Lending Eyesight to Language Models: Modeling and Probing Human Scanpath through Transformer Decoder

Junlin Li[1,2], David R. Reich[3,4], Yu-yin Hsu[1],
[1]The Hong Kong Polytechnic University, [2]MBZUAI *,
[3]University of Zurich [4]University of Potsdam,
junlin.li@mbzuai.ac.ce
davidrobert.reich@uzh.ch
yu-yin.hsu@polyu.edu.hk

Abstract

Human scanpaths offer rich and reliable clues about the cognitive mechanisms underlying language comprehension. Decoder-only language models, typically large language models (LLMs), have proven to exhibit striking parallels with human cognitive processes. In this study, we investigate to what extent language models can be endowed with human-like gaze shifts. Besides, by probing scanpath through eye model, analogous to probing language through language models, we ask whether such modeling can yield novel knowledge of the cognitive machinery of sense making.

This study presents a novel "plug-in" module, EyeLM, to transform an autoregressive language model into an autoregressive eye model, thus facilitating a probabilistic spatial modeling of human explicit attention. Our EyeLM module, powered by LLMs, achieves competitive performance with novel cognitive probing capabilities. By probing EyeLM, we can reach the predictability and uncertainty of the scanpath. Exhibiting aligned patterns with prior knowledge about human reading comprehension, these probabilistic measures of scanpath act as promising predictors of human comprehension skills. [1]

1 Introduction

Understanding a sentence is never achieved in just a single glance, as human attention, memory, and visual acuity are essentially limited (Reichle, 2011). To achieve adequate comprehension, readers have to shift their attention across serially presented linguistic units, producing a sequence of gazes and saccades, commonly referred to as a "scanpath" (von der Malsburg et al., 2012).

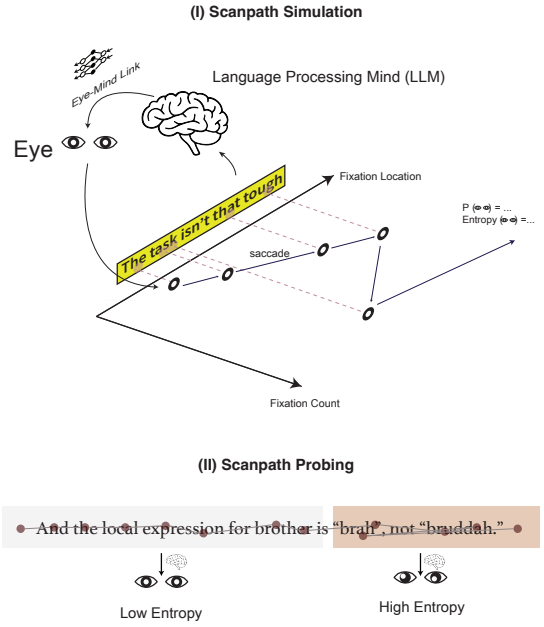

Figure 1: Simulating and probing human scanpath with LLMs

Central to scanpath modeling is the planning and execution of saccades. Saccades refer to the rapid, simultaneous movements of both eyes that allow the gaze to jump between different points in reading. Computational models of reading, such as E-Z Reader (Reichle et al., 2003) and SWIFT (Engbert et al., 2005), ascribe outgoing saccades, at least in part, to the processing difficulty of the currently fixated word and to higher-order integrative demands. For example, difficult words typically result in shorter outgoing (progressive) saccades and more frequent refixations (Liu et al., 2020). Models assuming more parallel lexical processing, including SWIFT and the more recent OB-1 Reader (Snell et al., 2018), further associate saccade generation with the competitive activation in the foveal and parafoveal fields (Richter et al., 2006). Harder and unpredictable words give rise to larger processing

*This work was begun when the first author was visiting University of Zurich as a visiting student from the Hong Kong Polytechnic University. His new affiliation is MBZUAI.

[1]Our code is available at http://github.com/CN-Eyetk/EyeLM.

Findings of the Association for Computational Linguistics: ACL 2026, pages 31807–31819

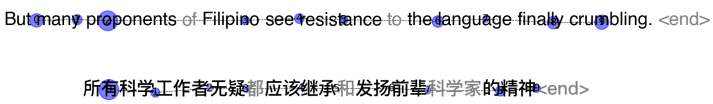

Figure 2: Two cases of scanpath simulated by `TinyLlama/TinyLlama-1.1B-Chat-v1.0`

Figure 3: An example of how lexical competition in parafoveal affects scanpath generation.

rates and inhibit the processing of their neighbors, leading to the "parafovea-on-fovea" effect of yet-to-be fixated words (Risse et al., 2014), as shown in Figure 3.

Inspired by these computational accounts, recent years have witnessed a growing interest in synthesizing human scanpaths using neural network models, especially for eye-movement-while-reading (Deng et al., 2023; Bolliger et al., 2023). These methods outperform traditional computational models (e.g., E-Z Reader) in terms of their ability to approximate human scanning behavior (Bolliger et al., 2025). More recently, decoder-only language models, with their internal or external probabilistic measurements, exhibit striking alignment with the human reading data (Oh et al., 2022; Salicchi et al., 2023; Wilcox et al., 2023; Wiechmann et al., 2022; Hale, 2001; Levy, 2008; Kuribayashi et al., 2025; Oh and Schuler, 2023; Nair and Resnik, 2023; Ding et al., 2025). Consequently, increasing attention has been paid to the effects of scale and processing depth (i.e., layer depth) on the cognitive plausibility of LLMs (Kuribayashi et al., 2025; Oh and Schuler, 2023).

Since LLMs' prior knowledge of human language processing has become increasingly eye-catching, it is naturally tempting to explore the following question:

- **RQ1:** To what extent can decoder-only language models, typically LLMs, simulate human scanpaths?

From discrete to probabilistic measures, decoder-based autoregressive modeling of language has substantially advanced our understanding of human language processing. Eye gaze, likewise, has been broadly compared to, and sometimes even counted as a manifestation of human language processing. However, as human scanpaths are still broadly represented using scalar or discrete measures, a probabilistic view of scanpaths potentially broadens our understanding of eye-movements while-reading. Hence, our second research question is:

- **RQ2**: Does the probabilistic measure of scanpaths, such as predictability and uncertainty, enable deeper insights into human mental processing states and reading comprehension abilities?

The contributions of the current paper are threefold:

- We propose **EyeLM**$_{Decoder}$, an eye-movement prediction head that decodes scanpaths from a decoder-only language model. This method tailors Pointer Networks (Vinyals et al., 2015) into a decoder-only mechanism and demonstrates compatibility with eye-movement-while-reading.

- Fitting **EyeLM**$_{Decoder}$ on naturalistic reading data, we introduce probabilistic measures of scanpath predictability and uncertainty. These measures strongly correlate with human gaze behavior and comprehension abilities.

2 Related Work

2.1 Computational Models of Reading

Eye movements offer profound insights into the cognitive processes that are engaged during reading (Rayner et al., 2012). Cognitive processing models, such as the E-Z Reader model (which assumes the serial allocation of attention) (Reichle and Sheridan, 2015) and the SWIFT model (which assumes parallel attention processing) (Engbert et al., 2005), have been proposed to account for the multi-staged information processing underlying reading comprehension. Specifically, the E-Z Reader model ascribes saccade targeting mainly to the familiarity check and lexical processing of the current word, which means that the word properties of the saccade target bear little importance in the targeting itself (Reichle et al., 2003). In contrast, the SWIFT model assumes lexical competition among words in the activation field (Nuthmann and Engbert, 2009), thus linking saccade targeting to the activation strength of upcoming words.

2.2 Machine learning models

With the advancement of machine learning and deep networks, statistical and neural modeling systems for eye-movement behaviors have attracted increasing attention. Remarkably, transformer-based scanpath prediction systems, powered by autoregressive generation paradigms (Deng et al., 2023) or diffusion processes (Bolliger et al., 2023), have achieved high accuracy in synthesizing human-aligned sentence-reading scanpaths. Specifically, Deng et al. (2023) formulates saccade targeting as a classification problem over saccade direction and distance, whereas Bolliger et al. (2023) treats saccade targeting as the denoising objective of a discrete diffusion process. These recent advances highlight the potential of transformer-based language models for probing the cognitive mechanisms underlying human reading comprehension. More recently, masked language modeling shows promising performance in reconstructing the human scan path (Sood et al., 2025).

2.3 Language Models as Cognitive Models of Reading

A growing body of work has used language model-derived norms as predictors of human eye-movement behaviors, leading to a significant enhancement in the accuracy of reading behavior modeling. The most widely-adopted language model measures include surprisal (Oh et al., 2022; Salicchi et al., 2023), contextual entropy (Wilcox et al., 2023; Wiechmann et al., 2022), attention scores (Oh and Schuler, 2022), and semantic representation (Salicchi et al., 2023). Among these, surprisal quantifies the amount of unexpected information conveyed by a word in context (Hale, 2001; Levy, 2008), and has been shown to closely correlate with both early (Frank and Thompson, 2012) and higher-order processing effects (Demberg and Keller, 2019; Rajkumar et al., 2016). Recent studies have begun to focus on a novel debate surrounding whether scaling transformer decoders enhances or degrades their ability to simulate human reading times (Kuribayashi et al., 2025; Oh and Schuler, 2023; Nair and Resnik, 2023). Despite mixed findings and close scrutiny of their internal reactions to human languages, it has been revealed that large language models still encode rich clues of cognitive plausibility inside their inner layers (Kuribayashi et al., 2025).

3 Problem Formulation

This section presents the formal formulation of an autoregressive position-wise retrieval model for scanpath generation. At the core of this formulation, we serially predict a position-wise activation distribution, taking the current scanpath state as the query and the word tokens in stimuli as the key.

Our method can be viewed as a decoder-only adaptation of Pointer Networks (Ptr-Net) (Vinyals et al., 2015), an architecture designed to produce neural solutions to combinatorial optimization problems such as the Travelling Salesman Problem (TSP).

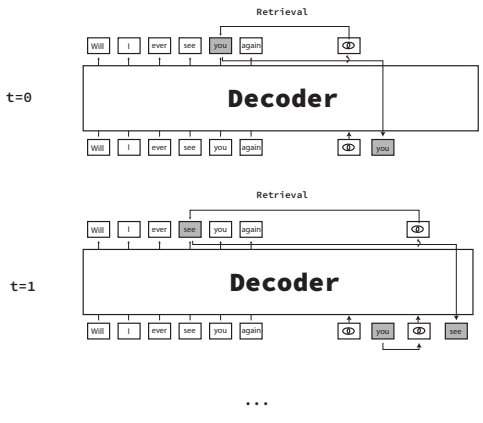

Figure 4: An autoregressive self-retrieval model, based on Transformer Decoder, generates scanpath.

Stimulus and Scanpath Representation Let

$$X_{1:m} = \langle x_1, x_2, \ldots, x_m \rangle$$

denote a naturalistic reading stimulus consisting of m word tokens in linear order. Each word x_j is encoded by a stimulus encoder into a contextualized representation

$$\mathbf{h}_j^x \in R^d, \quad j = 1, \ldots, m$$

and we denote the stacked stimulus representations by

$$\mathbf{H}^X = [\mathbf{h}_1^x, \mathbf{h}_2^x, \ldots, \mathbf{h}_m^x] \in R^{m \times d}$$

[2]

Let

$$y_{1:n} = \langle y_1, y_2, \ldots, y_n \rangle$$

denotes a scanpath composed of n fixations, where each fixation corresponds to a word in X.

Apparently, we can derive a sequence of fixation indices to link scanpath to stimuli

$$I = \langle i_1, i_2, \ldots, i_n \rangle, \quad i_t \in \{1, \ldots, m\},$$

such that the indexing operation $X[i_t] = y_t$ retrieves the word fixated at time step t.

Autoregressive Scanpath State Given the fixation history up to time step $t-1$, a retrieval model f_θ constructs a scanpath state

$$\mathbf{h}_t^y = f_\theta\big(\mathbf{H}^X, y_{<t}\big) \in R^d$$

where $f_\theta(\cdot)$ is an autoregressive encoder (e.g., a transformer or recurrent module) that summarizes the previous fixation sequence $\mathbf{y}_{<t}$ together with the stimulus context $\mathbf{H}^X$. The vector $\mathbf{h}_t^y$ serves as a query representation characterizing the reader's current attentional state.

Computation of Graded Activation Field At time step t, the model assigns a score to each stimulus position $j \in \{1, \ldots, m\}$ via a query–key matching function:

$$s_t(j) = \mathrm{score}\big(\mathbf{h}_t^y, \mathbf{h}_j^x\big) = \frac{(\mathbf{h}_t^y)^\top \mathbf{h}_j^x}{\tau}$$

where $\mathbf{h}_j^x$ is the stimulus representation of word x_j and $\tau > 0$ is a temperature parameter.

[2]For stimuli words spanning multiple tokens, we take the mean representation based on each subtoken.

The scores are normalized across all stimulus positions to define a categorical distribution over fixation targets:

$$p_\theta(i_t = j \mid y_{<t}, x_{1:m}) = \frac{\exp(s_t(j))}{\sum_{k=1}^m \exp(s_t(k))} = \mathrm{softmax}_j\Big(\mathbf{h}_t^{y\top} \mathbf{H}^X\Big)$$

Autoregressive Generation of Scanpath through Retrieval At time step t, suppose the model retrieves the word at position j, we append the new fixation y_t, where $y_t = X[j]$, into the ongoing sequence of fixation as $Y_{<t+1}$, which continues to generate until the final fixation position.

Because fixation targets are predicted over the full index set $\{1, \ldots, m\}$, this formulation naturally accommodates word skipping ($i_{t+1} > i_t + 1$), regressions ($i_{t+1} < i_t$), and refixations ($i_{t+1} = i_t$) within a unified retrieval-based decision mechanism.

4 Method

This section presents the detailed implementation of the **EyeLM** method. To facilitate understanding, we first introduce the inference procedure and then explain the training procedure.

For clarity, we present an example of input-output pair in box 5

Input-Output example of EyeLM

Input: This is an example. < end >
Output: < eye > 0 This < eye > 0 This < eye > 2 an < eye > 3 example. < eye > < end >

Figure 5: An Example of input and output

4.1 Inference Procedure

Stimuli State We initialize the inference procedure with the stimuli X itself. We represent its hidden state through the Transformer Decoder:

$$\mathbf{H}^x = \texttt{Decoder}(X_{1:m}) \tag{1}$$

In particular, we set the final unit x_m as < end >. When < end > is retrieved, we terminate generating the scanpath.

Eye State At each timestep t, we append a special "Eye Token" < eye > as the final token, to function as the query for the retrieval process. The Transformer Decoder represents < eye > as the hidden state of the current scanpath h_t^y

$$\mathbf{h}_t^y = \texttt{Decoder}(X_{1:m}, y_{<t})[-1] \tag{2}$$

Autoregressive Retrieval The hidden state of eye token < eye > retrieves the fixation area with the highest activation through a dot product process:

$$s_t = (\mathbf{h}_t^y)^\top \cdot \mathbf{H}^X \in R^m, \quad (3)$$

$$y_t = X_{1:m}[\texttt{argmax}(s_t)]. \quad (4)$$

The label of y_t will be appended to the ongoing scan path $y_{<t}$.

4.2 Training Objective

We train the model by minimizing the negative log-likelihood of the gold fixation positions under the predicted retrieval distribution:

$$\mathcal{L}_{\text{ret}} = -\sum_{t=1}^{n} \log p_\theta(i_t \mid y_{<t}, X_{1:m}).$$

This position-wise retrieval loss directly supervises the model to assign a high probability mass to the word positions that are fixated on by human readers at each time step.

4.3 Scanpath Probing

With a spatial transition distribution $p(i_t \mid y_{<t}, X_{1:m}) \in R^m$, we take the logits of ground-truth fixation as its predictability. We also take its entropy as the uncertainty of this saccade:

$$\mathrm{H}(i_t) = -\sum_{i_t=1}^{m} p(i_t) \log p(i_t). \quad (5)$$

5 Experiment

To compare with existing systems, we conduct a comparative experiment on human reading scanpath data.

5.1 Datasets

To compare with most studies, we select the CELER L1 (*Corpus of Eye Movements in L1 and L2 English Reading*, using the L1 subset) (Berzak et al., 2022) and BSC (*Beijing Sentence Reading Corpus*) (Pan et al., 2021) datasets. Both datasets feature single sentence reading and native subjects. Descriptive statistics of the two datasets are in Appendix 2.

5.2 Selection of Language Models

For English, we select `gpt2` and `gpt2-medium` (Lagler et al., 2013). For Chinese, we select `gpt2-chinese` and `gpt2-chinese-medium` (Radford et al., 2019). For both languages, we additionally select `TinyLlama/TinyLlama-1.1B-Chat-v1.0` (Zhang et al., 2024), which is a Llama-backbone language model capable of processing both English and Chinese. All language models are licensed and accessible through `huggingface`.

5.3 Finetuning and Adapter Tuning

For gpt-based models, we evaluate two settings: (1) full finetuning, which means that all parameters, together with additional eye modules, are finetuned, and (2) frozen-layer finetuning, which means that all hidden layers are frozen. This is to evaluate the extent to which the fundamental model encodes prior knowledge about human language processing. For `TinyLlama`-based models, we only try adapter tuning with mixed precision. Specifically, adapters are injected into the query and value projections inside transformer hidden layers, while the eye modules are fully finetuned. More details about hyperparameter selection and training are available in Appendix A.2.

5.4 Evaluation

We follow a 5-fold cross-validation, splitting across both subjects and sentences into consideration. Concretely, we select 80% of the subjects and 80% of the sentences as the training set, leaving the remaining 20% of subjects and 20% of sentences as the test set. Thus, each test-fold comprises unseen subjects and unseen sentences.

5.5 Metrics and Reference Systems

In parallel with previous settings, we consider the normalized Levenshtein distance (NLD) and negative log likelihood (NLL) as evaluation metrics. Lower NLD and lower NLL suggest better performance.

We select `Eyettention` (Deng et al., 2023) and `ScanDL` (Bolliger et al., 2023, 2025) as reference systems. They are both generative scanpath models that are trained on human eye-movement-while-reading data, using the identical train-test split policy. We refer to their "*New Reader/New Sentence*" results.

5.6 Results

Table 1 presents the performance of different **EyeLM** systems in comparison with EYETTENTION and SCANDL. The results indicate the impressive performance of decoder language models in simulating human eye-movements while-

Dataset	CELER L1		BSC	
Metrics	NLL	NLD	NLL	NLD
Eyettention (Deng et al., 2023)	2.297 ± 0.011	0.568 ± 0.004	1.84 ± 0.017	0.545 ± 0.004
ScanDL (Bolliger et al., 2025)	-	0.515 ± 0.014	-	0.41 ± 0.01
EyeLM (gpt2)	1.852 ± 0.057	**0.484 ± 0.012**	1.302 ± 0.054	0.34 ± 0.01
- *frozen hidden layers*	1.823 ± 0.049	0.488 ± 0.009	1.614 ± 0.070 (↑)	0.654 ± 0.017 (↑)
EyeLM (gpt2 medium)	1.791 ± 0.037	0.491 ± 0.014	1.299 ± 0.075	0.370 ± 0.024
- *frozen hidden layers*	1.780 ± 0.043	0.49 ± 0.013	1.591 ± 0.060 (↑)	0.589 ± 0.026 (↑)
EyeLM (Tiny Llama)	**1.616 ± 0.028**	0.496 ± 0.014	**1.084 ± 0.050**	**0.331 ± 0.019**

Table 1: Results of *New Reader/New Sentence* evaluation using 5-fold train-test split. ↑ denotes a significant increase of NLD and NLL when finetuned without updating the transformer hidden layer parameters. For BSC, we use `gpt2-chinese-cluecorpussmall` base and medium models.

reading, suggesting the promising role of decoder-only LLMs in modeling human attention shifts in processing sentence stimuli.

We identify three interesting tendencies from Table 1:

Firstly, larger LMs function as better probabilistic models than smaller LMs, as larger models consistently outperform smaller models in minimizing negative log likelihood loss.

Secondly, pre-trained decoder-only models encode rich prior knowledge of human language processing, as we can see that the freezing of hidden layers exerts minimal effect on the performance of gpt models on CELER L1. Dramatically, we notice that Chinese reading (BSC) is hard for the layer-frozen model to fit, which potentially indicates that the pre-trained models lack language-specific knowledge of reading behavior.

Thirdly, the scaling up of language models does not improve simulating English reading as much as it does Chinese reading. This result highlights language type as a strong modifier of the scale-up effect on LLM cognitive plausibility.

6 Scanpath Probing Analysis

Comparable to probing analysis of language using language models, this section presents a probing analysis of scanpaths using our the scanpath model. We aim to answer the following questions:

- **Q1:** How does saccade planning across different languages affect word-level processing during sentence reading? Specifically, can we observe "**predictability effects**" or "**uncertainty effects**" in eye-movement behavior?
- **Q2:** How do the certainty and uncertainty of saccade planning reflect human comprehension skills? Do readers with higher and lower comprehension skills differ in terms of scanpath predictability and uncertainty?

For clarity, we use the logits of the ground-truth saccade j as a predictability measure, and the entropy of the saccade distribution over all interest areas as entropy.

6.1 Next Fixation Analysis

Focusing on fixation-level measurements, we investigate how saccade predictability and uncertainty exert an effect on the subsequent fixation, such as next fixation duration (NFD), next landing position (NLP), and next out-going saccade distance (NSCD)[3]. To answer Q1, we analyze the effects of the logits and entropy of saccade on these fixation measures.

According to the SWIFT reader, high activation (or high processing rate, which corresponds to high logits in our model) arises from increased word difficulty (Engbert et al., 2005). Besides, high processing load, according to the hypothesis of the global inhibition process, further suppresses the processing of words to the right (Schad et al., 2024). Based on this assumption, we hypothesize that:

- **H1.1** High predictability of saccades prolongs fixation duration, reduces latency period, and decreases saccadic completion duration.

As entropy is mostly negatively correlated with predictability, we assume that:

[3]All data are acquired from the official fixation report. We take the raw value of fixation duration from the official data with log transformation. The landing position is available in the raw BSC fixation report. For CELER L1, it is calculated as the difference between the fixation position on the x-axis and the left boundary of the corresponding interest area. For the out-going saccade, it is calculated as the "IA_INDEX" difference of the next fixation from the current fixation.

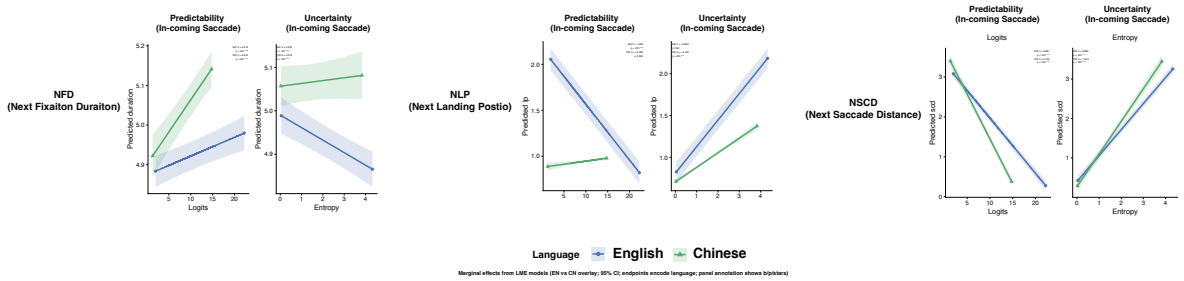

Figure 6: Correlation between saccade predictability/untertainty and gaze measures. We use Linear Mixed Effect Regression to measure the correlation. For example, we use the fomular "duration logits + position + length + (1 | subj_label) + (1 | sentence_id)" and "duration entropy + position + length + (1 | subj_label) + (1 | sentence_id)" to estimate the effect of logits and entropy on next fixation duration.

- **H1.2** High entropy reduces fixation duration, increases landing position, and increases outgoing saccades.

Predictability Effect Figure 6 displays how saccade predictability and entropy correlate with NFD, NLP, and NSCD in both English and Chinese reading. Specifically, high saccade predictability coexists with a longer next fixation duration (NFD), shorter next landing position (NLP) and shorter next outgoing saccade (NSCD), patterning the parallel lexical processing as depicted by SWIFT reader. The only exception is observed in the NLP regarding Chinese reading, which will be discussed in section 6.1.

The results regarding the predictability effect generally align with the predictions of the SWIFT reader model mentioned above, which suggests that our model reflects the graded parallel nature of human attention during reading.

Uncertainty Effect We also observe the expected uncertainty effect. As high uncertainty tends to coexist with shorter NFD, longer NLP, and longer NSCD. We still notice considerable exceptions in Chinese reading in terms of NLP, which will be discussed in section 6.1.

The uncertainty effect is not directly predictable from the SWIFT reader. Notwithstanding, our findings mostly align with the implications of the SWIFT reader.

Cross-lingual Divergence As we mentioned above, the next landing position (NLP) in English and Chinese reading shows an unequal response to saccade probability. While the tendency in English reading generally aligns with the SWIFT reader, we notice that predictable saccades in Chinese encourage users to look into the later part of the saccade target. This contradiction potentially reveals a stronger word length effect in Chinese if we accept the assumption that the fixation position linearly correlates with word length, based on the OVP (Optimal Viewing Position) hypothesis (O'Regan and Jacobs, 1992). Such a strong word length effect hints at a character-based processing mechanism in Chinese reading, especially the word-segmentation-in parafoveal hypothesis (Xie et al., 2025), namely, pre-lexical character-based processing in the parafoveal.

6.2 Correlation with Human Comprehension - A Case of L2 English Reading

To answer Q2, we examine how the probing analysis of human scanpaths predicts human comprehension skills. In other words, can saccade predictability and entropy reflect the comprehension capabilities of human readers, and thus be taken as a measure of human linguistic skills?

We take the L2 scanpath sampled from CELER L2 data as a hypothesis and calculate saccade entropy from the **EyeLM** model trained with L1 data fully-finetuned on `gpt2-medium` model, using the checkpoint with lowest NLD on L1 test set. The results suggest that saccade entropy serves as a reliable quantifier of L2 readers' reading comprehension capability (measured by Michigan VR and Michigan LG in CELER2 Data).

From Figure 7, it is apparent that L2 English proficiency ,indexed by MichiganLG ("Listening and Grammar sections of the Michigan Placement Test") and Michigan VR ("Vocabulary and Reading Comprehension") correlates with scanpath entropy. Figure 8 presents the paired-correlation among MichiganVR, MichiganLG, and saccade entropy, treating each individual subject as a sep-

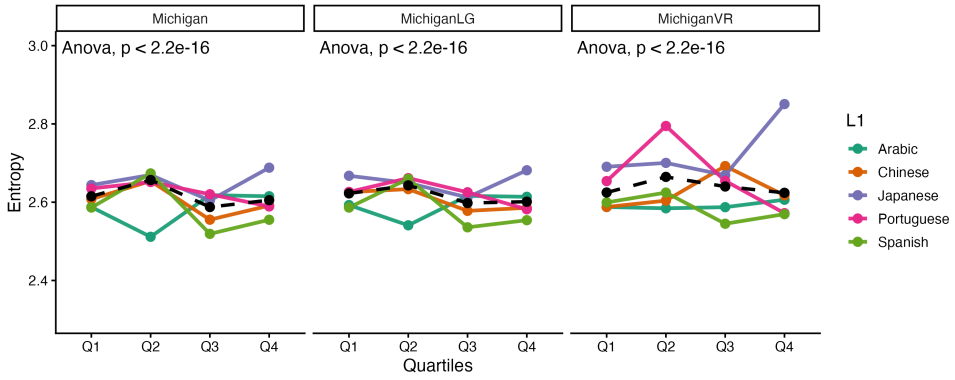

Figure 7: High Michigan test score generally leads to lower saccade entropy across L2 readers. (Michigan: The mean score of MichiganLG and MichiganVR, MichiganLG: Listening and Grammar sections of the Michigan Placement Test, MichiganVR: Vocabulary and Reading Comprehension sections.)

arate sample. The results indicate a considerable subject level correlation between L2 proficiency and model-generated probabilistic measures. High performance in reading and language tests is associated with lower entropy.

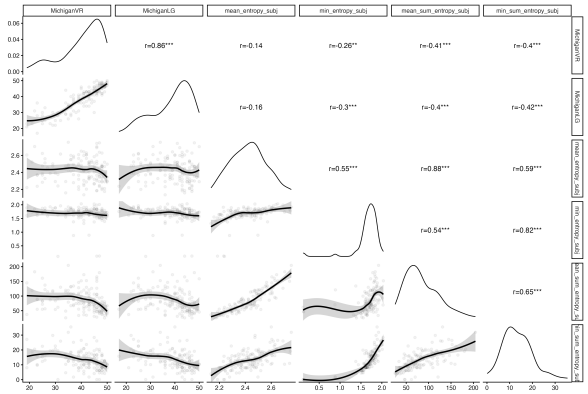

Figure 8: Saccade entropy predicts L2 English proficiency (mean_entropy_subj=The subject's average across the mean entropies (by fixation) of each trial. min_entropy_subj= The subject's min value across the mean entropies of each trial. mean_sum_entropy_subj= The subject's average across the total entropy of each trial. min_sum_entropy_subj= The subject's min value across the total entropy of each trial)

7 Case Studies

In Figure 9, we present a case of activation field dynamics in simulated scanpath. The activation spans multiple words close to the fovea and shifts dynamically to the right. In Figure 10, we visualize the activation field based on its correlation with the distance to the fovea. We notice a larger activation of the distal left field in Chinese reading, which aligns with the existing notion that regressions are more frequent in non-alphabetic languages (Chen et al., 2003; Liversedge et al., 2024).

8 Conclusion

This paper presents the first study that uses LLMs to model human activation fields and simulate scanpaths. It also proposes scanpath probing, which provides rich probabilistic measures that reveal the cognitive underpinnings of language comprehension and predict L2 proficiency.

Our scanpath simulation model achieves impressive performance in mimicking human attention shifts during sentence reading experiments. Besides, it offers effective measurements of scanpath fluency, such as predictability and uncertainty, from which we can validate existing language processing theories, delve into language diversity, and measure human comprehension capabilities.

9 Acknowledgement

We are grateful to the ARR reviewers for their thoughtful and constructive comments. The first author, when he started this work, was hosted by the University of Zurich as a visiting student from the Hong Kong Polytechnic University. We gratefully acknowledge the Computational Linguistics Department at the University of Zurich for making this productive visit possible. The second author was funded by the Swiss National Science Foundation under grant IZCOZ0_220330 (EyeNLG, PI: Lena Jäger). We are grateful to members of the Digital Linguistics Lab (DiLi Lab) at UZH, such as Prof. Lena Ann Jäger and Ms. Cui Ding, for their valuable suggestions and insightful comments on this work.

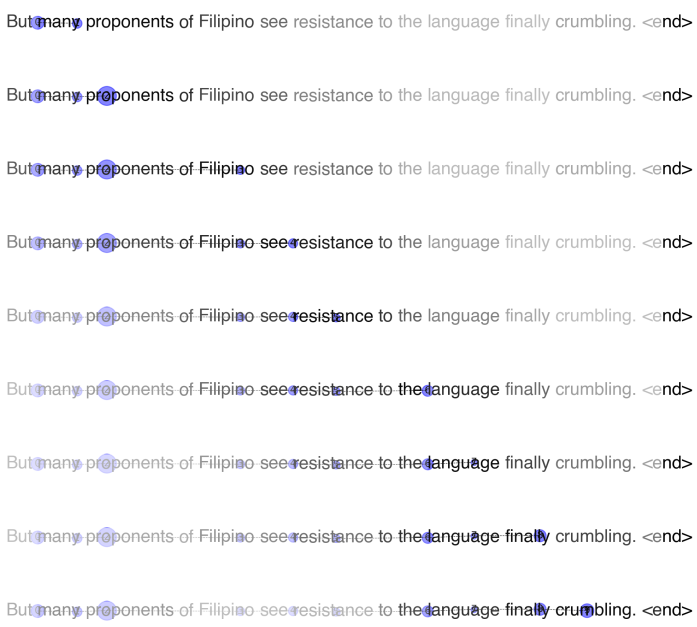

Figure 9: The dynamic change of horizontal activation field simulated by `TinyLlama/TinyLlama-1.1B-Chat-v1.0`. At each timestep, words receiving higher activations are colored darker. Larger scatter size denotes larger logits or predictability of the predicted fixation point.

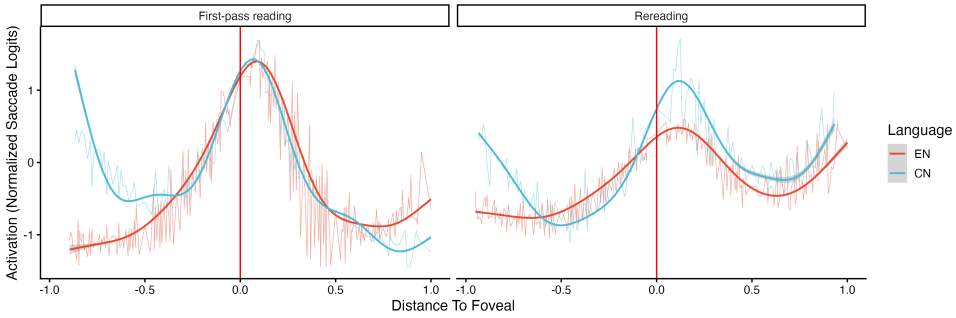

Figure 10: Activation Field: Eccentricity Effect in first-pass reading and re-reading

Limitations

We discuss the limitations of this paper as follows. Firstly, we only implement our study on English and Chinese sentence reading data, while reading itself can occur in various languages and complex discourse. Additionally, we did not incorporate subject information into the simulation process, which limits the generalizability of our model to human subject differences. Future studies may consider generalizing the proposed method to different languages and diverse data.

The second limitation is the training objective and evaluation metric. Initially, the proposed model, focused on spatial shifting, does not predict fixation duration and landing position, we cannot make use of the widely used scan-path measures, such as Multi-Match. Besides, the cognitive plausibility of machine learning models is still a tricky objective to evaluate. Multi-dimensional evaluation should be considered in future studies.

The third limitation is that we did not conduct a probing analysis on a broader set of data. Besides, the reading skill analysis is limited to L2 readers, as the CELER data do not include reading examination data for L1 readers. We expect future studies to extend our method to a wider diversity of reading data.

The fourth limitation is that the current model is still based on training, which means that the pre-trained weights of different language models are manipulated in approximation to human read-

ing data. We argue that this method allows us to achieve the upper bound of human-model alignment. However, the lower-bound alignment is still open to question. We hope future studies will develop training-free methods to address the problem of human-model alignment.

Finally, as a data-driven modeling study, we do not argue that our proposed model can serve as a theoretical reading model. Although it reproduces some assumptions of well-known sentence reading theories, we do not assert that it can outperform these theories in accounting for the cognitive underpinnings of human attention while reading. Neither do we assert that we successfully model all cognitive underpinnings of reading, which absolutely connect to multiple psychological and psycholinguistic constructs, such as memory, visual perception, and even problem solving.

Ethic Statement

Although this study does not recruit human subjects, several ethical issues should still be discussed. Firstly, the eye-movement data in this study includes human subject information. Although no harmful or identity-sensitive information is included in this database, there are still concerns regarding the data collection process. For example, readers must sit in front of the display for an extended period with controlled head movements. We assure that the eye-tracking data used were recorded using non-intrusive techniques compared to other data collection methods such as EEG and fMRI. All human subjects are compensated and recruited with informed consent. The experiments for data collection are approved by the ethics review board of concern.

Simulating human behavior is essentially risky. We do not use our model to conduct subject-sensitive modeling of human behavior. The central goal of this study is to discuss the human attention pattern that can be learned by autoregressive language models.

When writing this paper, the "Writefull" function in overleaf is activated, providing spelling checks and grammatical corrections.

References

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 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*.

Lena S Bolliger, David R Reich, and Lena A Jäger. 2025. Scandl 2.0: A generative model of eye movements in reading synthesizing scanpaths and fixation durations. *Proceedings of the ACM on Human-Computer Interaction*, 9(3):1–29.

Hsuan-Chih Chen, Hua Song, Wing Yin Lau, Kin Fai Ellick Wong, and Siu Lam Tang. 2003. Developmental characteristics of eye movements in reading chinese. *Reading development in Chinese children*, pages 157–169.

Vera Demberg and Frank Keller. 2019. Cognitive models of syntax and sentence processing. *Human language: From genes and brains to behavior*, pages 293–312.

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.

Cui Ding, Yanning Yin, Lena Ann Jäger, and Ethan Wilcox. 2025. Modeling bottom-up information quality during language processing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11720–11732.

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.

Stefan Frank and Robin Thompson. 2012. Early effects of word surprisal on pupil size during reading. In *Proceedings of the annual meeting of the cognitive science society*, volume 34.

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*.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *Transactions of the Association for Computational Linguistics*, 13:1743–1766.

Klemens Lagler, Michael Schindelegger, Johannes Böhm, Hana Krásná, and Tobias Nilsson. 2013. Gpt2: Empirical slant delay model for radio space geodetic techniques. *Geophysical research letters*, 40(6):1069–1073.

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

Zhifang Liu, Wen Tong, and Yongqiang Su. 2020. Interaction effects of aging, word frequency, and predictability on saccade length in chinese reading. *PeerJ*, 8:e8860.

Simon P Liversedge, Henri Olkoniemi, Chuanli Zang, Xin Li, Guoli Yan, Xuejun Bai, and Jukka Hyönä. 2024. Universality in eye movements and reading: A replication with increased power. *Cognition*, 242:105636.

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

Antje Nuthmann and Ralf Engbert. 2009. Mindless reading revisited: An analysis based on the swift model of eye-movement control. *Vision Research*, 49(3):322–336.

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5:777963.

Byung-Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9324–9334, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

J Kevin O'Regan and Arthur M Jacobs. 1992. Optimal viewing position effect in word recognition: A challenge to current theory. *Journal of Experimental Psychology: Human Perception and Performance*, 18(1):185.

Jinger Pan, Ming Yan, Eike M Richter, Hua Shu, and Reinhold Kliegl. 2021. The beijing sentence corpus: A chinese sentence corpus with eye movement data and predictability norms. *Behavior Research Methods*, pages 1–12.

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

Rajakrishnan Rajkumar, Marten Van Schijndel, Michael White, and William Schuler. 2016. Investigating locality effects and surprisal in written english syntactic choice phenomena. *Cognition*, 155:204–232.

Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr. 2012. *Psychology of reading*. Psychology Press.

Erik D Reichle. 2011. Serial-attention models of reading.

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

Erik D Reichle and Heather Sheridan. 2015. Ez reader: An overview of the model and two recent applications. *The Oxford handbook of reading*, pages 277–290.

Eike Martin Richter, Ralf Engbert, and Reinhold Kliegl. 2006. Current advances in swift. *Cognitive Systems Research*, 7(1):23–33.

Sarah Risse, Sven Hohenstein, Reinhold Kliegl, and Ralf Engbert. 2014. A theoretical analysis of the perceptual span based on swift simulations of the n+ 2 boundary paradigm. *Visual Cognition*, 22(3-4):283–308.

Lavinia Salicchi, Emmanuele Chersoni, and Alessandro Lenci. 2023. A study on surprisal and semantic relatedness for eye-tracking data prediction. *Frontiers in Psychology*, 14:1112365.

Daniel J Schad, Sarah Risse, Ralf Engbert, and Reinhold Kliegl. 2024. Individual differences during reading via process-based eye-movement modeling.

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.

Ekta Sood, Prajit Dhar, Enrica Troiano, Rosy Southwell, and Sidney K D'Mello. 2025. Scanez: Integrating cognitive models with self-supervised learning for spatiotemporal scanpath prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1132–1142.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in neural information processing systems*, 28.

Titus von der Malsburg, Shravan Vasishth, and Reinhold Kliegl. 2012. Scanpaths in reading are informative about sentence processing. In *Proceedings of the first workshop on eye-tracking and natural language processing*, pages 37–54.

Daniel Wiechmann, Yu Qiao, Elma Kerz, and Justus Mattern. 2022. Measuring the impact of (psycho-)linguistic and readability features and their spill over effects on the prediction of eye movement patterns. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5276–5290, Dublin, Ireland. Association for Computational Linguistics.

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

Fang Xie, Wanying Chen, Lei Zhang, Xiaohua Cao, and Kayleigh L Warrington. 2025. Exploring the role of word segmentation on parafoveal processing during chinese reading. *Journal of Cognitive Psychology*, 37(1):1–14.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

A Appendix: Implementation Details

A.1 Data Statistics

The descriptive statistics of datasets in this study are presented in Table 2.

A.2 Parameter Selection

Following SCANDL (Bolliger et al., 2023), we perform triple cross-validation and employ the normalized Levenshtein Distance to search for training parameters. For all gpt based models, we select a learning rate of `5e-4`, a batch size of `16` , and a training epoch of `10`. For tinyllama models, we select a learning rate of `1e-4`, a batch size of `1` , and a training epoch of `3`. The hidden dimension of all additional layers, such as the retrieval head, is consistent with the base model's dimension. For the optimizer, we use the official optimizer of `transformers.Trainer`. For the scheduler, we impose a warm-up of 100 steps and a linear decay policy after the warm-up.

A.3 Model size and time cost for training

The base model parameters for `gpt2` and `gpt2-chinese` is 117M, for `gpt2-medium` and `gpt2-chinese-medium` is 345M. The additional retrieval model, which includes two linear projection layers, requires $2 \times 768 \times 768$ additional parameters for `gpt2` and $2 \times 1024 \times 1024$ additional parameters for `gpt2-chinese`. The base model parameters for `TinyLlama-1.1B` include 1.1B parameters, for which the additional retrieval model requires $2 \times 2048 \times 2048$. The training procedure takes 10 to 20 minutes for gpt based model and around 40 minutes for tinyllama models.

B Appendix: Fixation Duration Modeling

Modeling fixation duration is a non-trivial objective of scanpath simulation. We tentatively conduct additional experiments by attaching the fixation duration head as follows:

$$\mathbf{h}_t^d = f_\phi(\mathbf{H}^X, y_{\leq t}) \tag{6}$$

And we conduct the same 5-fold validation using `gpt2-medium`, resulting in a ScaSim score of 2106.907 (±167.4572) on CELER-L1 and 1479.553 (±33.9372) on BSC.

Dataset	Eyetracker	Sent.	Words / Sent.	Readers	Language
BSC	EyeLink II (500 Hz)	150	11.2 ± 1.6	60 (L1)	Chinese
CELER L1	EyeLink 1000 (1000 Hz)	5,460	11.2 ± 3.6	69 (L1)	English
CELER L2	EyeLink 1000 (1000 Hz)	23,166	11.0 ± 3.6	296 (L2)	English

Table 2: Summary statistics of the eye-tracking datasets used in this study.