Fine-Grained Spatio-Temporal Modeling of Reading Behavior

Anonymous ACL submission

Abstract

Reading is a process that unfolds across space and time. Standard modeling approaches, however, overlook much of the spatio-temporal dynamics involved in reading by relying on aggregated reading measurements—typically only focusing on fixation durations—and employing modeling techniques that impose strong assumptions. In this paper, we propose a model that captures not only *how long* fixations last, but also *where* they land in space and *when* they take place in time. This is achieved by considering reading as an alternating renewal process, in which the locations and durations of eye fixations are modeled separately yet cohesively. The location (and timing) of fixation shifts, so-called saccades, are modeled using a spatio-temporal Hawkes process, which captures how each fixation *excites* the probability of a new fixation occurring near it in time and space. Empirically, our Hawkes process model exhibits higher likelihood on held-out reading data than baselines. The duration time of fixation events is modeled as a function of fixation-specific features convolved across time, thus capturing non-stationary delayed effects. We evaluate goodness-of-fit across various time-to-event distributions and find evidence that previous convolution-based approaches (Shain and Schuler, 2018, 2021) are insufficiently expressive for modeling disaggregated durations. Finally, testing surprisal theory on disaggregated data, we find that it is weakly predictive of where fixations land but has virtually no predictive power for individual fixation durations.

1 Introduction

Reading is a cognitively complex skill that unfolds across both space and time. As we read, our eyes move through an interdigitated sequence of **fixations**, brief pauses that allow for the perception and processing of linguistic material, and **saccades**, rapid movements that shift focus to the next point of interest. A longstanding premise in psycholinguistic research is that eye movements during reading provide a direct window into the cognitive processes underlying human language comprehension (McConkie, 1979; Just and Carpenter, 1980; Rayner et al., 1989; Findlay and Walker, 1999). Indeed, eye-tracking experiments have emerged as one of the most effective paradigms for testing and refining theories of language processing (Rayner, 1998; Frank et al., 2013). Data collected in eye-tracking trials consist of sequences of fixations across texts displayed within a two-dimensional coordinate space (e.g., a screen), along with their durations and onset time.

In modern computational psycholinguistic studies, the raw data is typically aggregated into summary measurements, e.g., **total fixation duration**, the summed duration of all fixations on a chosen linguistic unit, and **gaze duration**, the summed duration of all fixations between landing on a word and moving to another (see §2.1 for further details on aggregations). These summary measurements are then treated as dependent variables in a (generalized) linear model (Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020). Such aggregation, however, is an inherently lossy process. From a temporal perspective, combining multiple fixations in a single measurement may conflate several factors that underlie the aggregated behaviors. For example, total fixation time includes first fixations as well as regressions, which are fixations where the reader moves backwards spatially and correspond to a different cognitive process (Wilcox et al., 2024). From a spatial perspective, aggregations inherently rely on pre-defined regions of interest (Giulianelli et al., 2024). Aggregating fixations, which is most commonly word level, discards any information about *where* saccades land within the boundaries of a word and hinders investigations into smaller linguistic units, such as syllables or morphemes. In sum, while aggregations help simplify the challenge of modeling and interpreting

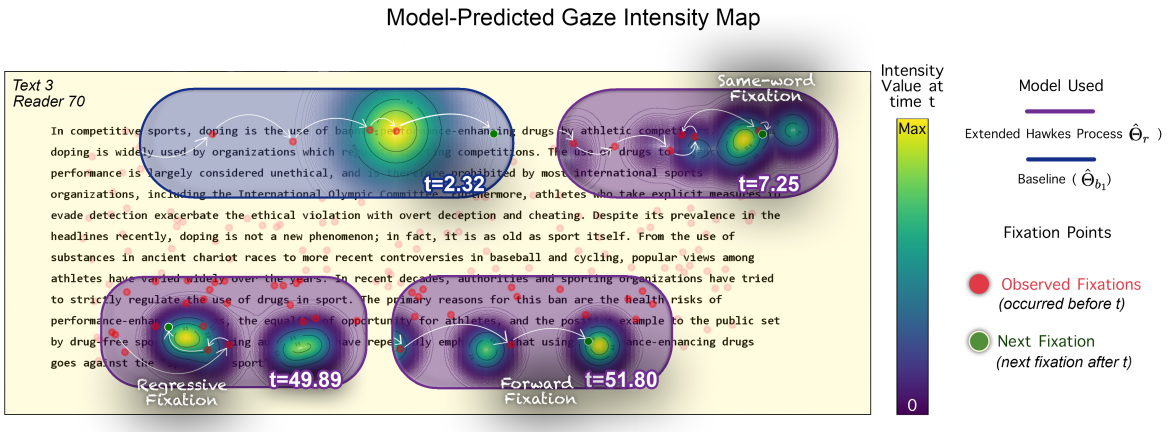

Figure 1: Illustration of how the density of fixations evolve over time within a reading session, and how it compares to held-out data. The intensity function is visualized at different timestamps and shows the predicted fixation intensity of our extended Hawkes process and a baseline model. Red dots indicate observed fixations before time t, while green dots represent the next fixation after t. Note how the following effects are captured by the Hawkes process: forward fixations ($t = 51.80$), backward regressions ($t = 49.89$), and re-fixations on the same word ($t = 7.25$).

the complex spatio-temporal dynamics of fixations and saccades, they inevitably result in a loss of information when compared to the raw reading data.

Beyond the fact that how one aggregates the data can result in loss of information, it also has a significant impact on the empirical support for a given theory. For example, surprisal theory (Hale, 2001; Levy, 2008) suggests that contextual word predictability should have a more pronounced effect on gaze duration than total fixation time, as the latter can be influenced by material from the right context (e.g., through regressive saccades). In contrast, surprisal has been found empirically to be a stronger predictor of total fixation time than gaze duration (Wilcox et al., 2023). Since both gaze and total duration are aggregate measures, providing a precise explanation for such counterintuitive results is challenging. Another theoretical concern with aggregations is that they tend to conflate cognitive and oculomotor control processes. For example, while cognitive processes contribute to reading slowdowns, the oculomotor system imposes an inherent time delay between successive saccades. To incorporate such effects, these models must include additional features describing previous units, so-called spillover variables. However, the number of spillover variables to include is hard to motivate empirically, as the time lag between the onset of an event and its effect on a subsequent fixation will depend on context as well as which type of stimuli that are modeled, i.e., they are non-stationary.

In this paper, we advocate a unified approach that jointly models *when* fixations occur, *where* they land, and *how long* they last. For saccade timing and fixation locations (*when* and *where*), we employ a Hawkes (1971) process, which captures how the density of future fixations changes in response to preceding ones in both time and space. For fixation durations (*how long*), we adopt survival analysis with a log-normal distribution and a convolution-based approach inspired by Shain and Schuler (2021)—originally developed for aggregated durations. We evaluate our framework on the Multilingual Eye-movements Corpus (MECO), assessing its ability to jointly model spatio-temporal disaggregated fixation and duration patterns. Our findings highlight the importance of explicitly modeling the coordination of oculomotor control (e.g., mechanical left-to-right shifts) and cognitive processes (e.g., through surprisal predictors), as well as inter-subject variability in reading strategies.

2 Reading Behavior

While reading, our eyes make progress through the text via brief, rapid movements called saccades. Very little visual information is extracted during a saccade (Ishida and Ikeda, 1989). Instead, most information is extracted during the pauses that occur between saccades, where the eyes remain stationary. These pauses, which are longer than saccades, are called **fixations**. During a reading session, our eyes alternate between fixations and saccades.

Examining reading behavior is key to understanding the cognitive mechanisms that underlie

reading. For example, fixations are known to reflect lexical access (Lima and Inhoff, 1985), syntactic parsing (Frazier and Rayner, 1982), and semantic integration (Ehrlich and Rayner, 1983). A common way to measure reading behavior is through eye-tracking studies (Rayner, 1998), which record high frequency gaze samples that are segmented into discrete fixations. Each fixation ξ is characterized as a tuple $(t, \mathbf{s}, d)$ consisting of an **onset time** t, a **spatial location** $\mathbf{s}$, and a **duration** d. The onset time $t \in \mathbb{R}_+$ is the starting time of the fixation relative to some reference point, typically the start of the reading session. The spatial position lives in a two-dimensional coordinate space $\Omega \subset \mathbb{R}^2$, e.g., a screen. Finally, the duration $d \in \mathbb{R}_+$ captures how long the eye remains still before initiating the next saccade. We consider a reading session $\mathcal{T}$ to be a set of N fixations, i.e.,

$$\mathcal{T} = \{\, \xi_1, \dots, \xi_n \,\}, \tag{1}$$

where $t_i < t_j$ if $i < j$. For each t_i, we define the history $\mathcal{H}_{t_i}$ as

$$\mathcal{H}_{t_i} \;=\; \{\, (t_j, \mathbf{s}_j, d_j) \;\mid\; t_j \leq t_i \}. \tag{2}$$

Note that in addition to fixations' onset times, locations, and durations, this sequence encodes the onset times and durations of saccades as well.[1]

2.1 Modeling Aggregated Reading Data

Rather than as a raw sequence of fixation events, reading data are typically preprocessed into reading time variables at the word level (see, e.g., Frank et al., 2013), though aggregations around other regions of interest are also used, especially when studying specific types of reading behavior like skip rates (Rayner et al., 2011). The most common word-level aggregations are first-fixation time, the duration of the first fixation that lands on a word; first-pass time, the summed duration of all fixations between landing on a word's region and leaving it; and total fixation duration, the summed duration of all the fixations on the word. For more details, see App. A. Disentangling the oculomotor and inferential processes that contribute to each aggregate measurement is a challenging task, particularly given that they are hierarchically ordered, with each variable including the previous.

[1] In particular, the onset of the i-th saccade can be inferred by adding the i-th duration to the i-th fixation onset. The duration of the saccades can be inferred by taking the difference between the fixation onsets and the saccade onsets.

The standard approach to analyzing these variables is to use linear modeling or generalized additive models (GAMs; Kliegl, 2007; Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020, 2023; Gruteke Klein et al., 2024, *inter alia*). When multiple reading times per stimulus are available, i.e., when multiple participants read the same stimulus in their respective trials, this is typically modeled with linear mixed-effects models (Aurnhammer and Frank, 2019; Xu et al., 2023), using individual reading time as a random effect to account for variability across participants.

3 Modeling Spatio-temporal Reading Data

We propose a model of the raw reading session $\mathcal{T}$ rather than aggretations as discussed in §2.1. Thus, we explicitly model *when* a new fixation begins, *where* it lands, and *how long* it lasts. In particular, we model the reading session as an **alternating renewal process** (Cox, 1967) that cycles between two phases, a saccade phase and a fixation phase:

1. **Saccade phase (§3.2).** This is the phase where the reader's eyes move from one fixation location to the next. Saccades are modeled using a spatio-temporal Hawkes process, which predicts the density of the next fixation onset and its location (i.e., the time of the current saccade's termination and its landing location, respectively) conditioned on the fixation history. The Hawkes process encodes how recent fixations spark a short-term increase in the probability of new fixation events nearby in time and space, so-called *self excitation*.

2. **Fixation phase (§5.4).** This is the phase where the reader's eyes dwell on a fixed location. Fixation durations are modeled as influenced by a history of fixation events using some parametric model, which can be selected by the modeler.

These two phases are combined into a unified generative procedure, described in §3.4, which produces the entire sequence of fixations and saccades for a reading session.

Why separate them? We argue that saccades and fixation durations should be modeled separately due to their differing characteristics. (1) Fixations are typically longer and concern a single location in the text, whereas saccades are short movements connecting two locations. (2) Saccade times

naturally exhibit a self-exciting property: more recent events have a greater influence on the likelihood of future events in time and space. This does not hold for fixation durations, as a longer fixation does not necessarily cause subsequent fixations to be longer or shorter. (3) We further observe that while the decision of where to land next is largely determined during a fixation (Aslin and Shea, 1987; Abrams and Jonides, 1988), the probability of actually moving to a different location remains zero until the fixation ends. Because a saccade can only occur after a fixation terminates, modeling fixation durations, saccade times, and saccade landing spots jointly can be achieved without loss of generality via a two-step approach: first modeling fixation durations, and then modeling saccade times and landing spots conditioned on the outcome of the former.

3.1 Transforming Time to Remove Fixation Durations

We decouple saccade duration from fixation duration in the following way. Consider the onset times $t_1, t_2, \dots, t_n$. We now remove fixation durations from the timeline via a mapping of onset times $h_{\mathcal{T}}\colon \mathbb{R}_+ \to \mathbb{R}_+$, defined as

$$h_{\mathcal{T}}(t_i) \;=\; t_i \;-\; \sum_{(t,\mathbf{s},d)\,\in\,\mathcal{H}_{t_i}} d\,. \tag{3}$$

In words, $h_{\mathcal{T}}(t_i)$ gives the onset time of the i-th saccade when the fixation durations are excluded from the timeline. In this *transformed* domain, each fixation onset t_i is mapped to $h_{\mathcal{T}}(t_i)$. We will refer to $h_{\mathcal{T}}(t)$ with t^* for brevity. Collecting these gives a new sequence

$$\mathcal{T}^* \;=\; \Big\{\big(t_1^*,\, \mathbf{s}_1,\, d_1\big), \dots, \big(t_n^*,\, \mathbf{s}_n,\, d_n\big)\Big\},$$

where the $\mathbf{s}$ and d values remain as in Eq. (1). Because all fixation durations are subtracted out, consecutive fixations in $\mathcal{T}^*$ appear as if they had zero duration; thus, the time elapsed *between* events in $\mathcal{T}^*$ correspond precisely to the saccade durations.

3.2 Hawkes Process for Saccade Onsets

We now define a **spatio-temporal Hawkes process** on the transformed times t^* and locations $\mathbf{s}$ in two-dimensional space. Let $\mathcal{H}_{t^*}$ be the history of events in the transformed domain up to t^*. The Hawkes process assumes that the probability of a new event occurring in the time interval $[t^*, t^* + dt)$ and in space $[\mathbf{s}, \mathbf{s} + d\mathbf{s})$ is $\lambda(t^*, \mathbf{s} \mid \mathcal{H}_{t^*})dtd\mathbf{s}$, where $\lambda\colon \mathbb{R}_+ \times \Omega \to \mathbb{R}_+$ is called the intensity function. Conditioning on the past events, it captures how past fixations $\big(t^{*\prime}, \mathbf{s}'\big)$ influence the probability of new fixations (i.e., *self-excite*) in an additive manner:

$$\lambda(t^*, \mathbf{s}\,|\mathcal{H}_{t^*}) \;= \tag{4}$$
$$\nu + \sum_{(t^{*\prime},\,\mathbf{s}')\,\in\,\mathcal{H}_{t^*}} \phi\big(t^* - t^{*\prime}\big)\; f(\mathbf{s} \mid \mathbf{s}'),$$

where $\nu \in \mathbb{R}_+$ is the base intensity, $\phi\colon \mathbb{R}_+ \to \mathbb{R}_+$ is an exponentially decaying temporal kernel governing the influence of past events, and $f(\mathbf{s} \mid \mathbf{s}')$ is a distribution over Ω. Note that omitting the sum in Eq. (4) yields a Poisson process (Palm, 1943).

Temporal Kernel. We let each past fixation contribute an exponentially decaying influence on the intensity of future fixations. The exponential decay kernel is defined as

$$\phi(\Delta) \;=\; g\big(\mathbf{x}_{t^*}^T\alpha\big)\; \exp\Big(-g(\mathbf{x}_{t^*}^T\beta)\,\Delta\Big), \tag{5}$$

where $\Delta = t^* - t^{*\prime}$ is the time elapsed between transformed onset times, $\mathbf{x}_{t^*} \in \mathbb{R}^p$ is a vector of predictor values, $\alpha, \beta \in \mathbb{R}^p$ are learnable parameters, and $g\colon \mathbb{R} \to \mathbb{R}_+$ is a linking function that ensures the non-negativity of the parameters, e.g., a ReLU. Note that the linking function is used twice in Eq. (5); $g\big(\mathbf{x}_{t^*}^T\alpha\big)$ quantifies how much a fixation increases the probability of subsequent fixations (excitation strength), while $g(\mathbf{x}_{t^*}^T\beta)$ determines how quickly the influence of the fixation diminishes over time (decay rate). Since they depend on an event-specific vector of predictor values $\mathbf{x}_{t^*}$, the strength of excitation and decay rate may vary across different spatio-temporal conditions.

Spatial Distribution. We model the spatial component $f(\mathbf{s} \mid \mathbf{s}')$ with a normal distribution with a mean $\mu(\mathbf{s}')$ which depends on $\mathbf{s}'$. We assume a scalar variance σ^2 (i.e., a spherical Gaussian), yielding the density function:

$$f(\mathbf{s} \mid \mathbf{s}') = \frac{1}{2\pi\,\sigma^2}\exp\left(-\tfrac{\|\mathbf{s}-\mu(\mathbf{s}')\|^2}{2\,\sigma^2}\right). \tag{6}$$

Overall Hawkes Density. These design choices result in the intensity function λ being a mixture of Gaussian components, each corresponding to a past fixation event. Thus, we obtain a multimodal spatial distribution. This aligns well with reading behavior, since it captures that eye fixations may jump either forward or backward in the text with varying probabilities. The overall density is defined as

$$f_{hp}(t^*, \mathbf{s} \mid \mathcal{H}_{t^*}) \;=\; \frac{\lambda(t^*, \mathbf{s} \mid \mathcal{H}_{t^*})}{Z}, \tag{7}$$

where the normalizing constant Z is

$$Z = \exp\left(\int_{t_{\mathcal{H}}^-}^{t^*} \int_{\Omega} \lambda(u, \boldsymbol{z} \mid \mathcal{H}_u)\, d\boldsymbol{z}\, du\right), \quad (8)$$

with $t_{\mathcal{H}}^- \stackrel{\text{def}}{=} \max_{t^*}\{t^* \mid (t^*, \mathbf{s}, d) \in \mathcal{H}_{t^*}\}$, i.e., the onset of the saccade in the history that occurred latest in time.

3.3 Parametric Distribution for Fixation Durations

We model fixation duration as a nonnegative random variable with density $f_d(t \mid \mathcal{H}_t)$, where $\mathcal{H}_t$ denotes the history of the reading session up to time t (here referring to the original, non-transformed session). To capture time-to-event data such as fixation durations, we employ a survival analysis framework. Within this framework, the density is factorized into a hazard function and a survival function:

$$f_d(t \mid \mathcal{H}_t) = h(t \mid \mathcal{H}_t)\, S(t \mid \mathcal{H}_t), \quad (9)$$

where

$$S(t \mid \mathcal{H}_t) = \exp\left(-\int_0^t h(s \mid \mathcal{H}_t)\, ds\right). \quad (10)$$

The hazard function $h(t \mid \mathcal{H}_t)$ represents the instantaneous rate at which a fixation terminates at time t, given that it has persisted until then. For example, the hazard is typically low at the onset of a fixation but increases as the reader proceeds through the stages of processing, eventually peaking at an optimal moment when the reader is most prepared to shift attention. The survival framework accommodates a variety of hazard functions. Furthermore, we parameterize f_d via a linear model and a linking function which incorporates predictors from previous fixations through a temporal convolution, as in Shain and Schuler (2021).

3.4 Unified Generative Model

Putting it all together, the reading process unfolds as follows:

(a) **Draw a new fixation onset** $(t_i, \mathbf{s}_i)$ from the Hawkes process in the *transformed* timeline:

$$({t^*}_i, \mathbf{s}_i) \sim f_{hp}(\cdot \mid \mathcal{H}_{{t^*}_i}), \quad t_i = h_{\mathcal{T}}^{-1}({t^*}_i),$$

(b) **Sample fixation duration** $d_i \sim f_d(\cdot \mid \mathcal{H}_{t_i})$ using Eq. (9).

(c) **Update the history** by adding the new sample $\xi_i = (t_i, \mathbf{s}_i, d_i)$ to

$$\mathcal{H}_{t_{i+1}} = \{\mathcal{H}_{t_i} \cup \xi_i\}.$$

This process is then repeated for any new fixation up to an ending time $T \in \mathbb{R}_+$. We implement this framework in PyTorch to learn parameters shared across different reading sessions through automated differentiation.

4 Experimental Setup

4.1 Data

We use the MECO dataset (Siegelman et al., 2022) as a source of reading data. It contains 12 short text excerpts from Wikipedia. These texts were displayed on a 1920×1080 screen in monospaced Consolas 22pt font. As summarized in Tab. 1 (App. B), the number of characters per text ranges from 831 to 1230 (average 1093), and the number of lines from 8 to 12 (average 10.5). We use the Python Tesseract OCR library[2] to identify characters and their bounding boxes. More details in App. B.

Fixation Data. The MECO dataset provides gaze measurements for multiple readers and reading sessions. For each session, we consolidated the measurements into a sequence $\mathcal{T}$ for each reader. This yielded a dataset where each fixation event ξ could be aligned with a bounding box in space and a specific time interval. We split the dataset into an 80% training, 10% validation, and 10% test. The training set contains $78,033$ fixation samples.

Surprisal. We are interested in whether word- and character-level surprisal influences the spatio-temporal dynamics of reading. Let Σ be a finite, non-empty set of lexical units (i.e., characters or words), called an **alphabet**, and Σ^* be the set of all strings that can be formed by concatenating units in Σ. We further assign a special symbol $\textsc{eos} \notin \Sigma$ to denote the end of an utterance, and define $\overline{\Sigma} \stackrel{\text{def}}{=} \Sigma \cup \{\textsc{eos}\}$. Following Shannon's (1948) formulation of information content, the surprisal of a unit $w_t \in \overline{\Sigma}$ in a preceding context of units $\boldsymbol{w}_{<t} \in \Sigma^*$ is defined as

$$s_t(w_t) \stackrel{\text{def}}{=} -\log_2 p(w_t \mid \boldsymbol{w}_{<t}),$$

where $p(\cdot \mid \boldsymbol{w}_{<t})$ is the true (albeit unknown) distribution over $\overline{\Sigma}$ conditioned on the preceding context

[2] https://pypi.org/project/pytesseract/

$\boldsymbol{w}_{<t}$. In practice, we estimate $p(\cdot \mid \boldsymbol{w}_{<t})$ using an autoregressive language model. Since most modern language models learn a distribution over *token* sequences, which may or may not correspond to standard linguistic units (Bostrom and Durrett, 2020; Gow-Smith et al., 2022; Beinborn and Pinter, 2023), we need a way of transforming the probability of a token sequence to one over words and characters. To obtain character-level surprisals, we use the algorithm proposed by Vieira et al. (2024); the surprisal values are taken from Giulianelli et al. (2024) and were computed using GPT-2 (Radford et al., 2019). Word-level surprisals are obtained simply by summing the surprisal values of the subword tokens that comprises the word; these values are taken from Opedal et al. (2024) and were computed using mGPT (Shliazhko et al., 2024).[3] The surprisal value associated with a fixation is taken as the surprisal of the corresponding character or word of the bounding box in which the fixation landed. We do not assign a surprisal value to fixations that did not land within a bounding box.

5 Modeling Saccade Onset and Fixation Location

We employ the framework introduced in §3.2 to model saccade onsets and fixation locations. First, we define three baseline models. We then propose a series of extensions intended to investigate the spatio-temporal dynamics of reading.

5.1 Baseline Models

The baseline models capture spatial and temporal dependencies by modifying the intensity function in Eq. (4). The first is a **Poisson Baseline** that assumes every fixation in space and time is equally likely and independent of past events. The second is a **Last-Fixation Baseline** that posits each new fixation is normally distributed around the previous fixation with constant variance. Finally, we introduce a **Standard Hawkes Baseline**, extending the Last-Fixation Baseline by incorporating past fixations with a temporally decaying influence on the next fixation. Further details on these models are provided in App. C.

[3]Alternative methods have been proposed to compute the probability of words (Oh and Schuler, 2024; Pimentel and Meister, 2024) or any character string (Vieira et al., 2024) from language models over tokens. While their impact on psycholinguistic predictive power is evident when considering sub- or super-word strings (Giulianelli et al., 2024), it remains less clear whether the additional runtime of these alternative methods provides the same benefits when focusing on words (Oh and Schuler, 2024; Pimentel and Meister, 2024).

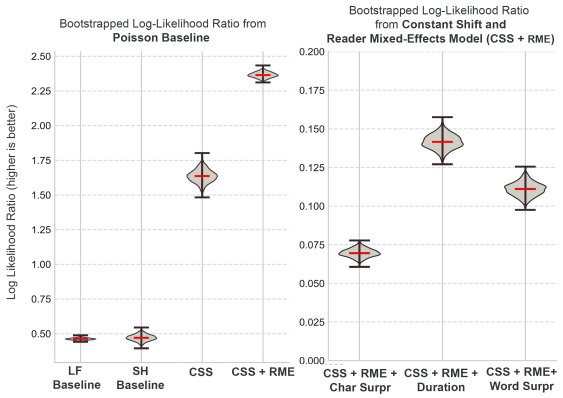

Figure 2: Bootstrap distributions for the log-likelihood ratio in the Hawkes Process. The left plot compares the Last Fixation Baseline (LF Baseline), the Standard Hawkes Process Baseline (SH Baseline), the Constant Spatila Shift (CSS), and the Constant Shift (CSS) + Reader Mixed Effects (RME) model with the Poisson baseline. The right plot compares the three models with added predictors (character-level surprisal, duration, and word-level surprisal) against the CSS + RME model.

5.2 Modeling Extensions

We consider three modeling extensions that incorporate a notion of mechanical spatial shift, reader-specific effects, and fixation attributes such as the fixation region's surprisal.

Constant Spatial Shift (CSS) Model. We consider an extension that evaluates whether a learned but constant *spatial shift* from previous fixations—independent of textual content—can account for eye movements. The underlying assumption is that eye movements follow a mechanical progression through text. From a modeling perspective, we modify the standard Hawkes process by allowing $\mu(\mathbf{s})$ to learn a translation, represented by a matrix $\mathbf{A} \in \mathbb{R}^{2\times 2}$ and an offset $\mathbf{b} \in \mathbb{R}^2$, from location $\mathbf{s}$, i.e.,

$$\mu(\mathbf{s}) = \mathbf{A}\mathbf{s} + \mathbf{b}.$$

The model parameters are $\Theta_c = \{\nu, \alpha, \beta, \sigma^2, \mathbf{A}, \mathbf{b}\}$.

Reader Mixed-Effects (RME) + CSS Model. The models introduced so far treat all reading sessions as though they come from a single, average reader. In practice, however, individuals often exhibit distinct reading styles. To capture these differences, we extend the model to include reader-specific deviations in both the temporal and spatial parameters. Specifically, for each reader r, we

introduce subject-level offsets γ_α^r and γ_β^r for the temporal kernel parameters :

$$\alpha^r = \alpha + \gamma_\alpha^r \quad \text{and} \quad \beta^r = \beta + \gamma_\beta^r,$$

where α and β are the global (population-level) temporal parameters, and $\gamma_\alpha^r, \gamma_\beta^r \in \mathbb{R}$ capture how reader r's temporal dynamics differ from the average. To account for spatial biases, we let the mean of the spatial distribution shift by a reader-specific vector $\mathbf{d}^r \in \mathbb{R}^2$:

$$\mu^r(\mathbf{s}) = \mathbf{A}\mathbf{s} + \mathbf{b} + \mathbf{d}^r.$$

Here, $\mathbf{d}^r$ represents an x- and y-offset indicating reader r's typical gaze shift relative to the global mean. Under this mixed-effects formulation, the overall parameter set is:

$$\Theta_r = \Theta_c \cup \{\gamma_\alpha^r, \gamma_\beta^r, \mathbf{d}^r \mid r = 1, \dots, R\},$$

where Θ_c collects the global parameters.

Incorporating Fixation Attributes. Finally, we treat each fixation's attributes—duration (d), character-level surprisal (s_c), and word-level surprisal (s_w)—as "marks" m_i that modulate the temporal kernel and spatial mean. For example, for reader r and fixation i:

$$\alpha_i^r = \alpha + \gamma_\alpha^r + \alpha_m\, m_i, \qquad \beta_i^r = \beta + \gamma_\beta^r + \beta_m\, m_i,$$

$$\mu_i^r(\mathbf{s}) = \mathbf{A}\,\mathbf{s} + \mathbf{b} + \mathbf{d}^r + \mathbf{g}_m\, m_i.$$

where $m_i \in \{d, s_c, s_w\}$. The new parameters $\{\alpha_m, \beta_m, \mathbf{g}_m\}$ capture how each attribute shifts temporal or spatial dynamics. We gather them in:

$$\Theta_{\text{marks}} = \{\alpha_m, \beta_m, \mathbf{g}_m\}_{m \in \{d, s_c, s_w\}},$$

so the full parameter set is $\Theta^m = \Theta_c \cup \Theta_{\text{marks}}$.

5.3 Results on Hawkes Process Model

We train the models until convergence under different hyperparameter values; see App. G for details. The final model candidate is selected based on likelihood on the validation set and subsequently tested on a held-out test set. The model performance on the test set is compared using log-likelihood ratios.

Fit on held-out data. First, we investigate how well our family of models fits with held-out reading data. Fig. 2 (left) shows the log-likelihood ratio under bootstrapped uncertainty intervals relative to the Poisson-process baseline for four different models: (i) the Fixation Baseline (Last-Fixation Model), (ii) the Standard Hawkes Baseline, (iii) the Constant Spatial Shift extension (CSS), and (iv) the Constant Spatial Shift + Reader Mixed-Effect extension (CSS + RME). First, we observe the last-fixation and Hawkes process baselines yield significant improvements over the Poisson process. Moreover, while the self-excitation behavior on its own is not sufficient to yield improvements over a model only accounting for the last fixation, the model incorporating a constant spatial shift does show a marked improvement in log-likelihood ratio. This supports the hypothesis that saccade generation involves memory of recent fixation locations and that reading unfolds in space and time with a strong mechanical component. Specifically, the estimated parameters indicate a global rightward shift of approximately 1.25 units per fixation. Given that the x-axis ranges from 0 to 20 on the screen, each new fixation is thus predicted to move about 5% further to the right. Further introducing reader-specific mixed effects reveals substantial individual variability (e.g., in skipping or regressing). Analyzing the temporal coefficients per reader yields a mean $\hat{\alpha}$ of 5.63 ($SD = 0.44$) and a mean $\hat{\beta}$ of 9.12 ($SD = 1.21$). The lower $\hat{\alpha}$ compared to $\hat{\beta}$ suggests that although recent fixations do influence the likelihood of another fixation in short succession, this influence decays relatively quickly over time. Fig. 1, shows the CSS + RME intensity at different times during a reading session. Qualitatively, the figure demonstrates how this model successfully captures the intensity around the next fixation, whether it is a forward fixation, a regressive fixation, or one that remains on the same word.

Influence of surprisal and past durations. Next, we study whether incorporating surprisal and past durations into the model helps predicting the location of the fixation. In Fig. 2 (right), we illustrate the gains from further adding these fixation-level attributes as compared to the RME + CSS model. We see a modest but consistent improvement beyond the mechanical and reader-dependent factors, suggesting that local cognitive load also helps shape eye-movement decisions. In all cases we find that $\hat{\alpha}_m > \hat{\beta}_m$, which suggests that past fixations with larger durations, character-level and word-level surprisals increase the temporal influence on the next fixation. We extend on this statement in the appendix in relation to the branching ratio of a Hawkes process (App. F).

5.4 Modeling Fixation Durations

We model fixation durations using a log-normal distribution, chosen for parameter interpretability through cross-validated comparison to alternatives like the gamma distribution (App. D). Let $t > 0$ denote duration with:

$$\log t \sim \mathcal{N}\left(\mu^d(\mathcal{H}_t), \sigma^2\right)$$

where $\mu^d(\mathcal{H}_t)$ encodes history-dependent effects (e.g., prior fixation dynamics) and σ^2 is a shared log-variance parameter. Our baseline estimates $(\mu_0^d, \sigma^2) \in \mathbb{R}^2$, with extensions proposed below. We show a full survival analysis derivation in App. D.

Mark-Dependent Past Influence. We let each fixation's duration depend on the history of prior "marks" (e.g., durations or surprisals). Let m_i denote a mark at time t_i. We define the discrete convolution:

$$(f_\Gamma * \eta^m)(t \mid \alpha_\gamma, \beta_\gamma, \delta_\gamma) = \sum_{(t', m') \in \mathcal{H}_t} m' f_\Gamma\big(t - t' \mid \alpha_\gamma, \beta_\gamma, \delta_\gamma\big). \tag{11}$$

where η^m is a discrete measure for mark m, and f_Γ is the density of a shifted-gamma distribution, as previously employed by Shain and Schuler (2018). We refer to App. E for details. For a fixation i, the mean log-duration μ_i^d is modeled via

$$\mu_i^d = \beta_0 + \beta_1^m \left(f_\Gamma * \eta^m\right)(t_i) + \gamma_r^d.$$

In the case of character or word-level surprisal, we also include an additional predictor for the current word's (or character's) surprisal value.

5.5 Results

We followed the experimental protocol described in §5.3 to evaluate the test-set likelihoods of our proposed models. Our primary objective is to assess the impact of three predictors—duration, character-level surprisal, and word-level surprisal—on individual duration modeling, while controlling for mixed reader effects. The results (Fig. 3) demonstrate that established methods for modeling aggregated durations only detect very weakly effects when applied to individual-level predictions. While bootstrapping confirms statistically significant likelihood ratios for each mark-dependent model, the small magnitude of these ratios (0.03–0.05) indicates limited predictive power in disaggregated settings. We note that these experiments were performed on unfiltered data, which could be noisy.

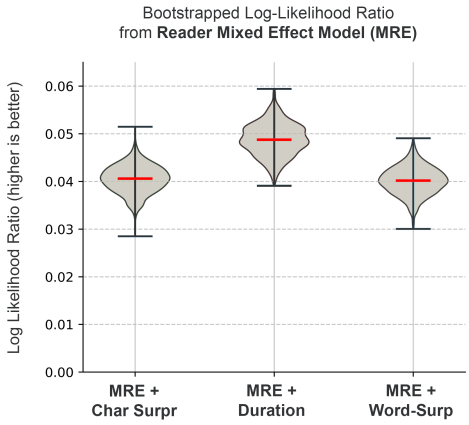

Figure 3: Bootstrapped distributions of the Log-Likelihood Ratio for the duration model, comparing the three models with added predictors (character-level surprisal, duration, and word-level surprisal) against the Reader-Mixed Effect (MRE) Duration model.

6 Conclusion

We introduced a unified probabilistic framework for modeling reading behavior, integrating fixation duration, location, and saccade timing within a single mathematical structure. Unlike models that focus solely on temporal patterns (Shain and Schuler, 2018, 2021), our alternating renewal process captures both spatial and temporal dynamics, enabling a detailed examination of both oculomotor control and cognitive processes underlying reading. Our results highlight the significant role of mechanical processes, such as systematic left-to-right shifts, in shaping reading behavior. They also emphasize the importance of explicitly modeling inter-subject variability, particularly when analyzing disaggregated data. Moreover, while predictors like surprisal effectively model aggregate reading patterns, we find they struggle to generalize to individual fixation-level dynamics. Specifically, word and character surprisal exhibit only moderate predictive power for saccade timing and landing position and very weak predictive power for individual fixation durations, suggesting a potential oversimplification of the link between cognitive mechanisms and reading behavior. Overall, our work challenges the direct applicability of aggregated cognitive models to individual-level analysis and provides a flexible foundation for more precise and scalable investigations of reading behavior.

Limitations

While our framework advances eye-movement modeling, four limitations merit discussion. First, our analysis uses English data from MECO, leaving the 10 other languages unexamined. Cross-linguistic validation is needed to assess whether our findings generalize to languages with distinct orthographic or syntactic properties (e.g., languages that read from right to left). Second, we intentionally constrained linguistic predictors (e.g., surprisal features) to simple link-linear relationships (i.e., a linear transformation combined with a link function) to maintain interpretability; however, this may oversimplify the cognitive-ocular relationships compared to high-capacity neural architectures. Our open-source PyTorch implementation enables future exploration of nonlinear or hierarchical feature interactions. Third, for fixation durations, we compared previous work on aggregated data to the disaggregated setting but did not propose a method tailored specifically for the disaggregated setting. Finally, the Standard Hawkes process baseline and Last Fixation baseline showed comparable performance. While we augmented the Standard Hawkes model to capture spatial gaze patterns, we did not similarly extend the Last Fixation baseline. A more symmetrical evaluation—testing both models with spatial covariates—could clarify their relative advantages.

References

Richard A Abrams and John Jonides. 1988. Programming saccadic eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 14(3):428.

Richard N. Aslin and Sandra L. Shea. 1987. The amplitude and angle of saccades to double-step target displacements. *Vision Research*, 27(11):1925–1942.

Christoph Aurnhammer and Stefan L. Frank. 2019. Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, 134:107198.

Lisa Beinborn and Yuval Pinter. 2023. Analyzing cognitive plausibility of subword tokenization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Yevgeni Berzak and Roger Levy. 2023. Eye movement traces of linguistic knowledge in native and non-native reading. *Open Mind*, 7:179–196.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Charles Clifton, Adrian Staub, and Keith Rayner. 2007. Chapter 15 - 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*, pages 341–371. Elsevier, Oxford.

D.R. Cox. 1967. *Renewal Theory*. Springer Netherlands.

A.G. de Varda, M. Marelli, and S. Amenta. 2024. Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. *Behavior Research Methods*, 56:5190–5213.

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

Kate Ehrlich and Keith Rayner. 1983. Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior*, 22(1):75–87.

John M. Findlay and Robin Walker. 1999. A model of saccade generation based on parallel processing and competitive inhibition. *Behavioral and Brain Sciences*, 22(4):661–674.

Stephan L. Frank, Irene Fernandez Monsalve, René L. Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45:1182–1190.

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.

Mario Giulianelli, Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Tim Vieira, and Ryan Cotterell. 2024. On the proper treatment of tokenization in psycholinguistics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18556–18572, Miami, Florida, USA. Association for Computational Linguistics.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Edward Gow-Smith, Harish Tayyar Madabushi, Carolina Scarton, and Aline Villavicencio. 2022. Improving tokenisation by alternative treatment of spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Keren Gruteke Klein, Yoav Meiri, Omer Shubi, and Yevgeni Berzak. 2024. The effect of surprisal on reading times in information seeking and repeated reading. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 219–230, Miami, FL, USA. Association for Computational Linguistics.

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*.

Alan G. Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Albrecht Werner Inhoff. 1984. Two stages of word processing during eye fixations in the reading of prose. *Journal of Verbal Learning and Verbal Behavior*, 23(5):612–624.

Taiichiro Ishida and Mitsuo Ikeda. 1989. Temporal properties of information extraction in reading studied by a text-mask replacement technique. *J. Opt. Soc. Am. A*, 6(10):1624–1632.

MA Just and PA Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review*, 87(4):329—354.

Reinhold Kliegl. 2007. Toward a perceptual-span theory of distributed processing in reading: A reply to Rayner, Pollatsek, Drieghe, Slattery, and Reichle (2007). *Journal of Experimental Psychology: General*, 136(3):530–537.

Patrick J Laub, Thomas Taimre, and Philip K Pollett. 2015. Hawkes processes. *arXiv preprint arXiv:1507.02822*.

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

Susan D. Lima and Albrecht W. Inhoff. 1985. Lexical access during eye fixations in reading: Effects of word-initial letter sequence. *Journal of Experimental Psychology: Human Perception and Performance*, 11(3).

George W. McConkie. 1979. *On the Role and Control of Eye Movements in Reading*, pages 37–48. Springer US, Boston, MA.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472, Miami, Florida, USA. Association for Computational Linguistics.

Andreas Opedal, Eleanor Chodroff, Ryan Cotterell, and Ethan Wilcox. 2024. On the role of context in reading time prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3042–3058, Miami, Florida, USA. Association for Computational Linguistics.

Conny Palm. 1943. *Intensitätsschwankungen im Fernsprechverkehr*. Ericsson technics. L. M. Ericcson.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.

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).

Keith Rayner, Sara C. Sereno, Robin K. Morris, A. Réne Schmauder, and Charles Clifton Jr. 1989. Eye movements and on-line language comprehension processes. *Language and Cognitive Processes*, 4(3-4):SI21–SI49.

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).

Leah Roberts and Anna Siyanova-Chanturia. 2013. Using eye-tracking to investigate topics in L2 acquisition and L2 processing. *Studies in Second Language Acquisition*, 35(2):213–235.

Cory Shain and William Schuler. 2018. Deconvolutional time series regression: A technique for modeling temporally diffuse effects. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2679–2689, Brussels, Belgium. Association for Computational Linguistics.

Cory Shain and William Schuler. 2021. Continuous-time deconvolutional regression for psycholinguistic modeling. *Cognition*, 215:104735.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

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.

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).

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

Tim Vieira, Ben LeBrun, Mario Giulianelli, Juan Luis Gastaldi, Brian DuSell, John Terilla, Timothy J. O'Donnell, and Ryan Cotterell. 2024. From language models over tokens to language models over characters. *Preprint*, arXiv:2412.03719.

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

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 1707–1713. Cognitive Science Society.

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

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, Singapore. Association for Computational Linguistics.

A Common Aggregations of Reading Data and their Interpretation

The most common word-level aggregation of reading data are first fixation time, first-pass time, and total fixation time. These are generally thought to reflect progressively later stages of language processing (Inhoff, 1984; Berzak and Levy, 2023). We describe them here, along with their standard interpretations. First-fixation time, the duration of only the first fixation that lands on a word, is associated with word identification and lexical processing (Clifton et al., 2007; Berzak and Levy, 2023) and tends to exhibit smaller surprisal effects (Wilcox et al., 2023; de Varda et al., 2024). First-pass time (or gaze duration), the summed duration of all fixations between landing on a word's region and leaving it, is thought to be indicative of early syntactic and semantic processing, and typically considered the aggregate to be most strongly associated with processing difficulty (e.g., Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020). Total fixation time, the summed duration of all the fixations on the word, including refixations of the region after it was left, is thought to be indicative of integrative processes (e.g., Demberg and Keller, 2008; Roberts and Siyanova-Chanturia, 2013) and sometimes exhibits, somewhat unexpectedly, stronger surprisal effects than first-fixation and first-pass time (Wilcox et al., 2023; Giulianelli et al., 2024).

B Optical Character Recognition (OCR)

We applied the Python Tesseract OCR library[4] to each image in the MECO dataset (Siegelman et al., 2022) to identify textual characters and their bounding boxes. The number of characters per text, the number of lines, and bounding box information are summarized in Tab. 1. Tesseract provides the position and dimensions of each recognized character. We set the heights to a constant value by adjusting them to match the tallest character in the image, and the widths to the 90th percentile of character widths. This ensures a consistent character grid for subsequent analysis. Because Tesseract does not detect whitespace as distinct regions, we identified whitespace ourselves by comparing gaps between adjacent character boxes. Whenever the horizontal gap was at least 80% of a typical single-character width, we considered the gap to be a whitespace and assigned it a bounding box of the same constant height.

	Avg	SD	Min	Max
Lines	10.5	1.2	8.0	12.0
Characters	1093.0	125.2	831.0	1231.0
BBox width	12.0	0.0	12.0	12.0
BBox height	22.3	2.2	18.0	24.0

Table 1: Summary statistics of raw MECO data, including the number of lines and characters per text and bounding box (BBox) dimensions.

C Detailed Description of Baseline Models

In this section, we provide a comprehensive description of the baseline models designed to capture progressively more sophisticated spatial and temporal dependencies in fixation behavior. Each model is defined via a modification of the intensity function introduced in Eq. (4).

C.1 Poisson Process

The Poisson process model assumes that every fixation is equally likely regardless of spatial location or temporal history, implying independent fixations in space and time. Its intensity function is

$$\lambda(\tau, \mathbf{s} \mid \mathcal{H}_\tau) = \nu,$$

where $\nu \in \mathbb{R}_+$ is the only learnable parameter. This serves as the simplest baseline, ignoring any spatial or temporal dependencies.

[4] https://pypi.org/project/pytesseract/

C.2 Last-Fixation Model

The last-fixation model introduces a basic spatial dependency by assuming that each new fixation is normally distributed around the most recent fixation. Let $\mathbf{s}_{t_{\mathcal{H}}^-}$ denote the location of the last fixation. Then, the intensity function is given by

$$\lambda(\tau, \mathbf{s} \mid \mathcal{H}_\tau) = \nu + f\big(\mathbf{s} \mid \mathbf{s}_{t_{\mathcal{H}}^-}, \sigma^2\big),$$

where $f(\cdot \mid \mathbf{s}_{t_{\mathcal{H}}^-}, \sigma^2)$ is the probability density function of a normal distribution centered at $\mathbf{s}_{t_{\mathcal{H}}^-}$ with variance σ^2. This model introduces two learnable parameters: the baseline intensity ν and the variance σ^2, making the set $\Theta_{b1} = \{\nu, \sigma^2\}$.

C.3 Standard Hawkes Process

The standard Hawkes process builds upon the last-fixation model by incorporating the influence of all past fixations with a temporally decaying impact. Specifically, each previous fixation contributes to the current intensity with an exponential decay. The intensity function is defined as

$$\lambda(\tau, \mathbf{s} \mid \mathcal{H}_\tau) = \nu + \sum_{t_i < \tau} \alpha\, e^{-\beta(\tau - t_i)}\, f\big(\mathbf{s} \mid \mu(\mathbf{s}_i), \sigma^2\big),$$

where:

- α and β are the temporal kernel parameters governing the strength and decay rate of the influence from past fixations.
- $f(\cdot \mid \mu(\mathbf{s}_i), \sigma^2)$ is the spatial component, with $\mu(\mathbf{s}_i) = \mathbf{s}_i$ representing the location of the i-th fixation.

The complete parameter set for this model is given by $\Theta_{b_2} = \Theta_{b_1} \cup \{\alpha, \beta\}$, where $\Theta_{b_1} = \{\nu, \sigma^2\}$ are the parameters inherited from the last-fixation model.

D Modeling Duration: A Survival Analysis Framework

Survival analysis provides an optimal framework for modeling fixation durations as it naturally handles right-skewed, time-to-event distributions. We evaluated six candidate distributions through K-fold cross-validation on training and validation data: Rayleigh, exponential, Weibull, normal, log-normal, and gamma.

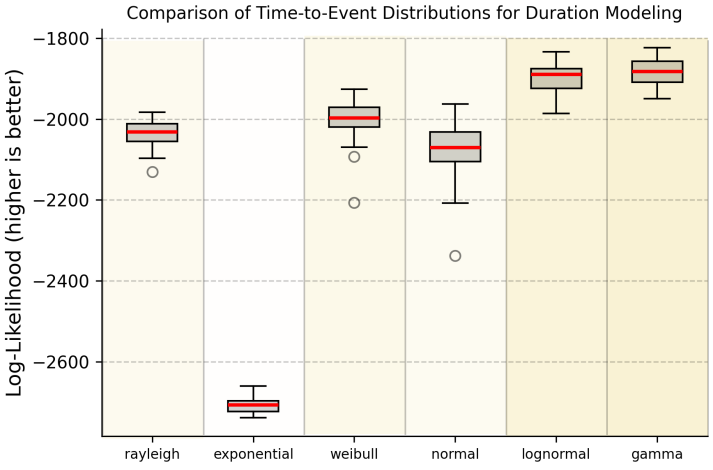

Figure 4: Goodness-of-fit comparison of candidate distributions for fixation durations. Both log-normal and gamma distributions showed superior performance compared to simpler other forms (higher log-likelihood indicate better fit). The log-normal was ultimately selected for its enhanced parameter interpretability.

While the log-normal and gamma distributions demonstrated comparable predictive performance (Fig. 4), we selected the log-normal distribution for two key reasons: (1) its parameters directly correspond to moments of the log-transformed durations, enabling more intuitive interpretation, and (2) it provides closed-form expressions for survival and hazard functions through relationship with the normal distribution.

D.1 Log-Normal Survival Model Specification

Let $t > 0$ represent fixation duration with associated survival characteristics:

- **Log-transformed duration**: $\ln t \sim \mathcal{N}(\mu^d(\mathcal{H}_t), \sigma^2)$
- **Probability density function (PDF)**:

$$f_d(t \mid \mathcal{H}_t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln t - \mu^d(\mathcal{H}_t))^2}{2\sigma^2}\right)$$

- **Survival function**:

$$S(t) = P(T \geq t) = 1 - \Phi\left(\frac{\ln t - \mu^d(\mathcal{H}_t)}{\sigma}\right)$$

  where $\Phi(\cdot)$ denotes the standard normal CDF

- **Hazard function**:

$$h(t) = \frac{f_d(t \mid \mathcal{H}_t)}{S(t)}$$

- **Expected duration**:

$$\mathbb{E}[t \mid \mathcal{H}_t] = \exp\left(\mu^d(\mathcal{H}_t) + \frac{\sigma^2}{2}\right)$$

Here $\mu^d(\mathcal{H}_t)$ represents the conditional mean of log-durations given the fixation history $\mathcal{H}_t$, while σ^2 captures residual variance on the log-scale. This parameterization enables direct interpretation of covariate effects in terms of proportional changes in duration expectation through the exponential relationship.

E Convolution-Based Duration Model Details

We provide full details on the convolution-based predictor described in the main text.

E.1 Measure Definition

We define the discrete measure η^m for a general mark m_i at time t_i as

$$\mathrm{d}\eta^m(\tau) \;=\; \sum_{(t_i, m_i) \in \mathcal{H}} m_i \, \delta(\tau - t_i),$$

where $\delta(\cdot)$ is the Dirac delta function, and $\mathcal{H}$ is the set of all past events.

E.2 Shifted Gamma Kernel

We use a shifted Gamma kernel $f\big(\tau \mid \lambda, \theta, \delta\big)$ (Shain and Schuler, 2018), which takes the form

$$f(\tau \mid \lambda, \theta, \delta) = \frac{(\tau - \delta)^{\theta - 1} \, \lambda^{\theta} \, e^{-\lambda\,(\tau - \delta)}}{\Gamma(\theta)}$$

where $\lambda > 0$, $\theta > 0$, and $\delta < 0$.

E.3 Discrete Convolution

The discrete convolution of the kernel f with the measure η^m at time t is

$$(f * \eta^m)\big(t \mid \lambda, \theta, \delta\big) \;=\; \sum_{(t', m') \in \mathcal{H}_t} m' \, f\big(t - t' \;\big|\; \lambda, \theta, \delta\big).$$

This sum effectively weighs each past mark m' by the kernel evaluated at the lag $t - t'$.

F Branching Ratio with Reader-Specific Effects

In a hawkes process with exponential temporal decay, the branching ratio (Laub et al., 2015) is defined as $\dfrac{\alpha}{\beta}$ and it quantifies the average number of events directly triggered by a single event. The higher the branching ratio, the more subsequent events are generated by each past events; that is the influence that past events have on future occurrences is greater. For reader r, our model specifies covariate-dependent parameters:

$$\alpha^r = \underbrace{\alpha + \gamma_\alpha^r}_{\text{Baseline}} + \alpha_m m_i, \quad \beta^r = \underbrace{\beta + \gamma_\beta^r}_{\text{Baseline}} + \beta_m m_i \tag{12}$$

where m_i is the mark (e.g., fixation duration), $\gamma_\alpha^r, \gamma_\beta^r$ are reader-specific random effects, and α_m, β_m are fixed coefficients. The *reader-specific branching ratio* for a fixation with mark m_i is:

$$\eta^r(m_i) = \frac{\alpha^r}{\beta^r} = \frac{(\alpha + \gamma_\alpha^r) + \alpha_m m_i}{(\beta + \gamma_\beta^r) + \beta_m m_i} \tag{13}$$

F.1 Mark Modulation of Influence

The derivative with respect to m_i reveals how marks modify temporal influence within readers:

$$\frac{\partial \eta^r}{\partial m_i} = \frac{\alpha_m(\beta + \gamma_\beta^r) - \beta_m(\alpha + \gamma_\alpha^r)}{\Big[(\beta + \gamma_\beta^r) + \beta_m m_i\Big]^2} \tag{14}$$

When $\alpha_m > \beta_m$ and the baseline ratio satisfies stationarity ($\frac{\alpha+\gamma_\alpha^r}{\beta+\gamma_\beta^r} < 1$) - which all our estimates do - the numerator simplifies to:

$$\alpha_m(\beta + \gamma_\beta^r) - \beta_m(\alpha + \gamma_\alpha^r) > \beta_m \left[\frac{\alpha + \gamma_\alpha^r}{\beta + \gamma_\beta^r}(\beta + \gamma_\beta^r) - (\alpha + \gamma_\alpha^r)\right] = 0 \tag{15}$$

where the inequality follows from $\alpha_m/\beta_m > (\alpha + \gamma_\alpha^r)/(\beta + \gamma_\beta^r)$. Thus, $\partial \eta^r / \partial m_i > 0$ - larger marks increase the branching ratio for fixations from the same reader.

G More Details on Experimental Setup

Our models are implemented in PyTorch and we estimated model parameters using gradient-based techniques. A systematic hyperparameter search was conducted over batch sizes $\{64, 128, 256, 512\}$, learning rates $\{0.0001, 0.001, 0.01, 0.1\}$, and weight decay values $\{0, 0.0001\}$. In our experiments, we compared different optimizers, including the Adam optimizer and SGD with Nesterov momentum.

To reduce the risk of converging to different local optima across experiments, we employed a progressive training strategy. We trained models sequentially, moving from simpler to more complex architectures in terms of the number of parameters. When training a more complex model, we initialized its shared parameters with the best values from the corresponding nested simpler model, and set any additional parameters to 0 (when possible). This approach was designed to mitigate the possibility of converging to an optimum that had not been explored in the simpler model, by ensuring that each more complex model started from the same local minimum as its less complex predecessor.