A Spatio-Temporal Point Process for Fine-Grained Modeling of Reading Behavior

Francesco Ignazio Re[1] **Andreas Opedal**[1,2] **Glib Manaiev**[1]
Mario Giulianelli[1] **Ryan Cotterell**[1]
[1]ETH Zürich [2]Max Planck Institute for Intelligent Systems, Tübingen
{francesco.re, andreas.opedal, ryan.cotterell}@inf.ethz.ch

Abstract

Reading is a process that unfolds across space and time, alternating between fixations where a reader focuses on a specific point in space, and saccades where a reader rapidly shifts their focus to a new point. An ansatz of psycholinguistics is that modeling a reader's fixations and saccades yields insight into their online sentence processing. However, standard approaches to such modeling rely on aggregated eye-tracking measurements and models that impose strong assumptions, ignoring much of the spatio-temporal dynamics that occur during reading. In this paper, we propose a more general probabilistic model of reading behavior, based on a marked spatio-temporal point process, that captures not only *how long* fixations last, but also *where* they land in space and *when* they take place in time. The saccades are modeled using a Hawkes process, which captures how each fixation excites the probability of a new fixation occurring near it in time and space. The duration time of fixation events is modeled as a function of fixation-specific predictors convolved across time, thus capturing spillover effects. Empirically, our Hawkes process model exhibits a better fit to human saccades than baselines. With respect to fixation durations, we observe that incorporating contextual surprisal as a predictor results in only a marginal improvement in the model's predictive accuracy. This finding suggests that surprisal theory struggles to explain fine-grained eye movements.

https://github.com/rycolab/spatio-temporal-reading

1 Introduction

Reading is a cognitively complex skill. 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 psycholinguistics is that eye movements during reading provide a direct window into the cognitive processes underlying language comprehension (McConkie, 1979; Just and Carpenter, 1980; Rayner et al., 1989; Findlay and Walker, 1999). Based on this premise, 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 studies, termed scanpaths, consist of sequences of the participants' fixations on a text displayed on a two-dimensional coordinate space, e.g., a screen placed in front of the participant, along with the fixations' durations and onset time. In modern computational psycholinguistic studies, such raw data is typically aggregated into summary measurements, such as 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. These summary measurements are then treated as dependent variables in a (generalized) linear model during analysis (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 fixations where the reader moves back to a previously fixated point (regressive saccades), which are posited to correspond to different cognitive processes (Wilcox et al., 2024). From a spatial perspective, aggregations inherently rely on pre-defined regions of interest (Giulianelli et al., 2024a). Aggregating fixations, which is most commonly done at the word level, discards any information about *where* saccades land within the boundaries of a word and hinders

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30518–30538

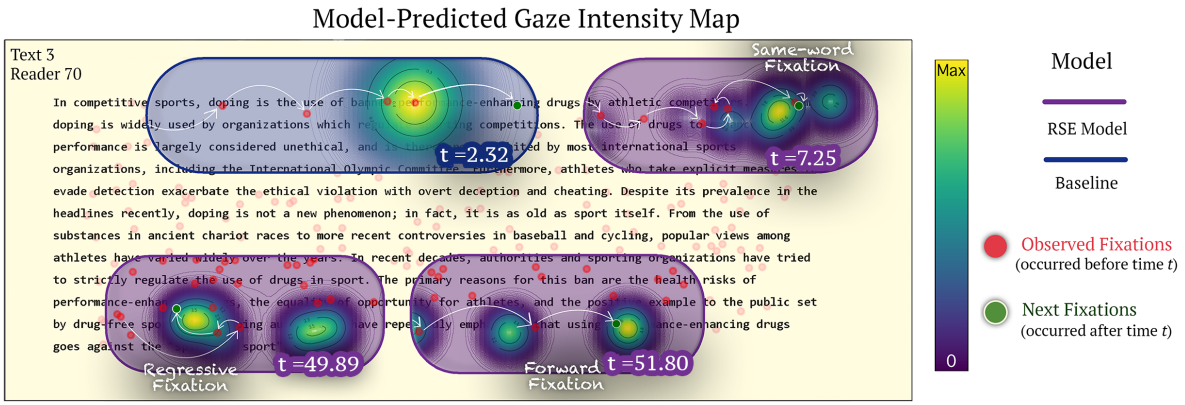

Figure 1: Illustration of how the intensity function of the saccade model evolves over time within a scanpath, and how model predictions compare to held-out data. The intensity function is shown at selected timestamps and visualizes the predicted fixation density of two models: our reader-specific effects (RSE) Hawkes process (in purple), which accounts for reader-specific effects and directional saccade tendencies (e.g., forward shifts in reading), and a last-fixation baseline model (in blue), which concentrates density around the most recent fixation location; see details in §3.1 and §4.1. Red dots indicate observed fixations prior to time t, while green dots mark the fixation immediately following t. Note how the RSE model captures key reading behaviors, including forward saccades ($t = 51.80$), backward regressions ($t = 49.89$), and re-fixations on the same word ($t = 7.25$).

investigations into smaller linguistic units, e.g., syllables or morphemes. In sum, while aggregations help simplify the challenge of modeling and interpreting the complex spatio-temporal dynamics of fixations and saccades, they inevitably throw away information compared to the raw reading data.

Moreover, the manner in which scanpaths are aggregated also impacts the empirical support for a given theory. E.g., surprisal theory (Hale, 2001; Levy, 2008), the theory that the processing effort for a linguistic unit depends on its in-context information content, suggests that contextual word predictability should have a more pronounced effect on gaze duration than total fixation time, because the latter can be influenced by material from the right context through regressive saccades. But, counterintuitively, next-word surprisal has been found empirically to be a stronger predictor of total fixation time than gaze duration (Wilcox et al., 2023). Because both gaze and total duration are aggregate measures, providing a precise explanation for such results is challenging. Another theoretical concern with aggregations is that they tend to conflate cognitive and oculomotor control processes. While cognitive processes, e.g., lexical access, contribute to reading slowdowns (Mollica and Piantadosi, 2017), the oculomotor system also imposes physiological time delays between successive saccades, suggesting that reading times may reflect an interplay of both cognitive and mechanical constraints (Salthouse and Ellis, 1980; Rayner et al., 1983).

In this paper, we advocate a unified approach that jointly models *when* fixations occur, *where* they land, and *how long* they last. We achieve this by constructing a marked spatio-temporal point process, which alternates between generating saccades and fixations. To model saccade timing and fixation locations (*when* and *where*), we employ a spatio-temporal Hawkes (1971) process, which captures how the density of future fixations changes in response to preceding ones in both time and space; Fig. 1 gives an illustration. To model fixation durations (*how long*), we adopt a log-normal distribution with a convolution-based approach inspired by Shain and Schuler (2018, 2021). We evaluate our probabilistic model on a reprocessed (App. A) version of the English portion of the MECO corpus (Siegelman et al., 2022), assessing its ability to jointly model spatio-temporal disaggregated fixation and duration patterns.

Empirically, we uncover several factors that yield better models of saccade planning, including, e.g., spatio-temporal dependencies on previous saccades, the ability to model the tendency towards a left-to-right progression of eye movement, and reader-specific effects. Interestingly, these improvements are the most prevalent when fixations that occur outside of a character's bounding box are included in modeling. When additionally including predictors like contextual surprisal, however, we observe only tiny improvements in the model's ability to perform saccade planning. Furthermore, in our experiments on fixation durations, we obtain results that may have negative implications for established theories on human language processing. First, we observe little improvement when incorporating the effect of unboundedly many previous fixations through the convolution-based technique in comparison to a model that employs a Markov assumption when modeling past fixations. This find-

ing suggests that the effect of previous fixations on subsequent ones may indeed be bounded. Second, when incorporating predictors such as contextual surprisal, unigram surprisal, and length, we obtain effects that are an order of magnitude smaller when modeling scanpaths as compared to modeling aggregated measurements. This finding suggests that some effects reported in the literature based on aggregated scanpaths may appear larger than they would if raw scanpath data were modeled directly.

2 Modeling Reading Data

While reading, our eyes make progress through the text via brief, rapid movements called **saccades**. Very little visual information is extracted during any one saccade (Ishida and Ikeda, 1989). Instead, most of the information is extracted during the pauses that occur between saccades, where the eyes remain (mostly) stationary.[1] These pauses, which are generally much longer than saccades, are called **fixations**. In these terms, reading can be thought of as alternating between fixations and saccades.

Examining reading behavior is one method to better understand the cognitive mechanisms that underlie reading and language processing more generally. 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.

Formalization. Formally, each fixation is characterized as a triple $(t, \mathbf{s}, d)$ consisting of an **onset time** t, a **spatial location** $\mathbf{s}$, and a **duration** d. The onset time $t \in \mathbb{R}_{\geq 0}$ is the starting time of the fixation relative to some reference point, typically the start of the recording. The spatial position lives in a bounded two-dimensional coordinate space $\Omega \subset \mathbb{R}^2$, for example, a screen. Finally, the duration $d \in \mathbb{R}_{>0}$ captures how long the eye remains still before initiating the next saccade. We consider a (**full**) **scanpath** $\mathcal{T}$ to be a set of N fixations, i.e.,

$$\mathcal{T} \stackrel{\text{def}}{=} \{(t_n, \mathbf{s}_n, d_n)\}_{n=1}^{N}, \tag{1}$$

where $t_m < t_n$ if $m < n$. For each $n \in \{1, \ldots, N\}$, we define the history $\mathcal{H}_n$ as

$$\mathcal{H}_n \stackrel{\text{def}}{=} \{(t_m, \mathbf{s}_m, d_m) \mid t_m \leq t_n\}. \tag{2}$$

Note that in addition to fixations' onset times, locations, and durations, this sequence encodes the onset times and durations of saccades as well.[2] Furthermore, we note that Eq. (1) includes the full sequence of fixations and saccades that occur in Ω during an eye-tracking session. This includes fixations that do not land on any word, i.e., those that land outside character bounding boxes, re-fixations on different parts of the same word, and regressive saccades. We also consider **filtered scanpaths**, i.e., the subsequence of a full scanpath that only contains fixations that fall within the bounding box of some word.[3]

2.1 Modeling Details

Aggregations. In most sentence processing experiments based on eye-tracking data, the raw data are typically aggregated into reading time variables at the word level; see (Frank et al., 2013).[4] Some of the most common word-level aggregations are:

- **first fixation duration**, the duration of the first fixation that lands on a word;
- **gaze duration**, the sum of the durations of all fixations that land on a word before leaving it the first time;
- **total fixation duration**, the summed duration of all the fixations on a word; and
- **scanpath duration**, the summed duration of all *consecutive* fixations that land on the same word.

Note that the first three aggregation strategies are ordered in a nested fashion, i.e., each strategy aggregates over at least as many fixations as the previous; for more details, see Inhoff (1984), Berzak and Levy (2023), and App. B.1. The final aggregation strategy, scanpath duration, is *not* nested in the sense above. Moreover, a sequence of scanpath durations may contain multiple durations

[1] The eyes do not remain *completely* stationary; there are constantly minor movements in the form of tremors, small drifts, and microsaccades (Rayner, 1998).

[2] In particular, the onset of the n^{th} saccade can be inferred by adding the n^{th} duration to the n^{th} fixation onset. The duration of the saccades can be inferred by taking the difference between the fixation onsets and the saccade onsets.

[3] Fixations outside word bounding boxes account for 52.3% of all fixations in our dataset. Note that both the size of a bounding box and the strategy for assigning fixations to words are determined at the discretion of the modeler.

[4] Aggregations into other regions of interest are also used (Cook and Wei, 2019; Brodbeck et al., 2022)—especially when studying specific types of reading behavior like skip rates (Rayner et al., 2011; Giulianelli et al., 2024a).

that correspond to the same word. See Shain and Schuler (2021) for more details.

Statistical analysis. The standard approach to analyzing aggregated measurements is to use linear modeling (Kliegl, 2007; Giulianelli et al., 2024b; Kuribayashi et al., 2024, 2025) or generalized additive modeling (GAMs; Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020, 2023; Gruteke Klein et al., 2024). When multiple reading times per stimulus are available, e.g., when multiple participants read the same stimulus in their respective trials, it is appropriate to apply a mixed-effects model (e.g., Aurnhammer and Frank, 2019; Xu et al., 2023), using random effects to account for variability across participants. In §4, we investigate whether incorporating reader-specific effects improves the modeling of saccades and fixation durations, estimating them as fixed effects rather than random effects.

Spillover effects. Beyond aggregation, there is another limitation of the above-mentioned modeling techniques worth noting. The effect that cognitive processing of a unit has on reading time is not necessarily instantaneous; it often leads to reading slowdowns for units that occur later in the text (Shain and Schuler, 2021). To incorporate such effects, these models must include additional predictors describing previous units, which are called **spillover variables**. We discuss the modeling of spillover variables in more detail in §3.2.

Predictors. It is common to use predictors derived from pre-trained language models when modeling reading data in order to estimate the effect that the prior context has on reading time (e.g., Oh and Schuler, 2023). Here, we focus on one such predictor called **contextual surprisal**. Let Σ be a finite, non-empty set of linguistic units, e.g., characters or words, called an **alphabet**, and let Σ^* be the set of all strings that can be formed by concatenating units in Σ, including the empty string ε. We further employ a special symbol $\text{EOS} \notin \Sigma$ to denote the end of a string, and define $\overline{\Sigma} \stackrel{\text{def}}{=} \Sigma \cup \{\text{EOS}\}$. Following Shannon's (1948) formulation of information content, the **contextual surprisal** of a unit $w_t \in \overline{\Sigma}$ given a preceding context of units $\boldsymbol{w}_{<t} \in \Sigma^*$ is defined as

$$s_t(w_t) \stackrel{\text{def}}{=} -\log_2 \overrightarrow{p}(w_t \mid \boldsymbol{w}_{<t}), \qquad (3)$$

where $\overrightarrow{p}(\cdot \mid \boldsymbol{w}_{<t})$ is the true (albeit unknown) distribution over $\overline{\Sigma}$ conditioned on the context $\boldsymbol{w}_{<t}$, defined using prefix probabilities; see, e.g., Opedal et al., 2024, for further technical discussion.

Surprisal theory. Surprisal theory (Hale, 2001; Levy, 2008) predicts that the cognitive processing effort for a linguistic unit w is a function of its contextual surprisal. This theory has received empirical support across several datasets and languages (Kuribayashi et al., 2021; Wilcox et al., 2023), and there is evidence suggesting that the linking function between surprisal and different aggregated measurements of processing effort is well-approximated by an affine function (Smith and Levy, 2013; Xu et al., 2023; Shain et al., 2024). Such studies typically include additional baseline predictors, like unigram frequency and word length (Opedal et al., 2024). Importantly, surprisal theory is a theory about computational demand, situated at the computational level of Marr's (1982) hierarchy. Thus, it neither makes direct predictions about the underlying cognitive mechanisms nor about eye movement control.[5] In our empirical study presented in §4.1, we explore whether surprisal and other predictors are predictive of more fine-grained eye movements than the theory itself considers, i.e., whether they are useful for **saccade planning** (Zingale and Kowler, 1987). App. B.2 gives a brief overview of related cognitive models of reading.

3 Modeling Fine-Grained Reading Data

To capture the fine-grained dynamics of human reading, we now propose a probabilistic model of scanpaths. At the core of our approach is the goal of generating the three key components of a fixation: *when* it begins, *where* it lands on the screen, and *how long* it lasts. Our probabilistic model takes the form of a spatio-temporal *marked* point process, consisting of two components.

- A **spatio-temporal point process** for fixation onsets and locations, characterized by a density function $f(t_n, \mathbf{s}_n \mid \mathcal{H}_{n-1})$ that models the likelihood that the n^{th} fixation in the scanpath $\mathcal{T}$ occurs at time t_n on screen location $\mathbf{s}_n$ given the history of preceding fixations up to time t_{n-1}. See §3.1 for details.
- A **probability distribution** for (non-negative) fixation durations (**marks**), which is characterized by a density function $g(d_n \mid \mathcal{H}_{n-1}, t_n)$

[5] For more discussion, see Ohams et al. (2025) who propose a mechanistic model of human sentence processing based on predictive coding (Friston and Kiebel, 2009).

that models the likelihood of the n^{th} fixation in the scanpath having duration d_n given the last fixation onset t_n and the history of preceding fixations up to time t_{n-1}. See §3.2 for details.

Under this model, a scanpath is iteratively sampled as follows. Note that $\mathcal{H}_0 \stackrel{\text{def}}{=} \{\}$.

(a) **Sample a fixation's onset time and location**:

$$(t_n, \mathbf{s}_n) \sim f(\cdot, \cdot \mid \mathcal{H}_{n-1}), \tag{4}$$

according to the spatio-temporal point process; see Eq. (13) for more details.

(b) **Sample a fixation's duration**:

$$d_n \sim g(\cdot \mid \mathcal{H}_{n-1}, t_n), \tag{5}$$

according to the density of fixation durations; see Eq. (15) for more details.[6]

(c) **Update the history with new fixation**: The full fixation $(t_n, \mathbf{s}_n, d_n)$ is added to the history:

$$\mathcal{H}_n \stackrel{\text{def}}{=} \mathcal{H}_{n-1} \cup \{(t_n, \mathbf{s}_n, d_n)\}. \tag{6}$$

Steps (a)–(c) are iterated until a predefined reading horizon $T \in \mathbb{R}_{\geq 0}$ is reached.[7] Thus, our model can be viewed as a **stopped** spatio-temporal marked point process (Daley and Vere-Jones, 1988).

3.1 Modeling Fixation Onsets and Locations

We model $f(t_n, \mathbf{s}_n \mid \mathcal{H}_{n-1})$ with a spatio-temporal **Hawkes** process (Hawkes, 1971).[8] A Hawkes process is a special case of a spatio-temporal non-homogeneous Poisson point process, which assumes that the probability that a new event occurs in the (infinitesimal) region $[t_n, t_n + \mathrm{d}t_n) \times [\mathbf{s}_n, \mathbf{s}_n + \mathrm{d}\mathbf{s}_n)$ is approximately $\lambda(t_n, \mathbf{s}_n; \mathcal{H}_{n-1})\mathrm{d}t_n\mathrm{d}\mathbf{s}_n$, where $\lambda\colon \mathbb{R}_{\geq 0} \times \Omega \to \mathbb{R}_{\geq 0}$ is called the **intensity function**. A Hawkes process allows past fixations to influence the probability of new fixations in an additive manner. This influence is often termed *self-excitation* in the literature. The intensity function of a spatio-temporal Hawkes process is defined as

$$\lambda(t_n, \mathbf{s}_n; \mathcal{H}_{n-1}) \stackrel{\text{def}}{=} \nu + \sum_{m=1}^{n-1} \phi_m\left(t_n - t_m - \delta(n, m)\right) \psi_m(\mathbf{s}_n), \tag{7}$$

where $\nu \in \mathbb{R}_{\geq 0}$ is the base intensity, $\phi_m\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ are exponentially decaying temporal kernels governing the influence of past events, $\psi_m\colon \Omega \to \mathbb{R}_{\geq 0}$ is a density over the coordinate space Ω, and

$$\delta(n, m) \stackrel{\text{def}}{=} \sum_{j=1}^{n-1} d_j - \sum_{j=1}^{m-1} d_j \tag{8}$$

is the cumulative duration between t_n and t_m. Notationally, we suppress the dependence of ϕ_m, ψ_m, and δ on the history $\mathcal{H}_{n-1}$ for simplicity. Subtracting $\delta(n, m)$ from the time difference in Eq. (7) ensures that the kernel $\phi_m(\cdot)$ only captures the intervals between saccades. Furthermore, we remark that, when generating the n^{th} fixation, the Hawkes process conditions on *all* past fixations $\mathcal{H}_{n-1}$, which is what enables the self-exciting behavior.

Temporal kernel. Our choice of ϕ_m allows each past fixation indexed by m to contribute to the intensity of future fixations. However, that influence decays exponentially the farther back it looks, as determined by $t_n - t_m - \delta(n, m)$ from Eq. (7). We consider an exponential kernel, parameterized as

$$\phi_m(\Delta) \stackrel{\text{def}}{=} h(\mathbf{x}_m^\top \boldsymbol{\alpha}) \exp(-h(\mathbf{x}_m^\top \boldsymbol{\beta}) \cdot \Delta), \tag{9}$$

where $\mathbf{x}_m \in \mathbb{R}^p$ is a column vector of predictor values associated with the m^{th} fixation (e.g., surprisal), $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are column vectors in $\mathbb{R}^p$ of learnable parameters, and $h\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a function that ensures the non-negativity of the dot product, e.g., a ReLU. Note that $h(\mathbf{x}_m^\top \boldsymbol{\alpha})$ quantifies how much a fixation increases the probability of subsequent fixations, its excitation strength, while $h(\mathbf{x}_m^\top \boldsymbol{\beta})$ determines how quickly the influence of the fixation diminishes over time, its decay rate. Because excitation strength and decay rate depend on an event-specific vector of predictor values $\mathbf{x}_m$, the strength of excitation and decay rate may vary across different spatio-temporal conditions.

Spatial density. We model the spatial component ψ_m as a bivariate spherical Gaussian distribution centered at $\mu_m(\mathbf{s}_m)$, where $\mu_m\colon \mathbb{R}^2 \to \mathbb{R}^2$ is a

[6] The density $g(\cdot \mid \mathcal{H}_{n-1}, t_n)$ could also have been conditioned on the n^{th} screen location $\mathbf{s}_n$; we chose not to do so based on results obtained from preliminary experiments.

[7] An application of our model that we do not emphasize here is generating synthetic scanpaths. Besides being interesting for cognitive modeling, synthetic scanpaths have been used both to enhance language representations (Barrett et al., 2018; Sood et al., 2020b) and to probe the internal behavior of neural NLP models (Sood et al., 2020a; Hollenstein et al., 2022).

[8] Spatio-temporal Hawkes processes (Reinhart, 2018) have several other applications, e.g., modeling earthquakes and their aftershocks (Ogata, 1999), urban crime (Mohler et al., 2011), and spread of infectious diseases (Meyer et al., 2012).

fixation-specific transformation function; we describe specific choices of μ_m in the paragraph below. These choices result in the following density

$$\psi_m(\mathbf{s}_n) \stackrel{\text{def}}{=} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{\|\mathbf{s}_n - \mu_m(\mathbf{s}_m)\|^2}{2\,\sigma^2}\right), \quad (10)$$

where $\|\cdot\|$ denotes the Euclidean (L_2) norm. This design choice allows us to interpret the intensity function introduced in Eq. (7) as a conical combination of Gaussian densities, each corresponding to a past fixation event. Thus, the intensity function can be seen as proportional to the density of a multimodal distribution over Ω. We contend that this property aligns well with what is known about human reading behavior because it captures that eye fixations may jump either forward or backward in the text with varying probabilities (Rayner, 1998).

We consider three different choices of the transformation function μ_m. As a baseline, we first consider $\mu_m^{\text{b}}(\mathbf{s}) \stackrel{\text{def}}{=} \mathbf{s}$, i.e., the identity function. However, this baseline yields an overly simple model: (i) it predicts that saccades are most likely to occur around previous fixations, and (ii) it does not incorporate other predictors that may influence the location of the next fixation. To address (i), we parameterize μ_m as an affine transformation of the previous fixation location $\mathbf{s}$. Mathematically, this means that $\mu_m(\mathbf{s})$ takes the following affine form

$$\mu_m^{\text{a}}(\mathbf{s}) \stackrel{\text{def}}{=} \mathbf{A}\mathbf{s} + \mathbf{b}, \quad (11)$$

where $\mathbf{A} \in \mathbb{R}^{2\times 2}$ is a matrix and $\mathbf{b} \in \mathbb{R}^2$ is a vector. To address (ii), we consider the same vector of predictors $\mathbf{x}_m \in \mathbb{R}^p$ applied into the temporal kernel and incorporate predictor-specific effects, yielding the fully parameterized transformation function

$$\mu_m^{\text{f}}(\mathbf{s}) \stackrel{\text{def}}{=} \mu_m^{\text{a}}(\mathbf{s}) + \mathbf{C}\mathbf{x}_m, \quad (12)$$

where $\mathbf{C} \in \mathbb{R}^{2\times p}$ is a matrix.

The density function. Given the intensity function defined in Eq. (7), the corresponding probability density function of a Hawkes process is given by

$$f(t_n, \mathbf{s}_n \mid \mathcal{H}_{n-1}) \quad (13)$$
$$\stackrel{\text{def}}{=} \frac{\lambda(t_n, \mathbf{s}_n; \mathcal{H}_{n-1})}{\exp(\Lambda(t_n; \mathcal{H}_{n-1}))} \mathbb{1}\{t_n \geq t_{n-1} + d_{n-1}\}.$$

The indicator function ensures that no probability mass is distributed to preceding time points and

$$\Lambda(t_n; \mathcal{H}_{n-1}) \quad (14)$$
$$\stackrel{\text{def}}{=} \int_{t_{n-1}+d_{n-1}}^{t_n} \int_{\Omega} \lambda(u, \boldsymbol{z}; \mathcal{H}_{n-1}) \mathrm{d}\boldsymbol{z}\mathrm{d}u.$$

See Daley and Vere-Jones (1988, §7.2) for a detailed technical discussion.

3.2 Modeling Fixation Durations

Fixation durations are modeled as non-negative random variables with a conditional density denoted by $g(d_n \mid \mathcal{H}_{n-1}, t_n)$. Let $d_n > 0$ be the duration of the n^{th} fixation. We assume a log-normal distribution for fixation durations,[9] i.e.,

$$g(d_n \mid \mathcal{H}_{n-1}, t_n) \stackrel{\text{def}}{=} \quad (15)$$
$$\frac{1}{d_n\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log d_n - \xi_n(t_n))^2}{2\sigma^2}\right),$$

where $\xi_n(t_n)$ is a fixation-specific function of the fixation onset t_n and the past fixation history $\mathcal{H}_{n-1}$. For clarity, again, we suppress the explicit dependence of $\xi_n(t_n)$ on $\mathcal{H}_{n-1}$ in the notation. Note that $\sigma^2 > 0$ is a fixed variance parameter, different than σ^2 in Eq. (10).

Conditional log-mean of duration. Similar to Shain and Schuler (2018, 2021), we model temporally delayed spillover effects through a convolution over past predictor values. Let $\mathbf{x}_m \in \mathbb{R}^p$ be a column vector of predictor values corresponding to the m^{th} fixation and $K \subseteq \{1, \ldots, p\}$ be a set of indices representing the predictors for which we model spillover effects. We then define the convolution function

$$\xi_n^{\text{c}}(t_n) \stackrel{\text{def}}{=} \mathbf{x}_n^\top \mathbf{w} \quad (16)$$
$$+ \sum_{k\in K} w'_k \sum_{m=1}^{n-1} x_{mk} \gamma\left(t_n - t_m \mid \alpha_k, \beta_k, \theta_k\right),$$

where $\mathbf{w} \in \mathbb{R}^p$ is a column vector of parameters, x_{mk} denotes the value of predictor k at fixation m, $w'_k \in \mathbb{R}$ is the spillover coefficient of predictor k, and the kernel $\gamma(\cdot \mid \alpha_k, \beta_k, \theta_k)$ is the density function of a shifted gamma distribution, defined as

$$\gamma(\tau \mid \alpha_k, \beta_k, \theta_k) \stackrel{\text{def}}{=} \quad (17)$$
$$\frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} (\tau + \theta_k)^{\alpha_k - 1} \exp\left(-\beta_k(\tau + \theta_k)\right),$$

with parameters $\alpha_k \in \mathbb{R}_{>1}$, $\beta_k \in \mathbb{R}_{>0}$ and $\theta_k \in \mathbb{R}_{\geq 0}$; $\Gamma(\cdot)$ denotes the gamma function.

[9] We evaluated several candidate distributions using K-fold cross-validation and found that the log-normal distribution offers a good trade-off between interpretability and goodness-of-fit; it outperformed alternatives such as the Weibull, exponential, and Rayleigh distributions. See App. D for details.

Markovian spillovers. Most previous work (see §2.1) incorporates spillover effects as additional predictors in a linear model or GAM. In our framework, this corresponds to replacing the gamma kernel in Eq. (16) with spillover parameters $w'_{mk} \in \mathbb{R}$, yielding the Markov model:

$$\xi_n^{\mathrm{M}}(t_n) \stackrel{\text{def}}{=} \mathbf{x}_n^\top \mathbf{w} + \sum_{k \in K} \sum_{m=n-l}^{n-1} x_{mk} w'_{mk}, \quad (18)$$

where l denotes the number of previous fixations that contribute to the spillover effect, i.e., the model's Markov order. Each predictor x_{mk} represents the value of feature $k \in K$ at time step m, and $w'_{mk} \in \mathbb{R}$ is the corresponding weight applied to that feature and lag. The double subscript emphasizes that the model learns distinct weights for each combination of feature and lag.

4 Experiments and Results

We employ the modeling framework introduced in §3 to run two suites of experiments distinguished by the scanpath components they target, i.e., saccade planning (§4.1) and fixation durations (§4.2). The experimental setup is detailed in App. A. In each of the subsequent subsections, we introduce the hypotheses we are testing, the model specifications we use, and present experimental results.

4.1 Modeling Saccade Planning

In this section, we empirically study how well the spatio-temporal Hawkes process fits to saccades in reader scanpaths. We are interested in a number of different questions. First, we ask which components of the model are useful for modeling saccades. To this end, we introduce several model specifications, including a learned constant spatial shift and reader-specific effects. We also introduce two simple baseline models. Second, we are interested in whether saccades can be modeled using filtered scanpaths, or whether this requires more fine-grained data. Finally, we turn our attention towards the predictors. We investigate whether incorporating word and character-level quantities as predictors associated with past fixations helps determine where and when the next saccade will land, beyond what is captured by constant displacements from past fixations and individual reader effects.

4.1.1 Model Specifications

Baseline models. We obtain baseline models by simplifying the intensity function from Eq. (7). The main baseline model is a **standard Hawkes process**. It defines the excitation strength α and decay rate β in Eq. (9) as scalar values, and $\mu_m(\cdot)$ from Eq. (10) as the identity function. We also include two simpler models: a **last-fixation** baseline which models each fixation as normally distributed around the previous one with constant variance, and a **Poisson** baseline which assumes fixations to be uniformly and independently distributed in space and time. Further details are given in App. C.

More sophisticated models. Next, we introduce models that are more expressive than our baselines. Our first model is the **constant spatial shift** (CSS) model, which predicts new fixation locations based on a constant displacement from previous fixations. This term captures a baseline tendency for gaze to progress linearly across the text, independent of specific word content or surrounding linguistic context. It uses the spatial mean from Eq. (11). This model and our baseline models treat all scanpaths as if a single average reader generated them; however, this treatment deviates from the reality that all individuals exhibit distinct reading styles. To capture such variations, we include reader-specific fixation effects in both the temporal and spatial parameters for each of the $R \in \mathbb{Z}_{>0}$ readers. We consider a vector of predictors

$$\mathbf{x}_m \stackrel{\text{def}}{=} \mathbf{1} \oplus \mathbf{u}_m, \quad (19)$$

where $\mathbf{1} \in \mathbb{R}^1$ is the 1-dimensional column vector with value 1 corresponding to the intercept, $\mathbf{u}_m \in \{0, 1\}^R$ is a unit vector with a single non-zero value that corresponds to a particular reader; $\oplus$ denotes column-wise concatenation. As a result, the dot products $\mathbf{x}_m^\top \boldsymbol{\alpha}$ and $\mathbf{x}_m^\top \boldsymbol{\beta}$ will be the sum of a global effect and a reader-specific effect. Note that this influences both the temporal (Eq. 9) and the spatial kernel (Eq. 12). We refer to this model as the **reader-specific effects** (RSE) model. Finally, we treat each predictor associated with a fixation as an effect that modulates both the temporal kernel and the spatial mean. In the case of a single effect z, we let the predictor vector be equal to

$$\mathbf{x}_m \stackrel{\text{def}}{=} \mathbf{1} \oplus \mathbf{u}_m \oplus z_m \mathbf{1} \oplus z_m \mathbf{u}_m. \quad (20)$$

Here, $z_m \mathbf{1}$ captures a global effect and $z_m \mathbf{u}_m$ captures its interaction with the reader identity encoded in $\mathbf{u}_m$. For saccades that do not land on a word and therefore lack a corresponding effect value, we include binary indicators denoting the

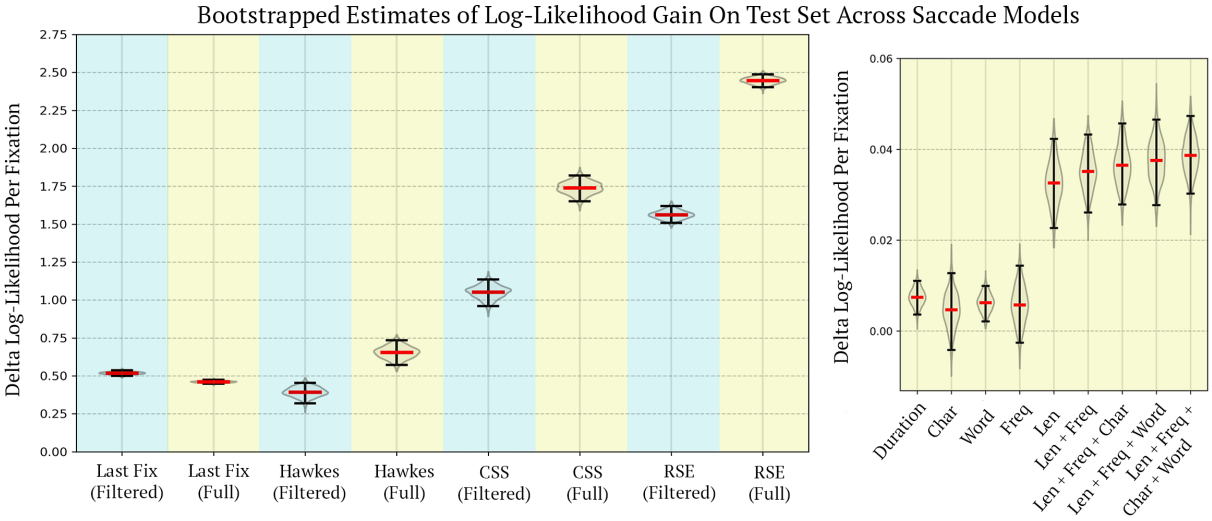

Figure 2: Bootstrapped distributions of per-fixation log-likelihood gains for different saccade models, evaluated relative to the Poisson baseline (left) and the RSE model (right). Higher values indicate better predictive performance in modeling saccade behavior. The models in the left-hand panel are described in §4.1, while the models in the right panel include the predictors introduced in App. A. We distinguish between whether the model was trained and evaluated on the full or the filtered scanpaths. The right-hand panel reports results only for the full dataset; an extended version is given in Fig. 4 (App. E). For the predictors, we write "word" for word-level surprisal, "char" for character-level surprisal, and "freq" for unigram surprisal.

presence of the effect, along with their reader interactions, in Eq. (12).

4.1.2 Results

Fig. 2 shows bootstrapped estimates of per-fixation log-likelihood improvements for the various saccade models evaluated on the held-out test set, as specified in App. A. In the left panel, we show the results when using the Poisson process as a baseline. The results demonstrate that the increasingly expressive models specified above, i.e., those resulting from the incorporation of temporal dependencies, the constant spatial shift, and reader-specific modeling, help the model to better generalize to held-out data. The best-performing model in the left panel, obtained by inserting Eq. (19) into Eq. (9) and Eq. (12), achieves an average per-fixation log-likelihood gain of 2.44 nats over the Poisson baseline. This corresponds to the model assigning approximately $(\exp(2.44) - 1) \times 100 \approx 1047\%$ higher likelihood to the true next fixation on average. Parameter estimates from this model suggest a consistent global rightward shift of roughly ≈ 10.61 characters (including white spaces) further to the right, a pattern consistent with the left-to-right progression of (English) writing. There is also evidence of self-excitation. Given estimates $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ of the parameters in Eq. (9), we find $h(\mathbf{x}_m^\top \widehat{\boldsymbol{\alpha}}) = 12.64 \pm 2.69$ and $h(\mathbf{x}_m^\top \widehat{\boldsymbol{\beta}}) = 16.24 \pm 3.44$, with values corresponding to the mean and standard deviation across individual readers, and $\mathbf{x}_m$ as defined in Eq. (19). However, the lower magnitude of $h(\mathbf{x}_m^\top \widehat{\boldsymbol{\alpha}})$ in comparison to $h(\mathbf{x}_m^\top \widehat{\boldsymbol{\beta}})$ suggests that, although recent fixations do influence the likelihood of another fixation in short succession, this influence decays relatively quickly over time.

The right panel of Fig. 2 shows the marginal effect of adding lexical predictors on top of the RSE model. While improvements are statistically significant, they are modest in magnitude: most predictors yield under 2% relative gain in per-fixation log-likelihood, a much smaller increase than what can be observed in the left panel. Word length displays the highest predictive power, leading to performance improvements by approximately 4%. This may be due to the fact that a longer word, in general, requires a longer saccade to move to future words.

Finally, in Fig. 2, we observe that models trained on full scanpaths place a higher probability on held-out data than those trained on filtered scanpaths. This pattern suggests that it is useful to preserve the full sequence of preceding fixations, rather than only those associated with words. See Fig. 4 for an extended comparison.

4.2 Modeling Fixation Durations

In this section, we focus on modeling fixation durations through the log-normal, convolution-based model given in Eq. (16), comparing it to the linear Markov model given in Eq. (18). In addition to modeling the full scanpaths, we also fit linear models to the three aggregated datasets described in App. A. We are interested in whether surprisal and other predictors are effective across several

levels of aggregation, or whether the empirical support for, e.g., surprisal theory, depends on the aggregation strategy. For the linear models, we log-transform the response variable.

4.2.1 Model Specifications

Similarly to in §4.1.1, we introduce reader-specific effects and the predictors described in App. A into the predictor vector $\mathbf{x}_n$. Specifically, we define $\mathbf{x}_n$ analogously to Eqs. (19) and (20), respectively, where each predictor is associated with the n^{th} fixation. Each predictor vector is then used in Eq. (16) and Eq. (18). When applying Eq. (18), we set the number of previous fixations $l = 2$. The baseline models contain intercepts, including reader-specific ones when modeling non-aggregated data, as well as durations of the fixations in the history. The latter are included as controls in order not to attribute the effects of past durations to the predictors. We also perform experiments in which the baseline *excludes* the prior fixation durations to get a sense of how much the effect sizes reduce.

4.2.2 Results

Fig. 5 (in App. E) plots bootstrapped distributions of per-fixation log-likelihood improvements for models that incorporate spillover effects via a convolution term, evaluated on the full and filtered scanpath datasets. Note that these are the only datasets that retain fixation-level timing. We observe very small and mostly insignificant effect sizes: With the exception of the model combining length, unigram suprisal, and character-level surprisal, none of the intervals exclude zero, and all average effect sizes remain below 0.004. Comparing the same models on the full versus filtered scanpaths, we generally observe slightly larger average improvements for the filtered scanpaths. This raises the question of whether different data aggregation strategies could recover the larger effect sizes previously reported (e.g., Wilcox et al., 2023).

In Fig. 6 (App. E), we present results for the Markov linear models introduced in Eq. (18) across differently aggregated datasets and across predictors. First, comparing the log-likelihood improvements with those in Fig. 5 shows that the convolution-based model and the linear Markov model achieve similar fit. Notably, the convolution-based model, despite its greater expressiveness, does not yield larger improvements than the linear model. These results suggest that it is effective in practice to model fixation durations with a linear model and Markov assumption on the spillover effects, at least on the MECO dataset.

Comparing across aggregation schemes, we find that averaging gaze durations across readers leads to substantially larger mean log-likelihood improvements across folds, ranging from approximately 0.023 to 0.037 when the baseline includes duration spillovers (see Fig. 6); this is indeed consistent with the effect sizes reported in Fig. 1 from Wilcox et al. (2023). In contrast, disaggregated (reader-specific) models yield improvements that are more than an order of magnitude smaller, with effect sizes of up to only 0.003 relative to the same spillover baselines. Aggregation also introduces greater variance, with the effect sizes in the aggregated setting ranging from 0.01 to 0.05 across folds, when accounting for duration spillovers. As for the filtered and full scanpaths, the effect sizes reduce even further. Furthermore, we find that effects are consistently lower when past durations are included in the baseline model (cf. §4.2.1). We thus advocate for the continued use of such more stringent baselines in future studies. Finally, we note that the absolute effects on aggregated data are somewhat smaller than what has been reported in previous studies on MECO data (Figs. 4 and 5, Opedal et al., 2024). This may be due to differences in preprocessing: while we map fixations to words via bounding boxes (see App. A), Siegelman et al. (2022) appear to have used a different method.

5 Conclusion

We introduced a probabilistic model of reading behavior, through which we investigate modeling strategies for saccade planning and evaluate whether fine-grained models of individual fixation durations outperform approaches based on aggregated gaze measurements. With respect to saccade planning, our findings highlight the importance of including the complete fixation history while also accounting for systematic rightward tendencies in gaze as well as reader-specific behavior. We also observe that surprisal-based predictors contribute minimal additional predictive power. When modeling fixation durations, we similarly observe that incorporating surprisal-based predictors in unfiltered scanpath models only yields marginal improvements in predictive power. Taken together, our findings suggest that commonly used aggregation and data processing procedures can substantially influence the outcomes of reading time analyses.

Limitations

This study contributes to the literature on modeling techniques for psycholinguistic data, but several limitations warrant discussion. First, in terms of data, our analysis is restricted to the English portion of MECO, leaving the other ten 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 are read from right to left). In addition, the relationship between the predictors and the responses was intentionally constrained to be affine, consistent with findings from previous work; see §2.1. This may oversimplify the relationship between cognitive processes and eye movements compared to more expressive modeling choices.

With respect to modeling choices, another limitation—raised during the reviewer discussion—concerns our assumption of isotropic variance in the kernel for the spatial density function (Eq. 10). While this assumption was made out of simplicity, it may limit the model's ability to capture directional biases in eye movements. A natural extension would be to replace the isotropic kernel with an anisotropic one, or to apply non-linear warping functions to better accommodate complex spatial patterns in fixation behavior.

One reviewer also raised the importance of capturing structured spatial effects, particularly those related to line transitions, and we explored one such direction. Specifically, we experimented with augmenting the model with a feature encoding the distance from the current fixation to the right margin, activated via a ReLU function to signal likely transitions to a new line. However, we encountered difficulties during training, as this feature failed to produce stable non-zero gradients, possibly due to the sparsity of fixations near line endings. For this reason, we chose not to include this model variant in the final set of results presented here.

Additional directions for extending the modeling framework can be drawn from existing literature. For instance, the framework could be extended to support neural Hawkes processes (Mei and Eisner, 2017), including for the modeling of individual reader effects (Boyd et al., 2020). Another interesting direction would be to study counterfactual reading scenarios—for example, asking questions about how scanpaths are influenced by external disturbances that may disrupt reading. There already exist techniques to estimate causal effects under temporal point processes (Gao et al., 2021; Noorbakhsh and Rodriguez, 2022; Zhang et al., 2022), which could be applied and extended to the setting considered here.

Regarding the experimental results, when training the PyTorch models, we conducted a grid search over a range of hyperparameters to identify the best-performing configurations. However, we cannot guarantee that these configurations correspond to globally optimal parameter estimates. We also note that the character- and word-level surprisal estimates are derived from different language models. Thus, any performance differences between them cannot be attributed solely to the granularity of representation, i.e., character- vs. word-level. Additionally, the standard Hawkes process baseline and the last-fixation baseline exhibited comparable performance. While we extended the Hawkes model to capture spatial gaze patterns, we did not apply similar enhancements to the last-fixation baseline. A more balanced comparison—where both models are equipped with spatial effects—could offer clearer insights into their relative strengths.

Acknowledgments

We thank the reviewers from ACL Rolling Review for their valuable feedback and discussion. AO was supported by the Max Planck ETH Center for Learning Systems. MG was supported by an ETH Zürich Postdoctoral Fellowship. RC gratefully acknowledges support from the Hasler Foundation.

References

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

Maria Barrett, Ana Valeria González-Garduño, Lea Frermann, and Anders Søgaard. 2018. Unsupervised induction of linguistic categories with records of reading, speaking, and writing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2028–2038. Association for Computational Linguistics.

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

Klinton Bicknell and Roger Levy. 2010. A rational model of eye movement control in reading. In *Pro-*

ceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1168–1178. Association for Computational Linguistics.

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

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

Alex Boyd, Robert Bamler, Stephan Mandt, and Padhraic Smyth. 2020. User-dependent neural sequence models for continuous-time event data. In *Advances in Neural Information Processing Systems*, volume 33, pages 21488–21499.

Christian Brodbeck, Shohini Bhattasali, Aura A.L. Cruz Heredia, Philip Resnik, Jonathan Z Simon, and Ellen Lau. 2022. Parallel processing in speech perception with local and global representations of linguistic context. *eLife*, 11:e72056.

Charles Clifton, Adrian Staub, and Keith Rayner. 2007. Eye movements in reading words and sentences. In Roger P., G. Van Gompel, Martin H. Fischer, Wayne S. Murray, and Robin L. Hill, editors, *Eye Movements*, chapter 15, pages 341–371. Elsevier.

Anne E. Cook and Wei Wei. 2019. What can eye movements tell us about higher level comprehension? *Vision*, 3(3).

Daryl J. Daley and David Vere-Jones. 1988. *An Introduction to the Theory of Point Processes*. Probability and its applications. Springer-Verlag.

Andrea Gregor de Varda, Marco Marelli, and Simona 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.

Ralf Engbert, Antje Nuthmann, Eike M. Richter, and Reinhold Kliegl. 2005. SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, 112(4):777–813.

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.

Karl Friston and Stefan Kiebel. 2009. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B*, 364(1521):1211–1221.

Tian Gao, Dharmashankar Subramanian, Debarun Bhattacharjya, Xiao Shou, Nicholas Mattei, and Kristin P. Bennett. 2021. Causal inference for event pairs in multivariate point processes. In *Advances in Neural Information Processing Systems*, volume 34, pages 17311–17324.

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

Mario Giulianelli, Andreas Opedal, and Ryan Cotterell. 2024b. Generalized measures of anticipation and responsivity in online language processing. In *Findings of the Association for Computational Linguistics*, pages 11648–11669. 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*, pages 10–18. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Michael Hahn and Frank Keller. 2016. Modeling human reading with neural attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 85–95. 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.

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

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. *Journal of the Optical Society of America*, 6(10):1624–1632.

Marcel A. Just and Patricia A. 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.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2024. Psychometric predictive power of large language models. In *Findings of the Association for Computational Linguistics*, pages 1983–2005. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *Preprint*, arXiv:2502.01615.

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).

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Company.

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

Hongyuan Mei and Jason Eisner. 2017. The neural Hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pages 6754–6764.

Sebastian Meyer, Johannes Elias, and Michael Höhle. 2012. A space-time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics*, 68(2):607–616.

George O. Mohler, Martin B. Short, P. Jeffrey Brantingham, Frederic Paik Schoenberg, and George E. Tita. 2011. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108.

Francis Mollica and Steven T. Piantadosi. 2017. An incremental information-theoretic buffer supports sentence processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 39.

Yurii Nesterov. 1983. A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. In *Doklady Akademii Nauk SSSR*, volume 269, pages 543–547.

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

Kimia Noorbakhsh and Manuel Rodriguez. 2022. Counterfactual temporal point processes. In *Advances in Neural Information Processing Systems*, volume 35, pages 24810–24823.

Dennis Norris. 2006. The bayesian reader: Explaining word recognition as an optimal bayesian decision process. *Psychological Review*, 113(2):327–357.

Yosihiko Ogata. 1999. *Seismicity Analysis through Point-process Modeling: A Review*, pages 471–507. Birkhäuser Basel.

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.

Chiebuka Ohams, Sathvik Nair, Shohini Bhattasali, and Philip Resnik. 2025. A predictive coding model for online sentence processing.

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. 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).

Keith Rayner, Maria L. Slowiaczek, Charles Clifton, and James H. Bertera. 1983. Latency of sequential eye movements: Implications for reading. *Journal of Experimental Psychology: Human Perception and Performance*, 9(6):912—922.

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

Alex Reinhart. 2018. A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3):299–318.

Herbert Robbins and Sutton Monro. 1951. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

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.

Timothy A. Salthouse and Cecil L. Ellis. 1980. Determinants of eye-fixation duration. *The American journal of psychology*, 93(2):207—234.

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

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. Association for Computational Linguistics.

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

Claude E. 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.

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

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

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

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. In *Forty-Second International Conference on Machine Learning*.

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

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*, pages 15711–15721. Association for Computational Linguistics.

Yizhou Zhang, Defu Cao, and Yan Liu. 2022. Counterfactual neural temporal point process for estimating causal influence of misinformation on social media. In *Advances in Neural Information Processing Systems*, volume 35.

Carolina M. Zingale and Eileen Kowler. 1987. Planning sequences of saccades. *Vision Research*, 27(8):1327–1341.

	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.

A Experimental Setup

Data. We use the English portion of the MECO dataset (Siegelman et al., 2022) as a source of reading data. The MECO dataset provides scanpaths for multiple readers over 12 short text excerpts from Wikipedia; the English portion of the dataset includes scanpaths for 46 readers. The texts were displayed on a 1920×1080 screen in monospaced Consolas 22pt font. As summarized in Tab. 1, the number of characters per text ranges from 831 to 1230 (arithmetic mean 1093), and the number of lines from 8 to 12 (arithmetic mean 10.5). We use the Tesseract OCR library[10] in Python to identify characters and their bounding boxes; see the paragraph below for more details. For each session, we consolidated the gaze measurements into a scanpath sequence $\mathcal{T}$ for each reader. This process resulted in 97,742 fixations across texts and readers. Each fixation is then associated with the character whose bounding box contains its location coordinates or flagged as having landed outside the bounding box of any character. To facilitate comparison with coarser targets, we derived three additional datasets: (i) the filtered **scanpath-duration dataset** (§2), obtained by removing fixations outside word bounding boxes and merging consecutive fixations on the same word (46,511 data points); (ii) the **per-reader word-level gaze-duration dataset**, in which each record captures the gaze duration of a single reader on a single word (34,368 data points); and (iii) the **averaged word-level gaze-duration dataset**, produced by averaging gaze durations across readers for each word (2,097 data points). Gaze duration is defined in §2.1.

Optical character recognition (OCR). We applied the Python Tesseract OCR library[11] 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 90$^{\text{th}}$ 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.

Model training and evaluation. Our models were implemented in PyTorch and trained with gradient-based optimization using stochastic gradient descent (Robbins and Monro, 1951), with Nesterov momentum (Nesterov, 1983). The models were trained on a fixed split of 80% training, 10% validation, and 10% test data over 30 epochs, using early stopping with a patience of 5 epochs (i.e., training stopped if the validation loss did not improve for 5 consecutive epochs). The best hyperparameter configuration was selected based on validation performance and then evaluated on the held-out test set. For the models related to fixation onsets and locations, we performed a grid search over batch sizes $\{64, 128, 256\}$, learning rates $\{0.1, 0.01, 0.001\}$, and weight-decay coefficients $\{0, 10^{-4}\}$, for a total of 18 runs for each model and dataset type. For the convolutional model of fixation durations, the batch size was fixed at 128, while learning rates $\{0.01, 0.001, 0.0001\}$ and weight-decay values $\{0, 10^{-4}\}$ were explored. We further varied the initial parameters of the Gamma distribution that defines the convolution kernel in Eq. (16), testing $(\alpha_k, \beta_k) \in \{(2, 3), (3, 4), (3, 6)\}$ and an initial θ_k of 0.5, for all predictors $k = \{1, \dots, K\}$. These

[10] https://pypi.org/project/pytesseract/
[11] https://pypi.org/project/pytesseract/

values were selected empirically so that the resulting distributions would have means that align with different inter-arrival times (from 0.5 to 0.75 seconds). Durations were recorded in milliseconds, but onset times were rescaled to seconds before being passed to the convolution and saccade models. This scaling keeps inter-arrival values within a range that avoids gradient explosion. We employed warm-starting to help in the search for a good set of parameter values. More specifically, we trained the models sequentially, starting with the simpler models used as baselines and using those for initialization in more complex models. For the more complex models, we initialized shared parameters with the best values from the corresponding simpler model. We also performed additional analyses using a linear model. For those analyses, we employed five-fold cross-validation: on each fold, the model was fit using 80% of the data and evaluated on the remaining 20%. Goodness of fit was quantified by the log-likelihood ratio (i.e., delta log-likelihood) with respect to predefined baselines, specified in §4.1 and §4.2. For experiments evaluated on the fixed test split, predictive uncertainty was estimated via bootstrap resampling of the test-set predictions.

Predictor values. Each fixation that lands in a character's bounding box can be adorned with linguistic attributes the character itself and the unit, e.g., the word, it belongs to. We consider the following attributes: (i) the character-level surprisal given the context, (ii) the word-level surprisal of the corresponding word, (iii) the number of characters in the corresponding word, i.e., its length, and (iv) the unigram surprisal, i.e., the log frequency, of the corresponding word. For bounding boxes associated with whitespaces we can only compute character-level surprisal; fixations that do not land in any bounding box remain adorned with these attributes.[12] We distinguish character- and word-level surprisal as both have been considered as predictors in past work (Giulianelli et al., 2024a). To obtain the surprisal values, we must estimate $\overrightarrow{p}(\cdot \mid \boldsymbol{w}_{<t})$ from Eq. (3) using an autoregressive language model. Because most modern language models learn a distribution over *token* sequences, we use Vieira et al.'s (2024) algorithm to convert the language models to the character level. Word-level surprisal is obtained by summing the surprisal values of the subword tokens that comprise the word; this algorithm is correct when no subword token crosses a word boundary, which is the case in the language models we experiment with. Word-level surprisal estimates are derived from mGPT (Shliazhko et al., 2024), while character-level surprisal estimates are derived from GPT-2 (Radford et al., 2019). Finally, word lengths and word frequencies are obtained from Speer (2022).

B Discussion on Related Work

The below subsections give more context on aggregated eye-tracking measurements (App. B.1) and cognitive models of reading (App. B.2).

B.1 Common Aggregations of Reading Data and their Interpretation

We defined first fixation duration, gaze duration, and total fixation duration in §2.1. 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). Gaze duration (also called first-pass time), 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., 2024a). Scanpath durations, the summed duration of all *consecutive* fixations on a word, are different from the above aggregated measurements in that they may contain several measurements for a given word if the word was fixated

[12] An alternative approach would be to assign each fixation to the predictor values associated with either the previous or the closest bounding box; we do not explore these options.

on multiple times, excluding consecutive re-fixations. Thus, fixation events are ordered according to the temporal order in which the words were fixated on, rather than the sequential order in which they are written out in the text. They have been modeled, e.g., by Shain and Schuler (2021).

B.2 Cognitive Models of Reading

Several cognitive models of eye movement control during reading have been proposed. Among the most prominent are the E-Z Reader (Reichle et al., 2003) and SWIFT (Engbert et al., 2005), which are probabilistic models of eye movements that can also be framed as point processes. However, these models are not designed to capture the influence of linguistic processing on gaze behavior in a data-driven way. The Bayesian Reader (Norris, 2006), for instance, focuses on explaining effects such as the increased difficulty of processing low-frequency words. Beyond these, other approaches have used machine learning to predict which words are fixated during reading, i.e., skip rates (Nilsson and Nivre, 2009; Hahn and Keller, 2016; Wang et al., 2019; Bolliger et al., 2023). Some models also predict fixation durations directly (Bicknell and Levy, 2010; Bolliger et al., 2025). In this article, we model the scanpaths at a more granular level, considering the exact spatial locations of fixations rather than just the words they land on. Although our model can learn effects that are related to both cognitive and oculomotor control processes, it is not meant to provide a plausible explanation of the mechanisms underlying such effects.

C Intensity Functions for Baseline Models

In this section, we provide more details on the Poisson and last-fixation baselines. Similarly to how we characterized the Hawkes process through Eq. (7), we characterize each of these models through their intensity function. The density function from Eq. (13) is changed accordingly.

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

$$\lambda(t_n, \mathbf{s}_n; \mathcal{H}_{n-1}) \stackrel{\text{def}}{=} \nu, \tag{21}$$

where $\nu \in \mathbb{R}_{\geq 0}$ is the only learnable parameter. This serves as the simplest baseline, ignoring any spatial or temporal dependencies.

Last-fixation baseline. The last-fixation model introduces a basic spatial dependency by assuming that each new fixation is normally distributed around the most recent fixation. Let $\mathbf{s}_{n-1}$ denote the location of the last fixation. Then, the intensity function is given by

$$\lambda(t_n, \mathbf{s}_n; \mathcal{H}_{n-1}) \stackrel{\text{def}}{=} \nu + \psi_{n-1}\big(\mathbf{s}_n\big), \tag{22}$$

where $\psi_{n-1}(\mathbf{s}_n)$ is the probability density function of a normal distribution centered at $\mathbf{s}_{n-1}$ with variance σ^2. This corresponds to the spatial distribution in Eq. (10) when using $\mu_m^{\text{b}}(\mathbf{s}) \stackrel{\text{def}}{=} \mathbf{s}$, i.e., the identity function.

D Modeling Duration: Distribution Selection

In this section, we elaborate on our choice of the conditional density function for modeling fixation durations, as introduced in §3.2. We evaluated six candidate distributions commonly used for modeling right-skewed time-to-event data: Rayleigh, exponential, Weibull, normal, log-normal, and gamma. The evaluation was performed using 10-fold cross-validation on the training and validation data. While the log-normal and gamma distributions demonstrated comparable predictive performance (see Fig. 3), we selected the log-normal distribution as its parameters directly correspond to moments of the log-transformed durations, enabling more intuitive interpretation. Moreover, these parameters can also be interpreted as those of a normal distribution fitted to the log-durations, facilitating analysis through least squares estimators.

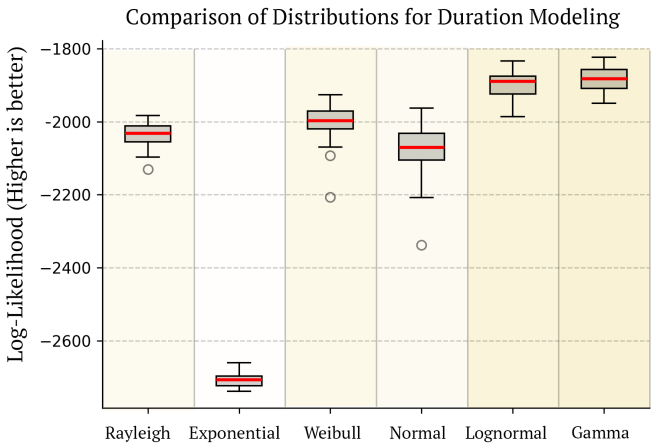

Figure 3: Goodness-of-fit comparison of candidate distributions for fixation durations. Both log-normal and gamma distributions showed superior performance compared to other alternatives; higher log-likelihood indicates better fit. The log-normal was ultimately selected for its enhanced parameter interpretability.

E Further Results

In this section, we provide additional plots and analyses to complement the main results: Fig. 4 expands upon Fig. 2, Fig. 5 shows outcomes from the convolution model of fixation durations, and Fig. 6 reports results for Markov linear models across various duration aggregation strategies.

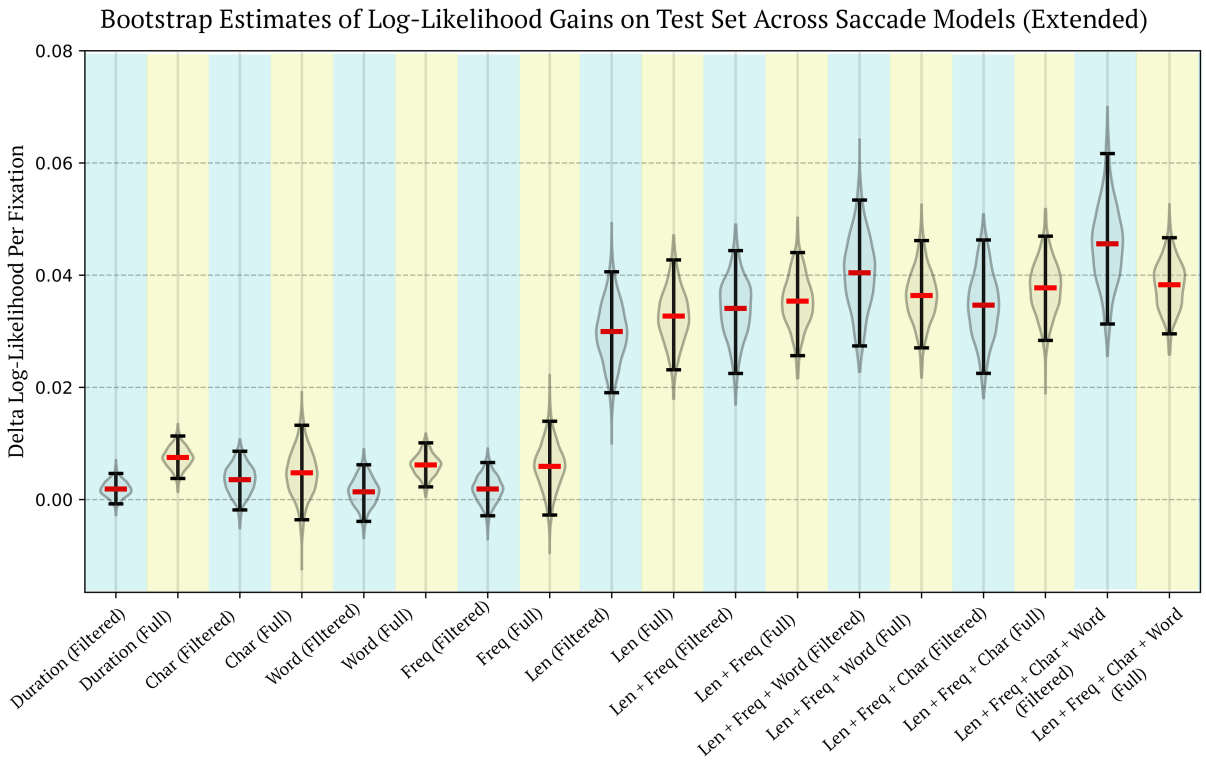

Figure 4: Bootstrapped distributions of per-fixation log-likelihood gains across saccade models computed relative to the RSE model. The models correspond to those specified in §4.1. The *"Raw"* and *"Filtered"* labels indicate whether the model was trained and evaluated on the raw or filtered scanpath dataset, respectively. We write "word" for word-level surprisal, "char" for character-level surprisal, and "freq" for unigram surprisal. Higher values indicate better predictive performance in modeling saccade durations. A shortened version of the plot can be found in Fig. 2.

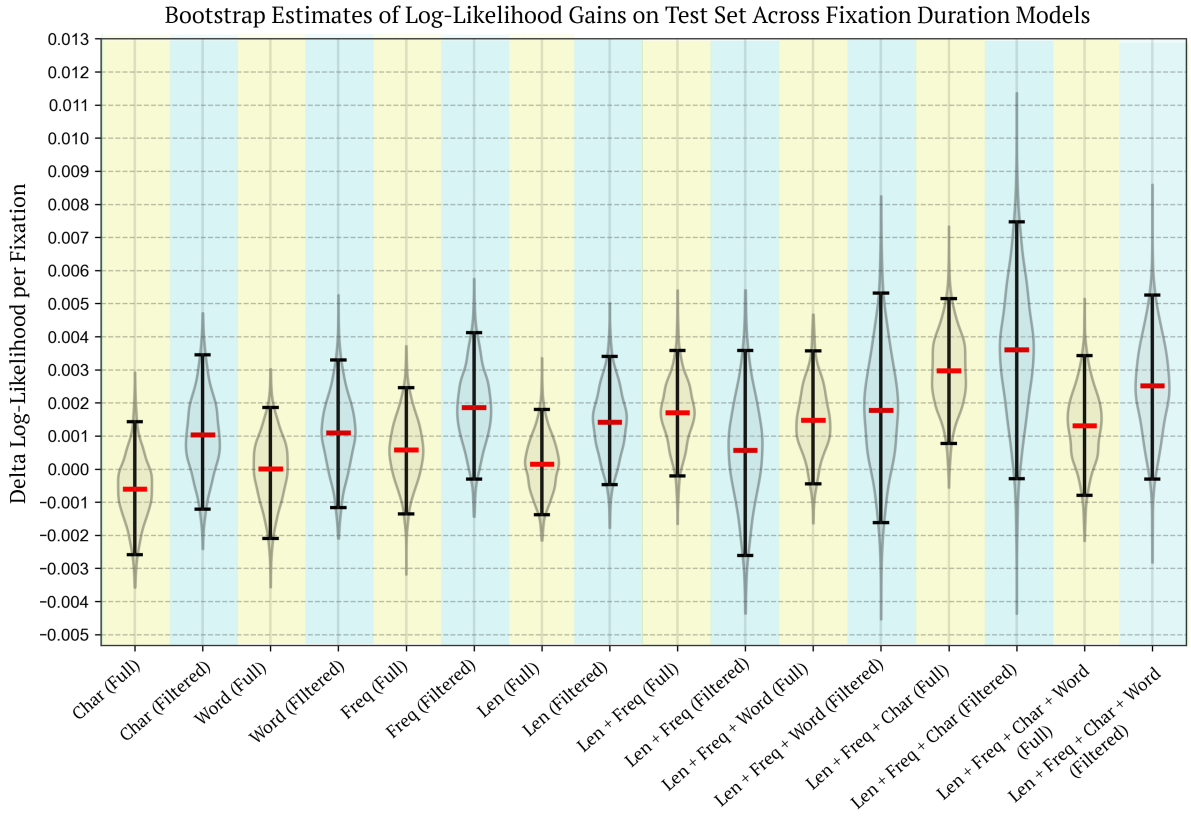

Figure 5: Bootstrapped distributions of per-fixation log-likelihood gains across duration models, as defined in Eq. (16). The base model includes reader-specific intercepts and duration spillovers via a convolution term. The full models incorporate the set of predictors indicated on the x-axis. For each predictor, the model includes its value at the current fixation, its interaction with the reader, and its spillovers through the convolution term. The models correspond to those described in §4.1. The labels *"Raw"* and *"Filtered"* indicate whether the model was trained and evaluated on the raw or filtered scanpath dataset, respectively. We write "word" for word-level surprisal, "char" for character-level surprisal, and "freq" for unigram surprisal. For models trained on the raw dataset, an additional binary predictor indicates whether the fixation occurs on a word. Higher values reflect better predictive performance in modeling fixation durations.

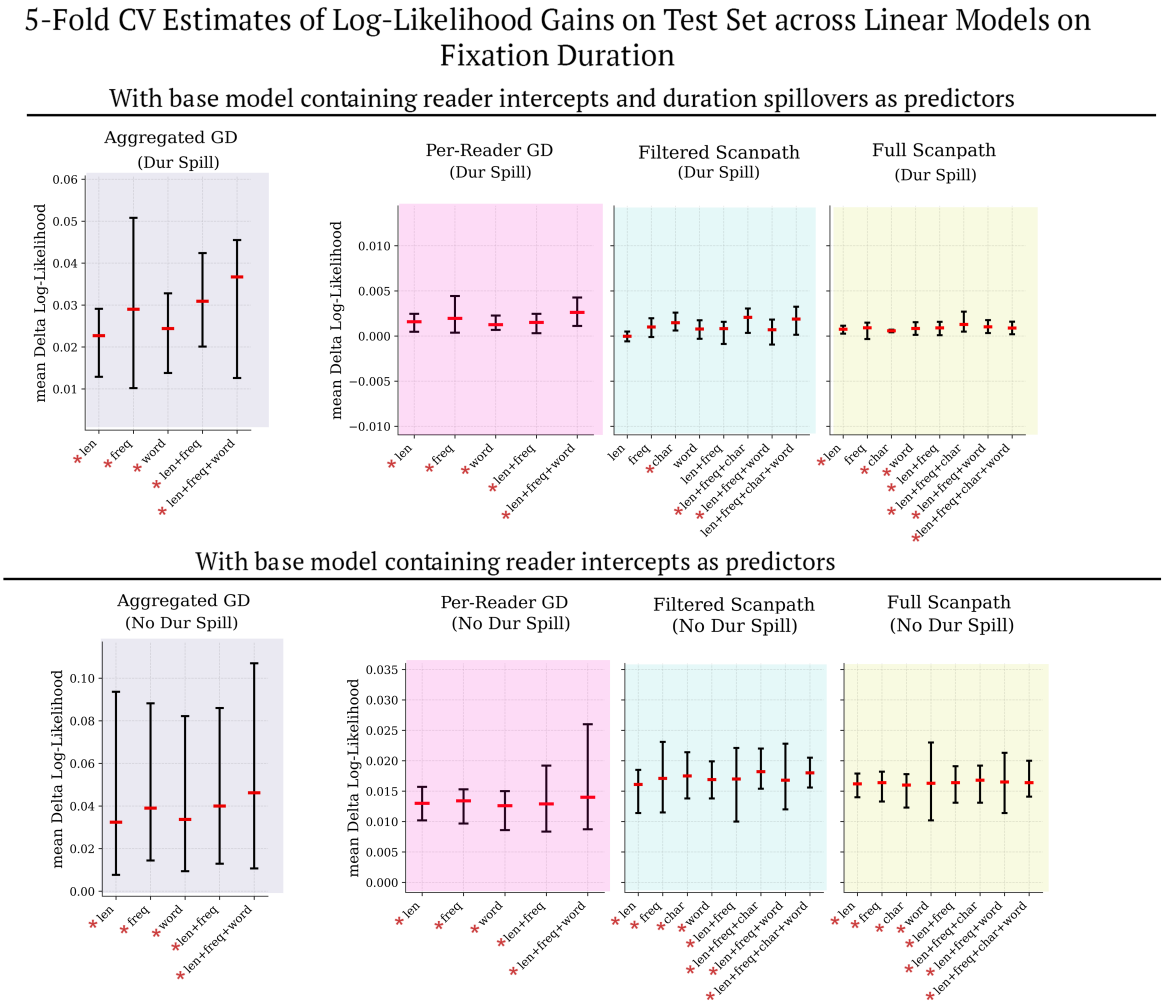

Figure 6: Bootstrapped distributions of per-fixation log-likelihood gains across linear Markov models of fixation duration, as defined in Eq. (18). Each delta log-likelihood is computed relative to a baseline model that includes: (i) reader-specific intercepts and spillover effects from previous fixation durations (models in the top panel of the figure), and (ii) only reader-specific intercepts (models in the bottom panel). The models with spillovers include them from the previous two fixations. The x-axis labels specify the set of predictors included in each model variant. Each full model incorporates the specified predictors with corresponding spillovers, as well as the predictors in the corresponding baseline model. We write "word" for word-level surprisal, "char" for character-level surprisal, and "freq" for unigram surprisal. Gaze duration is abbreviated as "GD". Each plot corresponds to one of the datasets introduced in App. A. Error bars represent the range—from minimum to maximum—observed across the five folds, with the central point indicating the arithmetic mean. We use asterisks (*) on the x-axis to highlight models that achieved a positive delta log-likelihood in all data folds.