c_i) \log_2 P(w \mid c_i) \quad (2)$$

where $c_i = w_1, \ldots, w_{i-1}$ represents the left context.

This measure captures the model's uncertainty about what word should come next, independent of the actual word that appears. Higher entropy values indicate greater uncertainty in the model's predictions.

3.3 Data Processing and Token Alignment

3.3.1 Pre-processing

We rebuilt sentence strings by concatenating word tokens and normalising surrounding punctuation. For eye-movement data, duration metrics kept only positive values, whereas count metrics kept zeros but dropped negatives. Outliers were trimmed with measure-specific cut-offs: the upper 99 % for durations and the upper 95 % for counts. Analyses were run only when a cell contained at least ten valid observations, ensuring stable statistics.

3.3.2 Character-Position Mapping Algorithm

The technical challenge involved aligning model subword tokens with human word boundaries. French words often tokenize into multiple subwords (e.g., "L'obstétrique" → ["L'", "obsté", "trique"]), but humans process complete orthographic words.

Our alignment algorithm proceeded as follows:
(1) Extract character spans for each token using the tokenizer's offset mapping
(2) Define word boundaries from whitespace-delimited text
(3) For each word, identify all overlapping tokens using character position intersection
(4) Sum surprisal values of overlapping tokens to obtain word-level surprisal
(5) Average entropy values across tokens within each word
(6) Handle edge cases (partial overlaps, missing tokens) with fallback procedures

This method generalizes across tokenization schemes and languages, enabling consistent surprisal calculation regardless of subword segmentation. The algorithm successfully aligned tokens with word boundaries across all experimental conditions.

3.4 Statistical Analysis

3.4.1 Pearson Correlation Coefficient

We employed Pearson product-moment correlation as our primary statistical measure to quantify the linear relationship between language model predictions and human eye-movement behavior. The Pearson correlation coefficient r is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (3)$$

where x_i represents individual language model predictions (surprisal or entropy values), y_i represents corresponding eye-movement measures, $\bar{x}$ and $\bar{y}$ are sample means, and n is the number of word-level observations.

3.4.2 Correlation Analysis Framework

For every participant – text – metric cell we computed Pearson correlations between each predictor and the corresponding eye measure. The fully crossed design produced 23 participants $\times$ 8 texts $\times$ 4 metrics $\times$ 2 predictors = 1 472 correlation tests (counterbalancing included).

Surprisal correlation : r between word-level surprisal and the eye metric.

Entropy correlation : r between contextual entropy and the eye metric.

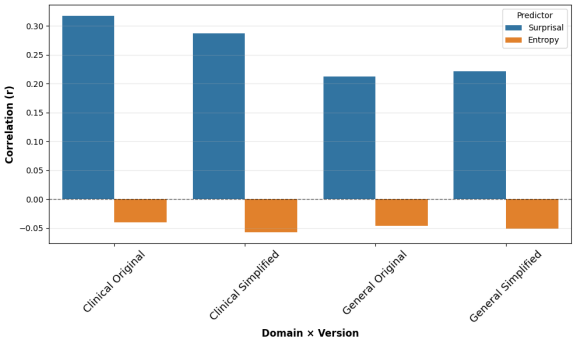

Figure 3: Domain Analysis

4 Results

Figure 1 summarises the aggregate correlations. In the left panel, surprisal (blue) is positive for every model, peaking in the two GPT-2 variants ($r \approx .25$) and tapering off as size increases. Entropy (orange) is negative and grows in magnitude, reaching $r \approx -0.14$ for the 7–8 B transformers. Hence small models best capture surprisal-driven slow-downs, while large models best capture the skim-when-uncertain effect indexed by entropy.

The right panel aggregates across models to compare eye-tracking metrics. Surprisal is strongest for *NFix* and *FPD*, weaker for *RRD*, and minimal for *DFF*. Entropy exhibits the reverse profile: it is most negative for *NFix*, moderate for *FPD*, and near zero on later measures – supporting the interpretation that surprisal indexes integration difficulty, whereas entropy reflects a strategic (skim-when-uncertain) allocation of attention.

Figure 2 contrasts a small (DistilGPT-2, 82 M), mid-size (Qwen-4B, 4 B) and large (Llama-8B, 8 B) model across the four eye metrics.

Surprisal. The ordering of effects is preserved across models, but magnitudes shrink as model size increases: DistilGPT-2 reaches $r = 0.48$ on *NFix* and $r = 0.33$ on *FPD*, whereas Llama-8B falls to $r = 0.38$ and $r = 0.24$, respectively. Small models therefore yield the clearest surprisal signal.

Entropy. The pattern is reversed. DistilGPT-2 shows near-zero correlations, Qwen-4B shows moderately negative correlations, and Llama-8B shows the strongest negative effects ($r = -0.25$ on *NFix*, $r = -0.18$ on *FPD*). The ranking of measures also flips: entropy effects are largest for fixation count and first-pass metrics, but minimal for *DFF* and *RRD*.

Figure 3 plots predictor strength by domain and simplification. Surprisal peaks in *Clinical Original* passages ($r \approx .32$), drops to $r \approx .29$ after simplification, and is lower overall in *General* texts ($r \approx .27 - .28$). Clinical terminology therefore amplifies error-driven slow-downs, and plain-language rewriting mitigates – but does not eliminate – this cost.

Entropy (orange) stays small and negative in every condition ($r \approx -0.03 - -0.05$) and shows no clear domain or version effect, implying that the skim-when-uncertain strategy is domain-invariant.

In short, simplification primarily reduces surprisal-based integration effort in specialized texts, while entropy-based allocation of attention remains unchanged.

5 Conclusion & Future Work

We demonstrate that LM-derived surprisal and entropy capture *different* aspects of French reading behavior, with effects that depend on text type: clinical originals produce the largest surprisal-driven slow-downs, while entropy effects remain modest

and stable across conditions. Small GPT-2 models best predict surprisal-based processing costs, whereas large 7–8B models best predict entropy-driven skimming behavior. Future work will (i) extend the corpus to longer passages and more readers, (ii) model text-level variation more explicitly by identifying which text properties modulate surprisal and entropy effects, (iii) investigate individual differences in reading strategies, and (iv) develop an automated simplification pipeline.

The current analysis is limited to clinical and general texts. Future studies will incorporate medical texts to examine domain effects more comprehensively.

Acknowledgments

This research is partially funded by MultiplEYE COST Action CA21131. Alamgir Munir Qazi is supported by the European Union's Horizon Europe programme under grant agreement No 101135757, project AI4Debunk[1].

This work was partially funded by the French National Agency for Research (ANR) as part of the CLEAR project (Communication, Literacy, Education, Accessibility, Readability), ANR-17-CE19-0016-01. Oksana Ivchenko is supported by the French National Agency for Research (ANR).

References

C. Christine Camblin, Peter C. Gordon, and Tamara Y. Swaab. 2007. The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language*, 56(1):103–128.

Susan F. Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655.

Susannah Fox. 2014. The social life of health information. Technical report, Pew Internet & American Life Project, Washington DC.

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*, pages 10–18, New Orleans, LA. Association for Computational Linguistics.

Natalia Grabar and Rémi Cardon. 2018. Clear – simple corpus for medical French. In *Workshop on Automatic Text Adaption (ATA)*, pages 1–11.

[1] https://ai4debunk.eu

Natalia Grabar, Clément Dalloux, and Vincent Claveau. 2020. CAS: corpus of clinical cases in French. *Journal of BioMedical Semantics*, 11(1):1–7.

Tal Linzen and T. Florian Jaeger. 2016. Uncertainty and expectation in sentence processing: Evidence from subcategorization probabilities. *Cognitive Science*, 40(6):1382–1411.

OCDE. 2015. *Guide de style de l'OCDE Troisième édition: Troisième édition*. OECD Publishing.

Byung-Doh Oh and William Schuler. 2023. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1770–1784, Singapore. Association for Computational Linguistics.

Tiago Pimentel, Clara Meister, Ethan Wilcox, Roger P. Levy, and Ryan Cotterell. 2023. On the effect of anticipation on reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1234–1248, Toronto, Canada. Association for Computational Linguistics.

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

Sylvie Rey, Aude Leduc, Xavier Debussche, Laurent Rigal, and Virginie Ringa. 2023. Une personne sur dix éprouve des difficultés de compréhension de l'information médicale. Études et Résultats 1269, Direction de la Recherche, des Études, de l'Évaluation et des Statistiques (DREES), Paris, France.

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

6 Appendix

Table 2: Comparison of Original and Simplified Texts

text_type	text_name	version	total_screens	total_sentences	total_words
clinical	toxico	original	4	19	398
		simplified	5	29	469
clinical	gastro	original	3	13	255
		simplified	3	13	336
general	weekend	original	9	31	844
		simplified	9	49	811
general	camelot	original	8	42	840
		simplified	8	58	880
medical	obstetrics	original	12	57	1104
		simplified	12	65	1202
medical	stroke	original	3	10	276
		simplified	3	22	328
medical	ulcer	original	15	77	1526
		simplified	15	92	1551

AlEYEgnment: Leveraging Eye-Tracking-while-Reading to Align Language Models with Human Preferences

Anna Bondar
Department of Computational Linguistics, Digital Society Initiative
University of Zurich
Zurich, Switzerland
anna.bondar@uzh.ch

David Robert Reich
Department of Computational Linguistics
University of Zurich
Zurich, Switzerland
davidrobert.reich@uzh.ch

Lena A. Jäger
Department of Computational Linguistics, Department of Informatics
University of Zurich
Zurich, Switzerland
lenaann.jaeger@uzh.ch

Abstract

Direct Preference Optimisation (DPO) has emerged as an effective approach for aligning large language models (LLMs) with human preferences. However, its reliance on binary feedback restricts its ability to capture nuanced human judgements. To address this limitation, we introduce a gaze-informed extension that incorporates implicit, fine-grained signals from eye-tracking-while-reading into the DPO framework. Eye movements, reflecting real-time human cognitive processing, provide fine-grained signals about the linguistic characteristics of the text that is being read. We leverage these signals and modify DPO by introducing a gaze-based additional loss term, that quantifies the differences between the model's internal sentence representations and cognitive (i.e., gaze-based) representations derived from the readers' gaze patterns. We explore the use of both human and synthetic gaze signals, employing a generative model of eye movements in reading to generate supplementary training data, ensuring the scalability of our approach. We apply the proposed approach to modelling linguistic acceptability. Experiments conducted on the CoLA dataset demonstrate performance gains in grammatical acceptability classification tasks when the models are trained in the gaze-augmented setting. These results demonstrate the utility of leveraging gaze data to align language models with human preferences. All code and data are available from Github.

1 Introduction

Direct Preference Optimisation (DPO, Rafailov et al., 2023) has recently emerged as a scalable, computationally efficient, stable method for aligning language models with human preferences. Unlike Reinforcement Learning from Human Feedback (RLHF, Christiano et al., 2017) or Reinforcement Learning from AI Feedback (RLAIF, Lee et al., 2024), it does not require a separate reward model and allows a policy model to internalise the preferences directly. However, DPO relies only on binary pairs of preferred and dispreferred responses, and this simplicity leads to a critical limitation: binary feedback provides no information about how strongly one response is preferred over another, limiting the model's ability to align with nuanced human judgements. Recent studies have demonstrated that integrating explicit, fine-grained preference labels—such as ranked lists or ordinal scores— into a DPO-based framework improves the alignment of a policy model with human preferences (Liu et al., 2025; Zhao et al., 2024). However, collecting explicit high-quality detailed annotations from humans at scale is labour-intensive and costly.

To address the outlined limitations, we introduce a method that leverages implicit human feedback from eye-tracking data collected during reading. Eye movements are considered the gold-standard method to investigate cognitive processes underlying language processing (Rayner, 1998; Clifton et al., 2007). Because eye movement patterns systematically reflect processing difficulty and readers' evaluations of linguistic input, these gaze signals can provide detailed, fine-grained indicators of human preferences, reducing reliance on explicit detailed human ranking or rewards from auxiliary models. In our approach, we integrate gaze-based signals into the DPO training pipeline, allowing the model to incorporate nuanced human feedback beyond binary supervision while retaining DPO's computational efficiency. Recent advances in generative models of eye movements in reading further support the scalability of this method (Prasse et al., 2023; Deng et al., 2023b; Bolliger et al., 2023, 2025) since they make it possible to generate synthetic human-like scanpaths and increase the training dataset without collecting data from humans.

The specific downstream task we focus on is

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025, pages 58–70, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-008

modelling grammatical acceptability (terms grammaticality and acceptability are used interchangeably) judgements. Models for this task are typically trained using supervised learning with binary labels that categorise sentences into acceptable versus unacceptable ones. However, psycholinguistic research (Lau et al., 2017; Francis, 2021) demonstrates that speakers perceive grammatical acceptability along a gradient rather than as a binary label. Eye-tracking data holds potential to inform the model about the degree of ungrammaticality, as psycholinguistic evidence demonstrates that eye movement patterns vary depending on the degree of grammar violations (Tuninetti et al., 2015; Rayner et al., 2004). Similarly, eye movements in reading can provide information on the type, strength of ungrammaticality (Braze et al., 2002) and its location within the utterance (Vasishth et al., 2013; Frazier and Rayner, 1982). Once an ungrammaticality is encountered, the reader's eye-movement patterns tend to exhibit longer fixation durations, an increased number of regressions, and disrupted saccadic movements—reflecting increased processing difficulty and reanalysis effort. These findings suggest that eye-tracking data can supply the fine-grained online signal missing from binary annotations and inform the model about the strength, locus, and characteristics of grammar violations as perceived by humans. Building on this foundation, we investigate two principal research questions: (i) whether integrating gaze signals into DPO during training improves model performance on grammatical acceptability; and (ii) whether increasing the amount of training data by adding synthetic gaze data leads to further gains in model performance.

2 Related Work

2.1 Human-Preferences Alignment

Direct Preference Optimisation has been proposed as a streamlined alternative to RLHF and RLAIF, as it trains directly on binary preferred–dispreferred pairs and does not require a learned reward model (Rafailov et al., 2023). However, this pairwise supervision limits the model's capacity to reflect how strongly one response is preferred over another. Recent work has introduced methods to incorporate finer-grained information. One of these methods—Ordinal Preference Optimisation (OPO)—replaces binary comparisons with ranked lists, enabling the model to capture relative distances among responses (Zhao et al., 2024). Another approach—Listwise Preference Optimisation (LiPO)—extends this idea by formulating alignment as a learning-to-rank problem (Liu et al., 2025). An alternative method, namely Margin Matching Preference Optimisation (MMPO), retains the pairwise format of the responses and attaches real-valued quality margins to each pair (Kim et al., 2024). All of these approaches rely on explicit, graded feedback, either from human annotations or reward models, which can be costly to obtain or may diverge from human judgements when using external LLMs to score the responses (Bavaresco et al., 2025).

2.2 Eye Movements in Reading as Indicators of Grammatical Violations

Eye-tracking studies have demonstrated that gaze patterns reliably identify the locus, type, and strength of grammatical violations (Schotter and Dillon, 2025). Readers precisely localise syntactic anomalies, leading to immediate regressions and increased fixation durations at points of structural disambiguation or grammatical inconsistency (Frazier and Rayner, 1982; Vasishth et al., 2013). Furthermore, distinct gaze signatures differentiate violation types: syntactic errors (e.g., agreement mismatches or structural ambiguities) typically cause rapid regressions and increased first fixation durations, whereas semantic and pragmatic anomalies predominantly affect later reading measures, such as regression-path duration and total fixation time (Braze et al., 2002).

Eye movements also systematically reflect the strength of violation. Strong violations, such as outright ungrammatical constructions or semantically impossible continuations, provoke immediate disruptions in first-pass reading times and frequent regressions. Conversely, milder violations, such as subtle semantic implausibilities or pragmatic errors, result in delayed and comparatively moderate reading disruptions, evident primarily through increased regression-path durations and cumulative reading times (Rayner et al., 2004; Tuninetti et al., 2015; Joseph et al., 2009; Schotter and Dillon, 2025; Schotter and Jia, 2016). Overall, these findings demonstrate the utility of eye-tracking as a fine-grained implicit feedback on the processing of grammatical violations in real-time language comprehension.

2.3 Eye-Tracking-while-Reading for Natural Language Processing

Eye movements in reading have been leveraged for model evaluation and interpretation, including the assessment of a model's and cognitive plausibility (Bolliger et al., 2024; Beinborn and Hollenstein, 2024; Haller et al., 2024; Goodkind and Bicknell, 2018; Bensemann et al., 2022; Eberle et al., 2022; Sood et al., 2020a; Hollenstein and Beinborn, 2021).

Besides model evaluation and interpretation, gaze signals have proven effective for training and evaluating NLP models. Recent research demonstrated that eye movements in reading can be leveraged as a supervisory signal to enhance model performance on various downstream NLP tasks. Early research employed eye-tracking data in the form of auxiliary input alongside the text embeddings for named entity recognition, sentiment analysis, sarcasm detection, part-of-speech tagging (Hollenstein and Zhang, 2019; Mishra et al., 2016; Barrett et al., 2016; Tiwari et al., 2023). Other studies integrated reading measures into models to guide attention mechanism directly for visual question answering, sentence compression and paraphrase generation, sentiment analysis (Sood et al., 2020b; Long et al., 2017; Sood et al., 2023). Further research utilised gaze data in transfer learning settings, tasking the models to predict reading measures as an auxiliary training objective for sarcasm detection, readability prediction, or machine reading comprehension (Yang and Hollenstein, 2023; Deng et al., 2023a; González-Garduño and Søgaard, 2018; Malmaud et al., 2020). A more recent line of research reordered the input sequence according to the scanpaths (Yang and Hollenstein, 2023; Deng et al., 2024) at the fine-tuning stage. All of the listed frameworks have demonstrated the utility of eye movements in reading for a wide range of NLP tasks and have exhibited comparable performance using either real human or synthetic eye-tracking data.

Most recently, eye-tracking data has been integrated into frameworks aimed at aligning human preferences, specifically in reward modelling within RLHF paradigms (López-Cardona et al., 2025). Eye movements have also shown promising results for constructing datasets reflecting human preferences (Kiegeland et al., 2024; Lopez-Cardona et al., 2025). Nevertheless, directly applying gaze data to preference alignment frameworks without relying on intermediate reward models remains unexplored. We address this gap and demonstrate the utility of eye-tracking-while-reading data for directly aligning large language models with human preferences.

3 Preliminaries

We first provide a short overview of Direct Preference Optimisation (DPO, Rafailov et al., 2023), a method for aligning language models with human preferences. This approach is a further development of RLHF (Ouyang et al., 2022) and relies on a policy model—the model being trained—a reference model—a frozen, pre-trained checkpoint used to regularise training and keep the policy close to its initial weights—and a reward model, which assigns rewards to outputs produced by the policy. DPO eliminates this reward model and instead optimises the policy to increase the (log-)probability of preferred over dispreferred responses directly.

Given a dataset of triples (r, x^1, x^0)—a prompt r with a preferred (chosen) response x^1 and a dispreferred (rejected) response x^0—DPO updates a policy π_θ relative to a fixed reference policy π_{ref} by maximising the Bradley–Terry log-likelihood:

$$\max_\theta \mathbb{E}_{(r,x^1,x^0)}\Big[\log\sigma\big(\beta\big(\log\frac{\pi_\theta(x^1|r)}{\pi_{\text{ref}}(x^1|r)} - \log\frac{\pi_\theta(x^0|r)}{\pi_{\text{ref}}(x^0|r)}\big)\big)\Big], \quad (1)$$

where σ is a sigmoid function that maps the difference in log-probability ratios between the policy and reference models to a value in $(0, 1)$, which can be interpreted as the probability that the policy assigns a higher probability to the preferred response x^1, β is a temperature parameter that controls the sensitivity of the model to small differences between the preferred and dispreferred options. This objective directly increases the model's relative log-probability of preferred over dispreferred responses.

4 Problem Setting

The task of linguistic acceptability classification is a supervised learning problem, where the goal is to determine whether a given natural language expression conforms to the grammatical norms of a particular language variety. Formally, let $\mathcal{X} \subset \Sigma^*$ denote all possible input strings over a finite vocabulary Σ. Each input $x \in \mathcal{X}$ is a sentence. The output is $\mathcal{Y} \in \{0, 1\}$, where $y = 1$ indicates an acceptable expression and $y = 0$ denotes an unacceptable

one. Given a dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$ sampled from $\mathcal{X} \times \mathcal{Y}$, the objective is to find a function $f_\theta : \mathcal{X} \rightarrow [0, 1]$ parametrised by θ, where $f_\theta(x)$ represents the predicted probability of acceptability. We investigate two questions: (i) whether incorporating human eye-tracking signals at training time improves performance on grammatical acceptability classification, and (ii) whether adding synthetic gaze provides further gains beyond human signals alone.

We evaluate our models and report the performance with accuracy, F_1, and Matthews correlation coefficient on held-out data.

5 Data

5.1 The CoLAGaze Corpus

We utilised the CoLAGaze eye-tracking-while-reading corpus (Bondar et al., 2025) to integrate implicit human feedback into the Direct Preference Optimisation framework. The dataset comprises eye-tracking data collected from 42 participants reading 153 pairs of (un)grammatical sentences manually selected from the Corpus of Linguistic Acceptability (CoLA; Warstadt et al., 2019). Each participant read either the grammatical or the ungrammatical counterpart of each sentence. These sentences span a diverse range of grammatical violations, including syntactic, morphosyntactic, semantic, and pragmatic anomalies. Detailed information on the original data collection procedure, preprocessing steps, and computation of reading measures can be found in Bondar et al. (2025). The full corpus contains 6,246 data points in total. For our analyses, we selected data from 38 well-calibrated participants, resulting in a total of 5,814 data points.

5.2 Synthetic Data

In addition to the human data provided by CoLAGaze, we trained Eyettention (Deng et al., 2023b), a state-of-the-art generative model of eye movements in reading, to produce synthetic scanpaths (i.e., sequences of fixations and saccades) for an additional 30 sentence pairs from CoLA (see Appendix B for more details), with the goal of extending the training set beyond the original CoLAGaze data and assess whether gaze-informed models can benefit from synthetic gaze signals during training.

6 Method

6.1 DPO with Binary Feedback

To address the linguistic acceptability classification task, we fine-tune a 7-billion-parameter instruction-tuned Mistral model within the Direct Preference Optimisation framework. To fit the DPO setup we form a set of 153 pairwise preferences $\mathcal{P} = \left\{\left(x_c^1, x_c^0\right)\right\}_{c=1}^{C}$, where x_c^1 represents a grammatical sentence and x_c^0 its ungrammatical counterpart, and C represents the total number of pairs. The DPO setup employs two pretrained LLMs: the policy model π_θ, initialised from the instruction-tuned Mistral checkpoint and fine-tuned during training, and the reference model π_{ref}, which shares the initial parameters of the policy model but remains frozen throughout fine-tuning to stabilise learning and avoid catastrophic forgetting. Given a prompt r that explicitly instructs the model to identify the grammatical sentence from the pair (x_c^1, x_c^0) (for details on the prompt, see Section 7), the policy model is trained to generate the grammatical sentence as output. We optimise the parameters of the model using the standard DPO objective (for details on standard DPO, see Section 3).

6.2 DPO Augmented with Gaze-Based Implicit Feedback

6.2.1 Eye-Tracking Feature Selection

We used sentence-level eye-tracking measures from CoLAGaze calculated after correcting for vertical drift. Specifically, we selected a subset of eye-tracking features most predictive of sentence-level acceptability across all violation types included in CoLAGaze. To select the subset from the CoLAGaze dataset, we fit a binomial generalised linear mixed model to predict sentence labels from the eye-tracking features and perform greedy backward (recursive) elimination, removing one feature at a time and refitting the model. Feature selection is guided by the Bayesian Information Criterion (BIC) (Schwarz, 1978). The final set of features is the one that minimises BIC.

Once the subset of the eye-tracking features is selected, we train our gaze-augmented large language models with two sets of eye-tracking measures (see Appendix C for a comprehensive list of measures and their definitions): measures based on event counts (e.g. number of fixations, number of regressions) and measures based on durations (e.g. total fixation duration, first-pass reading time). Models

augmented with synthetic eye-tracking data utilise only event-count based features, as the Eyettention model employed for synthetic data generation does not predict fixation durations.

6.2.2 Integration of the Eye-Tracking Data

To integrate the cognitive information into the DPO framework, we introduce an additional gaze-based loss term ℓ_{ET} to the original DPO loss function, that quantifies the alignment of the model's internal sentence representations h to cognitive (i.e., gaze-based) representations g derived from the sequence of eye-movement events s (see Figure 1 for a visualisation of how the gaze-based loss term is derived). To compute the eye-tracking based loss term, for a grammatical–ungrammatical pair (x_c^1, x_c^0), we obtain the sentence embeddings $h_c^1, h_c^0 \in \mathbb{R}^d$ from the policy model π_θ. To get the embeddings we tokenise the two sentences from each pair into two separate sequences $T_{x_c} = \{\, t_1, \ldots, t_{|T_{x_c}|} \,\}$, feed each of the sequences to the model π_θ, extract the hidden states of the last layer $H \in \mathbb{R}^{T_x \times d}$ from the model and use mean pooling to derive a sentence representation

$$h_c \;=\; \frac{1}{T_{x_c}} \sum_{q=1}^{T_{x_c}} H_q, \tag{2}$$

where q is a token position in a sequence. To integrate gaze data into the loss, we form eye-tracking feature vectors g_c, consisting of the selected sentence-level eye-tracking features from CoLAGaze (see 6.2.1). Let I_c^1 and I_c^0 denote the set of readers who saw x_c^1 and x_c^0, respectively[1]; for each reader $i \in I_c^1$, or $j \in I_c^0$, and for each sentence x_c^1, or x_c^0, we form a sentence-level gaze feature vector $g_{c,i}^1 \in \mathbb{R}^F$, or $g_{c,j}^0 \in \mathbb{R}^F$, where F is the number of gaze features. For each sentence pair (x_c^1, x_c^0) we form $K = 20$ cross-participant vector pairs by independently sampling indices $i_k \in I_c^1$ and $j_k \in I_c^0$ with replacement for each pair; these indices are fixed once at the start of training. (we treat the number of pairs K as a hyperparameter, see Appendix A for details) and compute the difference between them $\Delta_{\text{gaze}_c}^{(k)} = g_{c,i_k}^1 - g_{c,j_k}^0$. We then treat each tuple consisting of the prompt, the grammatical and ungrammatical sentences and the gaze vector difference $(r, x_c^1, x_c^0, \Delta_{\text{gaze}_c}^{(k)})$ as a separate gaze-augmented training instance. For each of the instances we compute the gaze-based loss term

$$\ell_{\text{ET}_c}^{(k)} = \cos(h_c^1, h_c^0)\, \left\|\Delta_{\text{gaze}_c}^{(k)}\right\|_2. \tag{3}$$

Because $\cos(h_c^1, h_c^0)$ is in the range $[-1, 1]$, ℓ_{ET_c} penalises the model when the sentence embeddings are too similar while the differences in human gaze patterns are large — in this case the cosine term is close to 1 and the Euclidean distance between the gaze vectors $\left\|\Delta_{\text{gaze}_c}^{(k)}\right\|_2$ is large, this results into large positive gaze-based loss. On the other hand, when the sentence representations are already well separated (cosine term closer to -1), the gaze-based loss term $\ell_{\text{ET}_c}^{(k)}$ becomes negative and implicitly rewards the model by decreasing the total loss. Training minimises the expectation over the final loss:

$$\mathcal{L}_{\text{total}} = \mathbb{E}_{(r, x^1, x^0, \Delta_{\text{gaze}})} \left[\, \mathcal{L}_{\text{DPO}}(\theta) + \alpha\, \ell_{\text{ET}} \,\right], \tag{4}$$

where α is a tuned hyperparameter. By training the model with a gaze-based loss term we intend to align the model's representations with human cognitive processing signals.

7 Experiments

7.1 Training Setup

We fine-tuned the 7-billion-parameter instruction-tuned Mistral model in several configurations to evaluate whether integrating implicit feedback derived from eye-tracking data into Direct Preference Optimisation enhances downstream performance on grammatical acceptability classification. See Figure 2 for a summary of the training and evaluation pipeline. Training details are available in Appendix A.

We model grammatical acceptability as a binary classification task, implemented as text generation with a decoder-only transformer. At training, for each item, both grammatical and ungrammatical sentences are presented in a single prompt:

```
Select the grammatically correct
sentence:
A) <sent_A>
B) <sent_B>
```

The assignment of the grammatical option to A or B is random to avoid position cues. The policy model π_θ computes log-probabilities for each sentence; grammatical sentences are treated as preferred responses and ungrammatical ones as dispreferred.

We augmented the DPO framework with gaze data in several configurations. First, we incorporated implicit human gaze feedback, where the

[1]The stimuli were presented in Latin square such that each reader saw either grammatical or ungrammatical version of each sentence

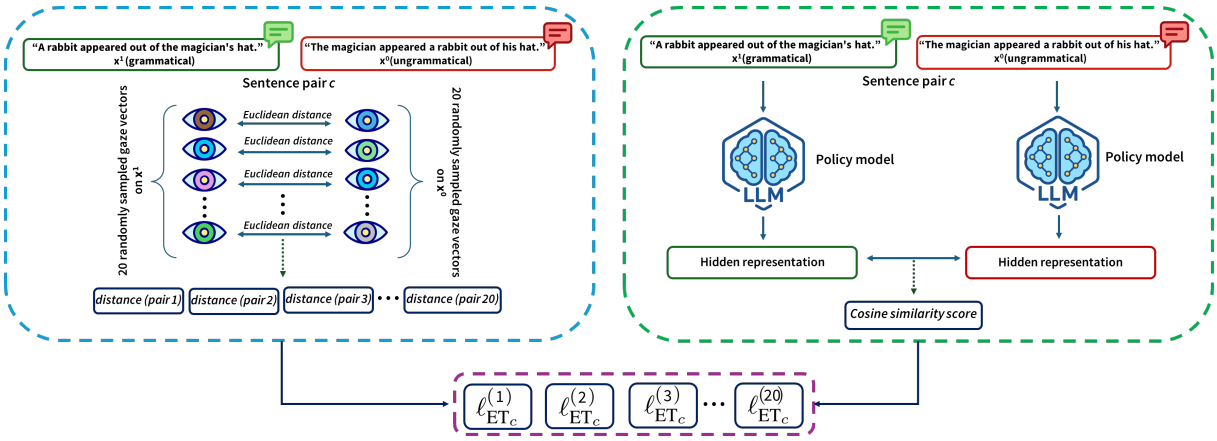

Figure 1: As depicted in the left blue box, for each grammatical–ungrammatical sentence pair c, we randomly sample 20 gaze vectors per grammatical sentence x^1 and per ungrammatical one x^0, drawn from different readers. From the sampled gaze vectors we then randomly pick a gaze vector for the grammatical sentence and a gaze vector for its ungrammatical counterpart and pair these gaze vectors, resulting into 20 gaze vector pairs. For each pair of gaze vectors, we compute the Euclidean distance between them to quantify the difference in gaze behaviour for the grammatical sentence compared to its ungrammatical counterpart. As shown in the right green box, we simultaneously, extract hidden representations of the sentences from the policy model and calculate the cosine similarity between the grammatical and ungrammatical sentences, reflecting their proximity in the model's internal representation space. Each gaze distance is then multiplied by the similarity of the hidden representations to produce a scalar eye-tracking-based loss $\ell_{\mathrm{ET}_c}^{(k)}$.

eye-tracking-based loss was derived from selected CoLAGaze features, using both event durations and event-count measures. Second, we experimented with using only count based features to assess whether duration based measures contribute to performance gains. Third, we extended this setup by incorporating synthetic eye-tracking data, by adding synthetic scanpaths on 30 additional sentence pairs generated by the Eyettention model (see B for details). Finally, we investigated whether averaging gaze features across all readers—representing an "average reader"—still leads to improved performance.

7.2 Baselines

We evaluated our method against three text-only baselines based on the instruction-tuned Mistral checkpoint. First, the Base model corresponds to the original checkpoint without any task-specific fine-tuning.

Second, we trained a Supervised Fine-Tuning (SFT) variant by optimising a cross-entropy loss on the 153 grammatical–ungrammatical sentence pairs from CoLA. When the policy model is trained in the SFT setting, it is fine-tuned to generate the acceptability label y from a prompt t containing a sentence x_n to be classified with a labely_n as being either grammatical (1) or ungrammatical (0) and a question *"Is this sentence grammatical?"*. SFT minimises the cross-entropy loss (negative log-likelihood) over dataset $\mathcal{D}$ defined in Section 4:

$$\begin{aligned}\mathcal{L}_{\mathrm{SFT}}(\theta) &= -\,\mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\log \pi_\theta\big(y \mid t, x\big)\big] \\ &\approx -\,\frac{1}{N}\sum_{n=1}^{N} \log \pi_\theta\big(y \mid t, x_n\big).\end{aligned} \tag{5}$$

Third, we trained a text-only DPO model (see Section 6.1 for details) using the same sentence pairs as in the previous training settings, relying solely on binary acceptability supervision without any cognitive signals.

7.3 Ablation

To further validate our findings, we conduct an ablation study eliminating the eye-tracking features in the additional loss term ℓ. In this variant, the standard DPO objective is augmented only with the cosine similarity between the two sentence embeddings, $\cos(h_c^1, h_c^0)$, omitting the gaze-difference term (i.e., effectively setting $\|\Delta_{\mathrm{gaze}_c}^{(k)}\|_2 = 0$).

7.4 Evaluation Setup

At test time, we use only the text data from the held-out CoLA training and development sets. Each test sentence is fed to the model alongside the following prompt:

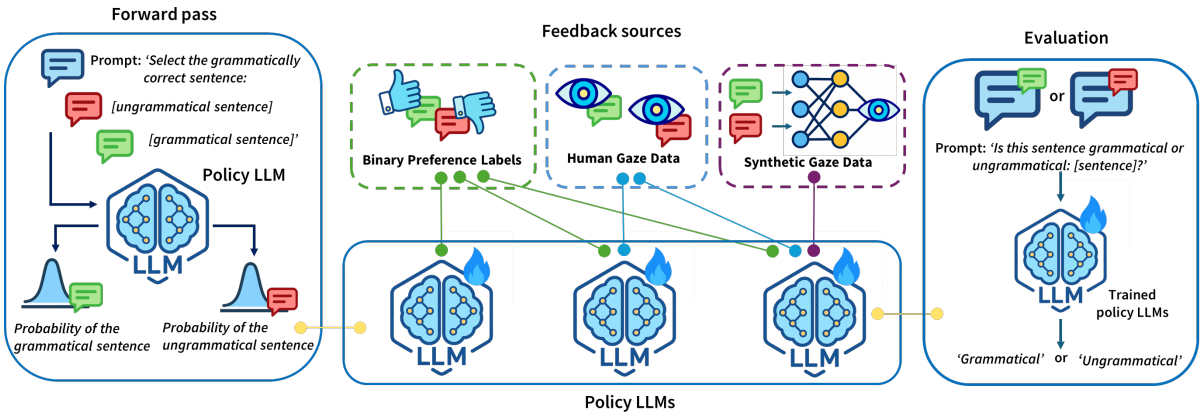

Figure 2: As depicted in the left part of the figure, during DPO training, the model receives a prompt containing a pair of sentences—one grammatical and one ungrammatical—and is tasked to select the grammatical one. We extract the probability of the grammatical and ungrammatical sentence being generated and use them for DPO. As the central block of the figure shows, we fine-tune the model and explore three training configurations: (i) using only binary preference labels (text-only), (ii) augmenting the labels with human gaze data, and (iii) further incorporating synthetic gaze data into the training. At evaluation, shown in the right side of the figure, the model is given a single sentence and prompted to classify it as grammatical or ungrammatical. Eye-tracking data is not used during evaluation.

```
Is this sentence grammatical or
ungrammatical?  <sent>
```

We report results separately on two subsets (both sourced from CoLA training and CoLA development set): sentences that share linguistic characteristics with the training data, such as similar syntactic constructions and lexical items (*in-domain* subset in the original CoLA dataset), and those that differ substantially from the training distribution (*out-of-domain* subset in the original CoLA dataset). Performance is measured using accuracy, F1 score, and Matthews correlation coefficient (MCC).

8 Results and Discussion

The results for grammatical acceptability classification on CoLA are summarised in Table 1.

Supervised Fine-Tuning performed notably worse than the instruction-tuned base model, obtaining an MCC of 0.463 in-domain and 0.410 out-of-domain. The model trained with text-only DPO also failed to surpass the base model's performance with an MCC of 0.460 and 0.406 for in-domain and out-of-domain, respectively. This drop in performance could be attributed in part to the small size of the dataset used for training (fine-tuning or DPO), which may have led the model to overfit and generalise poorly (Barnett et al., 2024). Additionally, the Supervised Fine-Tuning and DPO training was conducted using quantised low-rank adaptation (QLoRA, Dettmers et al., 2023), potentially further limiting the effective model capacity (Wang et al., 2024).

The best-performing model was Mistral fine-tuned using DPO augmented with eye-tracking features—both event-count and duration based. This model achieved an MCC of 0.510 in-domain and 0.502 out-of-domain. Relative to the baseline instruction-tuned Mistral model, gaze-augmented DPO improved the MCC by 0.037 points in-domain and 0.074 points out-of-domain. Similar improvements were observed for F1 and accuracy metrics. These results indicate benefits from integrating eye-tracking signals into the optimisation objective. In particular, this method appears useful in a low resource settings, as in our study the models were trained on only 153 sentence pairs. Finally, our ablation study demonstrates that, as expected, removing the gaze signal leads to inferior model performance.

We further compared gaze augmentation with all gaze-derived features against leveraging a reduced set containing only fixation- and saccade-count based features. The results showed an advantage of using all gaze features, suggesting that duration based gaze features contribute additional information beyond fixation counts alone.

Overall, the model trained with both gaze-event-count and duration based features outperformed the baseline and the models trained on text only. The results hold for all of the settings in which

Model	Gaze Data	Synthetic Data	Aggregated	Test Set	Accuracy↑	F1↑	MCC↑
Base	✗	✗	✗	in-domain	76.62	0.83	0.473
Base	✗	✗	✗	out-of-domain	72.87	0.79	0.428
SFT	✗	✗	✗	in-domain	$70.40_{2.45}$	$0.750_{0.033}$	$0.463_{0.009}$
SFT	✗	✗	✗	out-of-domain	$64.87_{2.36}$	$0.678_{0.037}$	$0.410_{0.019}$
DPO	✗	✗	✗	in-domain	$72.60_{0.19}$	$0.778_{0.004}$	$0.460_{0.006}$
DPO	✗	✗	✗	out-of-domain	$67.20_{0.41}$	$0.734_{0.03}$	$0.406_{0.008}$
Ablation	✗	✗	✗	in-domain	$76.24_{0.54}$	$0.8217_{0.0053}$	$0.4711_{0.0058}$
Ablation	✗	✗	✗	out-of-domain	$73.74_{0.68}$	$0.7929_{0.0083}$	$0.4521_{0.0029}$
DPO	all feat-s	✗	✗	in-domain	$80.17_{0.667}$	$0.864_{0.006}$	$0.510_{0.013}$
DPO	all feat-s	✗	✗	out-of-domain	$79.33_{0.872}$	$0.855_{0.008}$	$0.502_{0.008}$
DPO	count feat-s	✗	✗	in-domain	$79.53_{0.110}$	$0.858_{0.001}$	$0.499_{0.002}$
DPO	count feat-s	✗	✗	out-of-domain	$78.79_{0.469}$	$0.848_{0.004}$	$0.496_{0.009}$
DPO	all feat-s	✗	✓	in-domain	$79.76_{0.135}$	$0.868_{0.002}$	$0.490_{0.005}$
DPO	all feat-s	✗	✓	out-of-domain	$78.15_{0.070}$	$0.854_{0.001}$	$0.455_{0.001}$
DPO	count feat-s	✗	✓	in-domain	$76.24_{0.288}$	$0.821_{0.007}$	$0.472_{0.001}$
DPO	count feat-s	✗	✓	out-of-domain	$73.06_{0.551}$	$0.788_{0.008}$	$0.436_{0.003}$
DPO	count feat-s	✓	✓	in-domain	$76.15_{0.134}$	$0.821_{0.002}$	$0.471_{0.002}$
DPO	count feat-s	✓	✓	out-of-domain	$73.45_{0.820}$	$0.791_{0.009}$	$0.446_{0.007}$

Table 1: Results of training Mistral model on 153 sentence pairs from CoLA in different configurations: in-domain and out-of-domain subsets. Accuracy, F1 (positive = grammatical), and MCC are reported as mean_{SD} over 3 random seeds. *Gaze Data* indicates whether human eye-tracking features were used; *Synthetic Data* indicates whether synthetic gaze features were additonally used; *Aggregated* refers to whether gaze features were aggregated across readers. *All feat-s* in the *Gaze Data* columns means that both duration and event-count based features were used at training, *count feat-s* means that only event-count based features were leveraged.

the gaze data with all of the features was used — augmented with the data not aggregated across the readers, and with scanpaths aggregated across the readers. These findings are in line with the seminal work by Kliegl et al. (1982), who first showed that both duration and event-count based measures are informative about processing difficulty. The DPO training with event-count based features does not consistently lead to performance gains — while using the data not aggregated across the readers is beneficial, aggregating across participants leads to a decrease in performance in in-domain evaluation settings. Models where training was augmented with synthetic gaze data showed only marginal improvements over the base model on the out-of-domain test set. We attribute this to several factors, namely the small size of the synthetic dataset, the usage of a single gaze-feature vector per sentence, and the reliance on event-count reading measures only. Future research might investigate the integration of synthetic data with both duration and event-count based features, and explore the use of larger synthetic datasets.

Finally, future work may examine word-level eye-tracking features instead of sentence-level measures, as these have the potential to localise ungrammaticality within sentences and thereby provide the model with a more fine-grained and informative supervision signal.

9 Conclusion

We introduced a gaze-informed extension of Direct Preference Optimisation that aligns a large language model's internal representations with human cognitive processing signals. By integrating an eye-tracking loss term—derived from sentence-level differences in reading patterns observed on grammatical versus ungrammatical sentences—into the DPO objective, our approach injects graded, implicit feedback into training. Our experiments on CoLAGaze show that gaze-augmented models consistently outperform text-only baselines, and that both duration-based and count-based eye-tracking features provide useful signals beyond text alone.

Acknowledgements

This work was partially funded by the Digital Society Initiative at the University of Zurich via a PhD-scholarship granted to Anna Bondar, and the Swiss National Science Foundation under grant IZCOZ0_220330 (EyeNLG, PI: Lena Jäger). It was furthermore supported by the European Cooperation in Science and Technology (COST) under COST Action CA21131 (MultiplEYE).

References

Scott Barnett, Zac Brannelly, Stefanus Kurniawan, and Sheng Wong. 2024. Fine-Tuning or Fine-Failing?

Debunking Performance Myths in Large Language Models. *arXiv*, 2406.11201.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly Supervised Part-of-speech Tagging Using Eye-tracking Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–584, Berlin, Germany.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria.

Lisa Beinborn and Nora Hollenstein. 2024. *Cognitive Plausibility in Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Springer Nature, Cham, Switzerland.

Joshua Bensemann, Alex Peng, Diana Benavides-Prado, Yang Chen, Neset Tan, Paul Michael Corballis, Patricia Riddle, and Michael Witbrock. 2022. Eye Gaze and Self-attention: How Humans and Transformers Attend Words in Sentences. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–87, Dublin, Ireland.

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

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

Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. 2022. CELER: A 365-Participant Corpus of Eye Movements in L1 and L2 English Reading. *Open Mind*, 6:41–50.

Lena Bolliger, Patrick Haller, and Lena Jäger. 2024. On the Alignment of LM Language Generation and Human Language Comprehension. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 217–231, Miami, FL.

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*, page 15513–15538, Singapore.

Lena Bolliger, David Reich, and Lena Jäger. 2025. ScanDL 2.0: A Generative Model of Eye Movements in Reading Synthesizing Scanpaths and Fixation Durations. In *Proceedings of the ACM on Human-Computer Interaction*, 3, New York, NY.

Anna Bondar, David Reich, and Lena Jäger. 2025. CoLAGaze: A Corpus of Eye Movements for Linguistic Acceptability. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, New York, NY. Association for Computing Machinery.

David Braze, Donald Shankweiler, Weijia Ni, and Laura Conway Palumbo. 2002. Readers' Eye Movements Distinguish Anomalies of Form and Content. *Journal of Psycholinguistic Research*, 31(1):25–44.

Paul Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Proceedings of Advances in Neural Information Processing Systems 30*, pages 4299–4311, Long Beach, CA.

Charles Clifton, Adrian Staub, and Keith Rayner. 2007. Eye Movements in Reading Words and Sentences. In Roger Van Gompel, Martin Fischer, Wayne Murray, and Robin Hill, editors, *Eye Movements*, pages 341–371. Elsevier, Netherlands.

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.

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*, pages 217–224, Bangkok, Thailand.

Shuwen Deng, David Reich, Paul Prasse, Patrick Haller, Tobias Scheffer, and Lena Jäger. 2023b. Eyettention: An Attention-based Dual-Sequence Model for Predicting Human Scanpaths during Reading. In *Proceedings of the ACM on Human-Computer Interaction*, 162, New York, NY.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, LA.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do Transformer Models Show Similar Attention Patterns to Task-Specific Human Gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309, Dublin, Ireland.

Elaine Francis. 2021. The Problem of Gradient Acceptability. In *Gradient Acceptability and Linguistic Theory*. Oxford University Press, Oxford, United Kingdom.

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.

Ana González-Garduño and Anders Søgaard. 2018. Learning to Predict Readability Using Eye-movement Data from Natives and Learners. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, New Orleans, LA.

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*, pages 10–18, Salt Lake City, UT.

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.

Nora Hollenstein and Lisa Beinborn. 2021. Relative Importance in Sentence Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 141–150, Online.

Nora Hollenstein and Ce Zhang. 2019. Entity Recognition at First Sight: Improving NER with Eye Movement Information. 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 1–10, Minneapolis, MN.

Holly Joseph, Simon Liversedge, Hazel Blythe, Sarah White, and Keith Rayner. 2009. Word Length and Landing Position Effects during Reading in Children and Adults. *Vision Research*, 49(16):2078–2086.

Samuel Kiegeland, David Reich, Ryan Cotterell, Lena Jäger, and Ethan Wilcox. 2024. The Pupil Becomes the Master: Eye-Tracking Feedback for Tuning LLMs. In *ICML 2024 Workshop on LLMs and Cognition*, Vienna, Austria.

Kyuyoung Kim, Ah Jeong Seo, Hao Liu, Jinwoo Shin, and Kimin Lee. 2024. Margin Matching Preference Optimization: Enhanced Model Alignment with Granular Feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13554–13570, Miami, FL.

Reinhold Kliegl, Richard Olson, and Bruce Davidson. 1982. Regression Analyses as a Tool for Studying Reading Processes: Comment on Just and Carpenter's Eye Fixation Theory. *Memory & Cognition*, 10(3):287–296.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41(5):1202–1241.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In *Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, Peter Liu, and Xuanhui Wang. 2025. LiPO: Listwise Preference Optimization through Learning-to-Rank. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2404–2420, Mexico City, Mexico.

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. A Cognition Based Attention Model for Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 462–471, Copenhagen, Denmark.

Angela Lopez-Cardona, Sebastian Idesis, Miguel Barreda-Ángeles, Sergi Abadal, and Ioannis Arapakis. 2025. OASST-ETC Dataset: Alignment Signals from Eye-tracking Analysis of LLM Responses. In *Proceedings of the ACM on Human-Computer Interaction*, 3, New York, NY.

Ángela López-Cardona, Carlos Segura, Alexandros Karatzoglou, Sergi Abadal, and Ioannis Arapakis. 2025. Seeing Eye to AI: Human Alignment via Gaze-Based Response Rewards for Large Language Models. In *The 13th International Conference on Learning Representations*, Singapore, Singapore.

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

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2016. Leveraging Cognitive Features for Sentiment Analysis. In *Proceedings of the 20th Conference on Computational Natural Language Learning*, pages 156–166, Berlin, Germany.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Proceedings of Advances in Neural Information Processing Systems*, New Orleans, LA.

Paul Prasse, David Reich, Silvia Makowski, Seoyoung Ahn, Tobias Scheffer, and Lena Jäger. 2023. SP-EyeGAN: Generating Synthetic Eye Movement Data with Generative Adversarial Networks. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, New York, NY. Association for Computing Machinery.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY.

Keith Rayner. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, 124(3):372–422.

Keith Rayner, Tessa Warren, Barbara Juhasz, and Simon Liversedge. 2004. The Effect of Plausibility on Eye Movements in Reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(6):1290–1301.

Elizabeth Schotter and Brian Dillon. 2025. A Beginner's Guide to Eye Tracking for Psycholinguistic Studies of Reading. *Behavior Research Methods*, 57(68).

Elizabeth Schotter and Annie Jia. 2016. Semantic and Plausibility Preview Benefit Effects in English: Evidence from Eye Movements. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 42(12):1839–1866.

Gideon Schwarz. 1978. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Ekta Sood, Fabian Kögel, Philipp Müller, Dominike Thomas, Mihai Bâce, and Andreas Bulling. 2023. Multimodal Integration of Human-Like Attention in Visual Question Answering. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2648–2658, Vancouver, Canada.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020a. Interpreting Attention Models with Human Visual Attention in Machine Reading Comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 12–25, Online.

Ekta Sood, Simon Tannert, Philipp Mueller, and Andreas Bulling. 2020b. Improving Natural Language Processing Tasks with Human Gaze-Guided Neural Attention. In *Proceedings of Advances in Neural Information Processing Systems*, pages 6327–6341, Online.

Divyank Tiwari, Diptesh Kanojia, Anupama Ray, Apoorva Nunna, and Pushpak Bhattacharyya. 2023. Predict and Use: Harnessing Predicted Gaze to Improve Multimodal Sarcasm Detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15933–15948, Singapore, Singapore.

Andrea Tuninetti, Tessa Warren, and Natasha Tokowicz. 2015. Cue Strength in Second-language Processing: An Eye-tracking Study. *The Quarterly Journal of Experimental Psychology*, 68(3):568–584.

Shravan Vasishth, Titus von der Malsburg, and Felix Engelmann. 2013. What Eye Movements can Tell us about Sentence Comprehension. *WIREs Cognitive Science*, 4(2):125–134.

Sheng Wang, Liheng Chen, Jiyue Jiang, Boyang Xue, Lingpeng Kong, and Chuan Wu. 2024. LoRA Meets Dropout under a Unified Framework. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1995–2008, Bangkok, Thailand.

Alex Warstadt, Amanpreet Singh, and Samuel Bowman. 2019. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

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

Yang Zhao, Yixin Wang, and Mingzhang Yin. 2024. Ordinal Preference Optimization: Aligning Human Preferences via NDCG. *arXiv*, 2410.04346.

Appendix

A Implementation Details

The pretrained checkpoint was sourced from a publicly available Hugging Face repository. All models were trained on an NVIDIA GeForce RTX 4090 GPU. In all training configurations, we fine-tuned the 7 B instruction-tuned Mistral model with 4 bit weight quantisation; both policy and reference models were quantised. During training, we

applied parameter-efficient tuning and therefore updated only the LoRA parameters (rank r=16, α=32, dropout = 0.1, no bias). We optimised with the AdamW optimiser under a cosine schedule with a 10-step warm-up, batch size 8, and a maximum sequence length of 512 tokens.

We trained with three random seeds (17, 23, 42). Table 2 lists the hyperparameters explored in the grid search; the final setting was selected based on the lowest validation loss.

Table 2: Hyperparameter grid.

Hyperparameter	Values
Learning rate	2×10^{-6}, 3×10^{-6}, 5×10^{-6}
Weight decay	0.02, 0.03
Training steps	600, 700, 1000, 3000, 4700, 6120
β	0.2, 0.3
α	0.10, 0.05
Number of pairs	20, 30, 40

B Generation of Synthetic Eye Movement Data

To extend the eye movements dataset for training the model in the gaze-augmented DPO setting we generate the synthetic eye movements-while-reading data, particularly we predict the scanpaths for 30 sentence pairs from the CoLA dataset, preprocess the gaze data to extract the event count-based reading measures and train the models on both human and synthetic eye movements data.

Scanpath prediction is the task of mapping a tokenised sentence $x = (w_1, \ldots, w_T)$ to a variable-length sequence of eye-movement events $s = (e_1, \ldots, e_m)$, where each fixation event comprises the index $p_i \in \{1, \ldots, T\}$ of the fixated token. We formalise this as learning a conditional distribution $P(s \mid x; \theta)$, instantiated via autoregressive sequence models or structured prediction frameworks, by minimizing the negative log-likelihood:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{M} \log P(e_i \mid e_{<i}, x; \theta).$$

We used two corpora to train the Eyettention model: CELER (Berzak et al., 2022) and CoLAGaze (introduced above). CELER is a large-scale eye-tracking dataset comprising gaze recordings from 365 participants, including both native (L1) and non-native (L2) English speakers with varying levels of language proficiency and linguistic backgrounds. The participants read a total of 28,548 sentences, randomly sampled from Wall Street Journal (WSJ) newswire text. The dataset provides word-level fixation data, which we used to train the generative model of eye movements.

We generated synthetic fixation sequences for 30 CoLA training sentences using Eyettention, a dual-encoder Transformer for scanpath generation. We evaluated three training configurations: (i) pre-train on CELER (Berzak et al., 2022) and fine-tune on CoLAGaze, (ii) train on CELER only, and (iii) train on CoLAGaze only. Hyperparameters followed the original Eyettention setup. Each configuration used 5-fold cross-validation with the original "new sentence" split.

Training Data	**Fine-Tuning Data**	**Testing Data**	**NLD↓**
CELER	–	CoLAGaze$_{\text{all}}$	$0.493_{0.074}$
CoLAGaze	–	CoLAGaze$_{\text{all}}$	$0.487_{0.008}$
CELER	CoLAGaze	CoLAGaze$_{\text{all}}$	$0.491_{0.008}$
CELER	–	CoLAGaze$_{\text{ung}}$	$0.491_{0.014}$
CoLAGaze	–	CoLAGaze$_{\text{ung}}$	$0.484_{0.012}$
CELER	CoLAGaze	CoLAGaze$_{\text{ung}}$	$0.487_{0.017}$

Table 3: Eyettention training configurations and scanpath quality on CoLAGaze. Normalised Levenshtein Distance (NLD; lower is better) is reported as mean$_{SD}$ over readers. CoLAGaze$_{\text{all}}$ = all sentences; CoLAGaze$_{\text{ung}}$ = ungrammatical subset.

Performance was measured on a held-out CoLAGaze subset using normalised Levenshtein distance (NLD) between synthetic and human scanpaths: for each sentence–reader pair we computed the Levenshtein distance, divided it by the maximum scanpath length, and then averaged across readers. We report results for all sentences and for the ungrammatical subset. The three configurations performed similarly; the CoLAGaze-only model was marginally better on both subsets (Table 3). We therefore used this model to generate synthetic scanpaths. From these scanpaths we extracted the same event-count features as for human data, using the identical preprocessing pipeline, and integrated them into the DPO training pipeline.

C Reading Measures

To integrate human cognitive signals into the DPO framework, we extracted a diverse set of eye-tracking measures that capture different aspects of on-line reading behaviour. These measures reflect temporal and spatial dynamics of eye movements and have been shown in psycholinguistic research to be sensitive to lexical and syntactic properties of text. We report them in Table 4.

Reading Measure	Definition
Second pass duration (IQR, mean)	sum of fixation durations when a word is revisited after the first pass reading is complete, before the third pass
Go past time (mean, SD, IQR)	sum of all fixation durations from the first fixation on a word until the reader moves to a word to the right (progresses forward in the text)
First duration (median, IQR)	duration of the first fixation on a word, regardless of whether it was fixated in the first pass or not
Rereading time	duration of all fixation after the first pass
Gaze duration (SD, median, mean)	of all fixation durations on a word during first pass reading (before the eyes leave the word for the first time)
Normalised outgoing regressions count (SD)	number of regressions initiated from a word normalised by the total number of progressive saccades in a sentence
Saccade length (median, SD)	absolute horizontal distance of a saccade, measured in number of characters
Regression rate	proportion of regressions out of total incoming and outgoing saccades
Reading duration	total time spent reading each item, normalized by sentence length
Total fixation duration (SD)	sum of all fixation durations on a word across all passes
First fixation duration (SD, IQR, mean)	duration of the first fixation on a word during the first pass
Saccade duration (SD, IQR)	saccade duration in milliseconds
Normalised saccade duration (IQR)	saccade duration normalized by total reading time
Word in Fixed Context First and Total Fixation Duration (mean)	first and total fixation duration on a word in a fixed context (see Berzak et al., 2018 for more details) normalised by the context overall reading duration
Information Cluster First and Total Fixation Duration (mean, SD)	first and total fixation duration on a word in an information cluster (see Berzak et al., 2017 for more details) normalised by the cluster overall reading duration
Syntactic Cluster Total Fixation Duration (mean)	total fixation duration on a word in a syntactic cluster (see Berzak et al., 2018 for more details) normalised by the cluster overall reading duration

Table 4: Eye-tracking measures employed to augment DPO framework. For each measure we report its definition and the aggregation statistic(s) used to obtain a sentence-level vector (mean/median/SD/IQR).

Predicting Total Reading Time Using Romanian Eye-Tracking Data

Anamaria Hodivoianu[§] **Oleksandra Kuvshynova**[§]
Filip Popovici[⋆] **Adrian Luca**[⋆] **Sergiu Nisioi**[§*]

[§] Faculty of Mathematics and Computer Science
[⋆] Faculty of Psychology and Education Sciences
University of Bucharest
anamaria.hodivoianu@gmail.com
oleksandra.kuvshynova@unibuc.ro
sergiu.nisioi@unibuc.ro

Abstract

This work introduces the first Romanian eye-tracking dataset for reading and investigates methods for predicting word-level total reading times. We develop and compare a range of models, from traditional machine learning using handcrafted linguistic features to fine-tuned Romanian BERT architectures, demonstrating strong correlations between predicted and observed reading times. Additionally, we propose a lexical simplification pipeline that leverages these TRT predictions to identify and substitute complex words, enhancing text readability. Our approach is integrated into an interactive web tool, illustrating the practical benefits of combining cognitive signals with NLP techniques for Romanian, a language with limited resources in this area.

1 Introduction

Total Reading Time (TRT) refers to the cumulative duration a reader fixates on a given word, including all refixations. As an eye-tracking metric, TRT serves as a reliable indicator of the cognitive processing involved in both semantic and deep syntactic analysis during reading (Frazier and Rayner, 1982; Pickering et al., 2004). Unlike other reading-time metrics that may capture only initial attention, TRT reflects the full depth of engagement a word receives, offering valuable insight into processing difficulty.

The prediction of word-level reading times and their relationship to textual complexity have a long history of investigations. Previous studies demonstrate that models designed to estimate eye-tracking measures, such as first fixation duration and total reading time, can serve as effective indicators of text readability (González-Garduño and Søgaard, 2017). Furthermore, eye-tracking data has been used to improve neural network models; for example, Barrett et al. (2018) incorporate human attention patterns into recurrent neural networks, resulting in improved performance on a range of NLP tasks. More recently, research by Hollenstein et al. (2021) shows that large language models, including multilingual BERT, can approximate human reading behavior, supporting the integration of cognitive signals into language model development and evaluation. Additionally, it has been observed that transformer models inherently encode eye-tracking information during pre-training (Dini et al., 2025), and that intermediate fine-tuning with eye-tracking data does not negatively impact downstream task performance.

In this paper, we present a work-in-progress and several initial experiments on predicting word-level TRT using eye-tracking data collected from native Romanian speakers. Our work introduces the first dataset of Romanian eye tracking recordings collected in the framework of MultiplEYE[1] and we propose a variety of machine learning approaches to estimate TRT. All code is publicly available[2].

Accurate TRT prediction can inform a range of downstream applications, particularly in the development of cognitively informed tools such as lexical simplification systems and reading aids (Duffy et al., 1988).

2 Data

The dataset used in this study originates from the MultiplEYE project (Jakobi et al., 2025), and it represents the first eye-tracking corpus for reading in the Romanian language. It includes recordings from four participants, all of whom are native Romanian speakers.

[§] Corresponding authors.

[1] https://multipleye.eu/
[2] https://github.com/ana0101/eye-tracking

Proceedings of First International Workshop on Gaze Data and NLP associated with RANLP 2025, pages 71–75, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-104-2-009

The reading materials consist of 14 texts: 10 main texts, 2 practice texts, and 2 backup texts. These texts span a variety of genres, the majority being official Romanian translations from multiple source languages. Due to minor translation inconsistencies and updates across sessions, each participant read a slightly different version of the texts. The eye-tracking experiments were conducted using the EyeLink 1000 Plus system.

We process the raw gaze data using the *Pymovements* library (Krakowczyk et al., 2023) to extract fixations and their alignment to corresponding words in the text. For each word, the total reading time is computed as the average duration across all participants. To analyze how much the TRT varies between the participants, we calculate the coefficient of variation as the mean TRT divided by the standard deviation of the TRT. The coefficients are between 0 and 2, with a mean of 1.02 and a median of 0.98. The variation is quite high, which is expected given the small number of participants.

Figure 1 presents a histogram of the resulting, averaged TRT values. A significant number of words received a reading time of zero milliseconds, indicating that these words are skipped entirely during reading. This is a known and expected phenomenon, especially for short or high-frequency function words. Out of the 778 skipped words by all the participants, 689 are function words. At the opposite end of the spectrum, some examples of the words with the highest TRTs are *distorsiune* (distortion), *cosmodromică* (cosmodromic), *gravifice* (gravitational), and *premergătoare* (preliminary), which are all long, complex words. All TRT values were standardized to have a mean of 0 and a standard deviation of 1.

3 Results

We evaluate our models for predicting word-level TRT using several metrics: Mean Squared Error (MSE), R^2 score, Pearson and Spearman correlation coefficients, and Accuracy. Accuracy is defined as $100 - \text{MAE}$, where MAE is the Mean Absolute Error, with TRT values scaled to the $[0, 100]$ range, following established practices in eye-tracking prediction (Hollenstein et al., 2021).

We consider two primary modeling approaches: (1) traditional machine learning models trained on handcrafted features, and (2) fine-tuning pretrained BERT models.

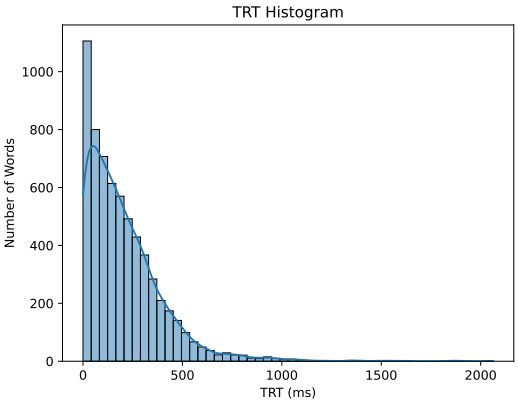

Figure 1: Histogram of TRT.

3.1 Feature Extraction and Analysis

For each word, we extract several features to aid in predicting the TRT. These include basic scalar attributes such as word length, the number of subword tokens generated by the Romanian BERT tokenizer (Dumitrescu et al., 2020), word frequency (obtained via the `wordfreq` library (Speer, 2022)), and the log probability of the word within its sentence context, estimated using a masked language modeling approach.

To calculate the log probability, we employ the pre-trained Romanian BERT model (Dumitrescu et al., 2020). The process involves first tokenizing the target word to determine its number of subword tokens. Then, these tokens are replaced in the sentence by an equal number of `[MASK]` tokens. The masked sentence is passed through the language model, which outputs probability distributions for each `[MASK]` token. The log probability for the word is taken as the negative logarithm of the probability assigned to the original first token by the model. While we also experiment with summing or averaging the log probabilities across all subword tokens, using only the first token's log probability yields better predictive performance.

In addition to scalar features, we derive contextual embeddings to capture semantic and syntactic information. We extract these embeddings from multiple layers of Romanian BERT: from the first, middle, last, and an average of all layers. Since BERT tokenizes words into subword units, we aggregate the embeddings of all tokens belonging to the same word by averaging them. This aggregation relies on character offset alignments to accurately map subword tokens back to their original words.

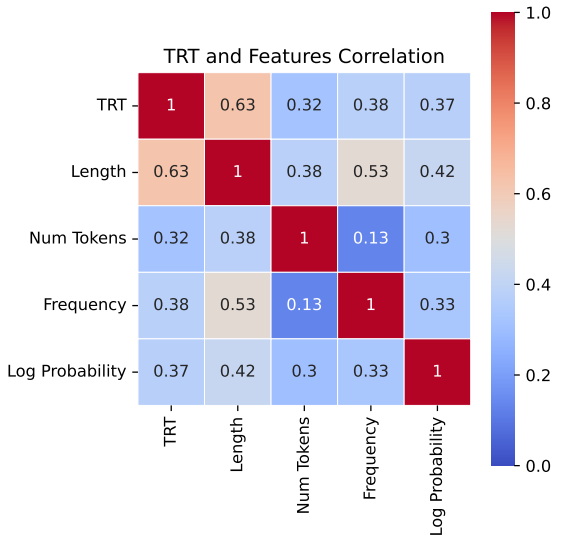

Figure 2: Pearson correlation between TRT and extracted features.

Figure 2 presents the Pearson correlation coefficients between TRT and the scalar features. Among these, word length shows the strongest correlation, followed by frequency, log probability, and number of tokens. Notably, the negative correlation between word length and frequency suggests that longer words tend to appear less frequently.

3.2 Traditional Regression Models

We train several regression models, including Linear Regression, Support Vector Regression, Random Forest, Gradient Boosting, Ridge Regression, Neural Networks, and more. Each model is trained on three feature sets: (1) only scalar features, (2) only embeddings, and (3) a combination of both. The data is split into train and test sets, with 80% of the data used for training and 20% for testing, which means 3796 words for training and 949 words for testing. When splitting the data, the words from the same sentence are not present in both the train and test sets. The features and reading times are standardized to have zero mean and unit variance.

All models achieve similar results: Pearson correlations between 0.6 and 0.7, accuracy between 70% and 95%, MSE between 0.4 and 0.7, Spearman correlation between 0.6 and 0.75, and R^2 scores ranging from 0.25 to 0.5.

Models trained on scalar features slightly outperform those trained on embeddings alone, although combining both types yields the best results overall. Among the embeddings, the average of all BERT layers generally performs best, so only the results with these embeddings were considered.

3.3 Fine-Tuning Pre-trained Language Models

We also fine-tune two Romanian BERT-based architectures:

- **BERT for Token Classification**: Modified to output a single regression value per token.
- **BERT with Regression Head**: Includes a linear layer, ReLU activation, layer normalization, dropout, and a final regression layer.

For token-level prediction, the TRT value of a word is assigned to each of its subtokens. During inference, token-level predictions are averaged to compute the word-level TRT.

The data is split into train, validation, and test sets, with 80% of the data used for training, 10% for validation, and 10% for testing. The train set contains 250 sentences, while the validation and test sets contain 25 sentences each. The reading times are standardized to have to have zero mean and unit variance.

Training is done in three phases using a gradual unfreezing strategy. For the first model, we unfreeze 4 additional layers every 8 epochs; for the second, we begin with only the regression head for 5 epochs and then unfreeze 6 layers every 10 epochs. Both models use the AdamW optimizer with a learning rate of 10^{-4}, weight decay of 10^{-4}, a cosine learning rate scheduler with warmup, batch size of 8, and dropout of 0.3. Padding tokens are ignored in the loss computation.

Table 1 summarizes the results. Both models perform comparably, achieving Pearson correlations around 0.7, Spearman correlations around 0.73, MSE near 0.5, and accuracy close to 90%, results that are similar to one of the best-performing traditional models, a neural network trained on all features.

4 Discussion

Our experiments demonstrate that predicting word-level reading times is feasible using both straightforward approaches, such as linear regression based on easily interpretable features like word length and frequency, as well as more sophisticated methods involving fine-tuning transformer-based language

Model	MSE	R^2	Pearson	Spearman	Accuracy
BERT for Token Classification	0.52	0.46	0.70	0.73	89.81
BERT with Regression Head	0.49	0.47	0.69	0.73	90.58
Neural network (all features)	0.41	0.44	0.69	0.74	90.43

Table 1: Performance of models.

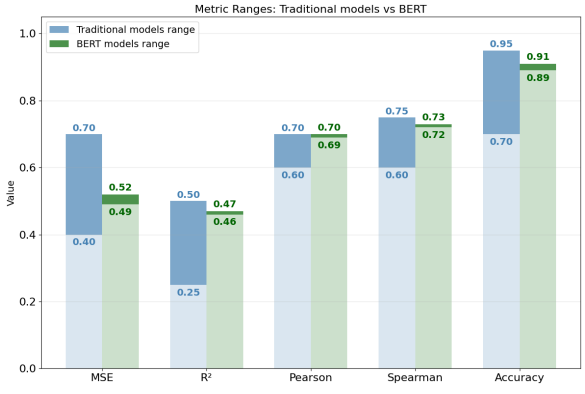

Figure 3: Metric ranges comparison between traditional models and BERT models.

models. The strong correlation between reading times and these features confirms their significance in capturing cognitive processing effort. Figure 3 illustrates the compared results of both methods.

One compelling application of accurate reading time prediction lies in lexical simplification. Since TRT effectively reflects the processing difficulty of words, it serves as a reliable indicator of lexical complexity. By identifying words with high TRT and substituting them with alternatives predicted to have lower TRT, we can enhance text readability and reduce overall reading effort.

To realize this, we implement a lexical simplification pipeline that first estimates the TRT for all words in a given text, selects candidates with elevated TRT, and generates potential replacements using the Romanian BERT masked language model (Dumitrescu et al., 2020). By masking the target word and leveraging the model's contextual predictions, we produce candidate substitutions. Inspired by Qiang et al. (2020), we experimented with concatenating the original and modified sentences in different orders to improve candidate quality, finding comparable improvements from both strategies. Before computing the predicted TRT for the candidates, we first make sure that the original word and the candidate are the same part of speech.

To make these capabilities accessible, we developed a user-friendly web interface called *Reading Time Estimator*. This tool enables users to input text, visualize predicted reading times on a word-by-word basis, and interactively simplify complex words by selecting suitable replacements with lower predicted TRT.

Overall, our work highlights the practical benefits of integrating cognitive signals such as eye-tracking data into NLP applications, particularly for languages like Romanian that have limited resources. A

5 Conclusions

In this paper, we introduced the first Romanian eye-tracking dataset focused on reading behavior, and demonstrated its utility in predicting word-level total reading time using both traditional machine learning and fine-tuned transformer-based models. Our experiments show that features such as word length and frequency are strong predictors of TRT, and that fine-tuned Romanian BERT models can achieve high predictive performance.

We also explored the practical implications of reading time prediction in the context of lexical simplification, proposing a pipeline that uses TRT estimates to identify and replace complex words. This system is implemented in an interactive web application that showcases the potential for user-centered NLP tools grounded in human reading behavior.

Our results highlight the value of eye-tracking data for advancing human-centered language technologies and pave the way for further work on Romanian and other low-resource languages in the domain of cognitive NLP.

Acknowledgments

This work was partially funded by the Romanian National Research Council (CNCS) through the Executive Agency for Higher Education, Research, Development and Innovation Funding (UEFISCDI) under grant PN-IV-P2-2.1-TE-2023-2007 (InstRead), and is supported by COST Action MultiplEYE, CA21131.

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, Brussels, Belgium. Association for Computational Linguistics.

Luca Dini, Lucia Domenichelli, Dominique Brunato, and Felice Dell'Orletta. 2025. From human reading to NLM understanding: Evaluating the role of eye-tracking data in encoder-based models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17796–17813, Vienna, Austria. Association for Computational Linguistics.

Susan A Duffy, Robin K Morris, and Keith Rayner. 1988. Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, 27(4):429–446.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online. 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.

Ana Valeria González-Garduño and Anders Søgaard. 2017. Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Deborah Noemie Jakobi, Maja Stegenwallner-Schütz, Nora Hollenstein, Cui Ding, Ramune Kaspere, Ana Matić Škorić, Eva Pavlinusic Vilus, Stefan Frank, Marie-Luise Müller, Kristine M Jensen de López, Nik Kharlamov, Hanne B. Søndergaard Knudsen, Yevgeni Berzak, Ella Lion, Irina A. Sekerina, Cengiz Acarturk, Mohd Faizan Ansari, Katarzyna Harezlak, Pawel Kasprowski, Ana Bautista, Lisa Beinborn, Anna Bondar, Antonia Boznou, Leah Bradshaw, Jana Mara Hofmann, Thyra Krosness, Not Battesta Soliva, Anila Çepani, Kristina Cergol, Ana Došen, Marijan Palmovic, Adelina Çerpja, Dalí Chirino, Jan Chromý, Vera Demberg, Iza Škrjanec, Nazik Dinçtopal Deniz, Dr. Inmaculada Fajardo, Mariola Giménez-Salvador, Xavier Mínguez-López, Maroš Filip, Zigmunds Freibergs, Jéssica Gomes, Andreia Janeiro, Paula Luegi, João Veríssimo, Sasho Gramatikov, Jana Hasenäcker, Alba Haveriku, Nelda Kote, Muhammad M. Kamal, Hanna Kundefineddzierska, Dorota Klimek-Jankowska, Sara Kosutar, Daniel G. Krakowczyk, Izabela Krejtz, Marta Łockiewicz, Kaidi Lõo, Jurgita Motiejūnienundefined, Jamal A. Nasir, Johanne Sofie Krog Nedergård, Ayşegül Özkan, Mikuláš Preininger, Loredana Pungă, David Robert Reich, Chiara Tschirner, Špela Rot, Andreas Säuberli, Jordi Solé-Casals, Ekaterina Strati, Igor Svoboda, Evis Trandafili, Spyridoula Varlokosta, Mila Vulchanova, and Lena A. Jäger. 2025. Multipleye: Creating a multilingual eye-tracking-while-reading corpus. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, ETRA '25, New York, NY, USA. Association for Computing Machinery.

Daniel G. Krakowczyk, David R. Reich, Jakob Chwastek, Deborah N. Jakobi, Paul Prasse, Assunta Süss, Oleksii Turuta, Paweł Kasprowski, and Lena A. Jäger. 2023. pymovements: A python package for processing eye movement data. In *2023 Symposium on Eye Tracking Research and Applications*, ETRA '23, New York, NY, USA. Association for Computing Machinery.

Martin J Pickering, Steven Frisson, Brian McElree, and Matthew J Traxler. 2004. Eye movements and semantic composition. In *The on-line study of sentence comprehension*, pages 33–50. Psychology Press.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lsbert: A simple framework for lexical simplification.

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

Author Index