Towards A Scanpath-Conditioned Surprisal Theory: Modeling Reader Information States

Michael Mooney, Edmond S. L. Ho
School of Computing Science
University of Glasgow
m.mooney.2@research.gla.ac.uk, Shu-Lim.Ho@glasgow.ac.uk

Abstract

Standard surprisal is typically computed from the linear text prefix, but human reading is non-linear and memory-constrained: readers skip words, regress, and do not retain prior context perfectly. We propose a formulation of surprisal conditioned on a reader-specific accessible information state given by the scanpath history and memory dynamics, rather than by the written prefix alone. Prior context is treated as only probabilistically accessible at each fixation, allowing predictability to depend on both non-linear exposure and forgetting. We evaluate the approach on eye-tracking corpora using held-out log-likelihood over standard duration-based reading measures. Across model variants, conditioning on accessible information states improves predictive fit over standard surprisal baselines. These results suggest that predictability in human reading is better characterized relative to the reader's evolving accessible information state than to the written prefix alone.

1 Introduction

How hard is a word to read? Research suggests it depends on predictability: unexpected words take longer to process than predictable ones (Rayner, 1998). *Surprisal* formalizes this intuition, a word w_i that is less probable given prior context has higher surprisal, $s(w_i) = -\log p(w_i \mid \mathcal{C}_{<i})$, and should incur greater processing cost. When estimated with language models, surprisal correlates strongly with reading times (Hale, 2001; Levy, 2008; Smith and Levy, 2013).

In empirical work, surprisal is computed by conditioning on the linear text prefix $w_{<i}$, corresponding to a reader with perfect access to all prior material. Eye-movement data reveals this assumption is misaligned with actual behavior: readers skip words, regress, and partially forget previously fixated material (Rayner, 1998; Engbert et al., 2005), while access to earlier context degrades with distance and interference (Lewis et al., 2006; McElree, 2000).

This misalignment is expected to be most severe for regressions which is a decision to return to prior material, driven by and simultaneously altering the reader's predictive state. Standard surprisal cannot represent this dynamic as it assigns the same predictability to a word whether encountered for the first time or revisited following comprehension difficulty. The conditioning context should instead reflect the reader's internal belief state. Whilst this state is latent, the scanpath is its observable trace.

Timkey et al. (2025) find a dissociation between surprisal and reading behavior where it explains forward reading but not regressions at disambiguation points. Building on this, we propose that part of this gap reflects how the conditioning context is defined rather than the limits of surprisal theory itself. We introduce *scanpath-conditioned surprisal*, which computes word predictability relative to a stochastic approximation of the reader's information state derived from the scanpath, incorporating probabilistic memory degradation over prior fixations. This preserves the standard interpretation of surprisal as negative log-likelihood, differing only in the conditioning context. Evaluated across two eye-tracking corpora and four reading regimes, scanpath-conditioned surprisal consistently improves predictive fit over standard LM surprisal, suggesting that the conditioning context, not surprisal theory itself, underlies its limited account of non-linear reading.

2 Background

Assuming that the processing cost of a word is given by its reading duration, then significant amount of results have reported a relationship between the surprise of a word and the words reading duration across various corpora and measures

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 41366–41380

(Smith and Levy, 2013; Wilcox et al., 2023; Shain et al., 2024). The exact computation of the distribution p of the surprise of a word is typically estimated using transformer based language models (LMs) (Oh and Schuler, 2025). However, as neural LMs improve in next-word prediction, their surprisal estimates often become less aligned with human reading times (Oh and Schuler, 2023; Oh and Linzen, 2025). One such reason is due to the unhuman-like memory that modern LMs poses, with studies showing near full recital of books (Ahmed et al., 2026).

This difference can be seen in repeated reading where Vaidya et al. (2023) found in a repeated reading cloze task that human reading speed and accuracy improve only modestly, language models do not. Including both readings in the same context effectively memorizes the first, unlike human memory. However, excluding it entirely assumes no carry-over between readings, which is equally wrong (Hyönä and Niemi, 1990; Raney and Rayner, 1995). Klein et al. (2024) further showed that when computed independently, surprisal becomes a weaker predictor during repeated reading than first reading.

Human memory for prior linguistic context is neither complete nor uniform. Memory-based accounts posit that earlier material is accessed through a cue-dependent retrieval process subject to decay and similarity-based interference, rather than maintained as a faithful record of the input (Lewis and Vasishth, 2005; McElree, 2000). Temporal distance reduces accessibility as representations held in working memory become harder to retrieve over time (McElree, 2000), while intervening material introduces similarity-based interference that further degrades retrieval accuracy (Lewis et al., 2006; Van Dyke and Lewis, 2003; Jäger et al., 2017).

Futrell et al. (2020) formalities this inside a surprisal framework, defining processing difficulty as surprisal given not the true preceding context but a *lossy* memory representation of it. Because that representation is incomplete, the comprehender reconstructs the probable context by combining noisy memory evidence with language-specific priors, naturally producing locality and structural forgetting effects within a single framework. The noise distribution remains a free parameter, however, and the model itself has no way of accounting for the individual reader or how they actually moved through the text.

A related limitation is during the processing of syntactically ambiguous structures, such as garden-path sentences. Readers naturally commit to an initial grammatical interpretation, but are forced to reanalyze the sentence when later words conflict with that assumption (Fodor and Ferreira, 1998). Recent work has shown that standard neural-LM surprisal substantially underestimates the magnitude of human garden-path slowdowns (Huang et al., 2024; Staub, 2025). However, this failure is not uniform across all reading behavior metrics. Timkey et al. (2025) showed a dissociation in which surprisal tracks forward reading relatively well, even in syntactically challenging sentences, but fails to explain regression-contingent rereading and structural reanalysis costs.

A further complication is that the accessible history is shaped not only by memory, but also by selective exposure. Standard surprisal assumes that every word in the written prefix contributes to the conditioning context. However, reading behavior is highly task-sensitive, and readers may skip substantial portions of text depending on their goals (Hahn and Keller, 2023; Shubi and Berzak, 2023; Shubi et al., 2025). This implies that the effective conditioning context for human prediction may be sparse, non-linear, and reader-specific (Rayner, 1998).

Whilst recent work has improved surprisal by redefining the outcome space (Giulianelli et al., 2023; Meister et al., 2024), we instead redefine the conditioning space. Predictability should be computed relative to what the reader has actually been exposed to and can access. For example, the occurrence of a regression fundamentally alters the global context, the reader gains access to downstream lexical evidence that was previously unavailable. Because standard surprisal is strictly constrained to the historical prefix, it cannot account for how this forward context updates the readers information state, even after they resume forward reading.

3 Scanpath Surprisal

We introduce scanpath-conditioned surprisal, defined over the oculomotor trace rather than the text sequence. Let $\mathbf{w} = (w_1, \ldots, w_N)$ be a word sequence. A scanpath $\mathbf{f} = (f_1, \ldots, f_T)$ is a temporally ordered sequence of fixations, each mapped to a word index $i_t \in \{1, \ldots, N\}$ where we denote $x_t = w_{i_t}$ for the fixated word at time t. The fixation history $\mathcal{H}_{t-1} = (x_1, \ldots, x_{t-1})$ is tempo-

rally ordered and may contain repeated tokens. Let $\mathcal{C}_{t-1} = \sigma(\mathcal{H}_{t-1})$ denote the scanpath-derived context, where σ is a model-dependent transformation of this history. Scanpath-conditioned surprisal at fixation t is then:

$$s(x_t) = -\log p(x_t \mid \mathcal{C}_{t-1})$$

The precise computation of $p(x_t \mid \mathcal{C}_{t-1})$ depends on the model instantiation, defined in the following sections.

3.1 Deterministic Information States

We first define two deterministic models that condition word predictability on the lexical information a reader may have actually encountered. Consider a reader that may skip from w_1 to w_5 and subsequently regress to w_2, then the information state at w_2 has fundamentally changed as it now incorporates downstream lexical content. To capture this, we introduce $M_{0;\text{dense}}$, which assumes that at time t, the reader has access to the entire preceding context up to the furthest forward fixated word reached at time t. Specifically, let $j_{t-1}^{\max} = \max_{1 \le k < t-1} i_k$ then the context is given by, $\mathcal{C}_{t-1}^{\text{dense}} = (w_1, \ldots, w_{j_{t-1}^{\max}})$ and thus, $s_{0;\text{dense}}(x_t) = s(x_t \mid \mathcal{C}_{t-1}^{\text{dense}})$.

The model $M_{0;\text{dense}}$ is sensitive to noisy distal fixations occurring deep into the stimuli, for example, w_1 to w_{N-1}, would result in the majority of $\mathbf{w}$ being conditioned. To evaluate a more conservative assumption, we formulate $M_{0;\text{sparse}}$, which restricts the information state to the tokens that have been fixated prior to time t. This results in a sparse information state consisting of the scanpath history, thus, $s_{0;\text{sparse}}(x_t) = s(x_t \mid \mathcal{H}_{t-1})$. This formulation is conceptually aligned with the task-based skipping framework of Hahn and Keller (2023) as well as the conditioning as given by Bicknell and Levy (2009).

Both $M_{0;\text{sparse}}$ and $M_{0;\text{dense}}$ assume perfect retention of the conditioning context. Although regressions are typically shorter in duration than forward fixations (Inhoff et al., 2019), suggesting surprisal should be lower but not negligible, perfect retention means a modern LM may assign near-zero surprisal to a regressed word it has already seen.

3.2 Stochastic Information States

The deterministic models and standard surprisal share a common assumption: perfect access to whatever context they condition on. From an information theoretic perspective, surprisal should be defined relative to the readers information state, what evidence is currently accessible at fixation t, not what evidence was ever observed. We therefore treat context as a random variable induced by memory limitations, and define surprisal by marginalizing over this uncertainty. Formally, let $Z^{(t)} = (Z_1^{(t)}, \ldots, Z_{t-1}^{(t)}) \in \{0,1\}^{t-1}$ denote a stochastic accessibility mask over the scanpath history H_{t-1}, where $Z_k^{(t)} = 1$ indicates that the lexical content from fixation k is accessible at time t. Given $Z^{(t)}$, we define the accessible context as the temporal subsequence

$$\tilde{H}_{t-1}(Z^{(t)}) = (x_k : \ 1 \le k < t, \ Z_k^{(t)} = 1)$$

which preserves scanpath order and permits duplicates. Let $\mathcal{Q}_{t-1}$ denote the distribution over $Z^{(t)}$ induced by a particular memory model. We define the scanpath conditioned probability of the currently fixated word by marginalizing over the uncertain accessible context. Specifically,

$$P_t(x_t) \ = \ \mathbb{E}_{\tilde{H} \sim \mathcal{Q}_{t-1}} \big[p(x_t \mid \tilde{H}) \big]$$

The corresponding scanpath surprisal is then given by,

$$S(t) = -\log \mathbb{E}_{\tilde{H} \sim \mathcal{Q}_{t-1}} \big[p(x_t \mid \tilde{H}) \big]$$

This formulation generalizes deterministic scanpath-conditioned surprisal as when Q_{t-1} is a point estimate at H^*, the expectation collapses to $p(x_t \mid H^*)$, giving, $-\log p(x_t \mid H^*)$. The stochastic formulation extends surprisal by marginalizing over uncertainty in the accessible history rather than committing to a point estimate.

Our formulation differs from lossy-context surprisal (Futrell et al., 2020), which defines processing difficulty as expected surprisal over uncertain contexts. We instead take the log of the expectation, corresponding to a reader whose predictions are formed by integrating over possible contexts rather than committing to one. By Jensen's inequality, our quantity is a lower bound. Under this view, the scanpath is observable evidence about a latent memory-constrained information state, and surprisal remains negative log predictive probability. We compare both objectives alongside a point estimate baseline in Appendix B.

As the expectation over contexts is latent, we estimate $P_t(x_t)$ by Monte Carlo sampling. Drawing M accessible histories $\tilde{H}^{(1)}, \ldots, \tilde{H}^{(M)} \sim \mathcal{Q}_{t-1}$,

giving,

$$\widehat{P}_t(x_t) := \frac{1}{M}\sum_{m=1}^{M} p_{\mathrm{LM}}(x_t \mid \tilde{H}^{(m)}),$$
$$\widehat{S}(t) := -\log \widehat{P}_t(x_t).$$

The next sections provide specific instantiations corresponding to different choices of $\mathcal{Q}_{t-1}$.

3.2.1 M_1: Kernelized Forgetting

We introduce our first main model, M_1. We first instantiate the scanpath information state $\mathcal{Q}_{t-1}$ with a forgetting mechanism. As reading is constrained by working memory (Miyake et al., 1994), evidence from earlier fixations becomes progressively less accessible with temporal distance (McElree, 2000; Bartek et al., 2011; Lewis et al., 2006). We capture this by treating accessibility as stochastic, each prior fixation event is either accessible or not at time t, with probability decaying as a function of lag. We parameterize $\mathcal{Q}_{t-1}$ via an independent Bernoulli accessibility mask over fixation events:

$$Z_k^{(t)} \sim \mathrm{Bernoulli}(\pi_k^{(t)}), \quad \pi_k^{(t)} := \exp\left(-\frac{t-k}{\tau}\right)$$

where $\tau > 0$ is a forgetting timescale. The accessible history $\tilde{H}_{t-1}$ is the temporal-order subsequence of H_{t-1} selected by $Z^{(t)}$ (duplicates preserved), giving $\tilde{H}_{t-1} \sim \mathcal{Q}_{t-1}$. Given the marginal surprisal definition from the previous section, we estimate $S(t)$ by Monte Carlo sampling. When $\mathcal{Q}_{t-1}$ is degenerative at a single history and set $\tau \to \infty$ we recover $M_{0;\text{sparse}}$.

3.2.2 M_2: Encoding × Retrieval

While M_1 makes accessibility a pure function of recency alone, retrieval-based accounts predict that availability depends on both temporal distance and encoding strength. Repeated attention strengthens encoding, increasing the likelihood that an item remains accessible (Van Dyke and Lewis, 2003; Jäger et al., 2017). We therefore drive accessibility by scanpath-dependent activation giving rise to model M_2. We define an activation trace over text positions $j \in \{1, \ldots, N\}$,

$$A_j^{(t)} = \sum_{k<t} \mathbf{1}[i_k = j]\,(t-k)^{-d},$$

where $d > 0$ controls the rate of decay. Activation is mapped to an accessibility probability via a sigmoid,

$$\alpha_j^{(t)} = \sigma\left(\frac{A_j^{(t)}}{\eta}\right),$$

with scale $\eta > 0$. Each prior fixation event k is then accessible with probability determined by the current accessibility of its text position,

$$Z_k^{(t)} \sim \mathrm{Bernoulli}\left(\alpha_{i_k}^{(t)}\right),$$

inducing $\mathcal{Q}_{t-1}$ over accessible histories. As in M_1, we estimate $S(t)$ by Monte Carlo sampling.

3.3 Distributional Uncertainty

Surprisal quantifies how unexpected the fixated word is under the scanpath-conditioned predictive distribution. We further extend M_1 and M_2 with two information-theoretic quantities, giving M_1^+ and M_2^+, bringing established KL and entropy-based accounts of processing difficulty (Levy et al., 2009; Hale, 2006) into the scanpath-conditioned framework. For $M \in \{M_1, M_2\}$ representing each model, then the scanpath-conditioned predictive distribution over the next word y is,

$$P_t^{(M)}(y) = \mathbb{E}_{\tilde{H}\sim\mathcal{Q}_{t-1}^{(M)}}\left[p(y \mid \tilde{H})\right].$$

After observing x_t, we form the updated predictive distribution $P_{t+1}^{(M)}(\cdot)$ by extending the history with x_t and applying the same accessibility mechanism. We then define the belief update magnitude as

$$BU_M(t) = D_{\mathrm{KL}}\left(P_t^{(M)} \,\|\, P_{t+1}^{(M)}\right),$$

and standard information gain (entropy change) (Cover and Thomas, 2005) as

$$IG_M(t) = H\left(P_t^{(M)}\right) - H\left(P_{t+1}^{(M)}\right).$$

where $IG_M(t)$ may be negative, corresponding to increased predictive uncertainty after observing x_t. We estimate $P_t^{(M)}(\cdot)$ and $P_{t+1}^{(M)}(\cdot)$ by Monte Carlo averaging over sampled accessible histories.

4 Dataset

We evaluate on two English eye-tracking corpora: OneStop (Berzak et al., 2025), comprising 360 native speakers reading 30 Guardian articles across four conditions crossing the reading goal (ordinary vs. information seeking) with prior exposure (first vs. repeated reading) and CELER (Berzak et al., 2022), comprising 365 speakers reading Wall Street Journal sentences, providing a sentence-bounded setting aligned with the classical surprisal paradigm. We follow the preprocessing pipeline of Klein et al. (2024), excluding skipped words, punctuation-containing words, paragraph boundary words, fixations exceeding 3,000 ms, and words whose base LM surprisal exceeds 20 bits.

5 Methods

Standard surprisal. We compute standard surprisal using autoregressive language models, where the word log-probability follows from the chain rule over subword tokens. We compute this using `wordsprobability` (Pimentel and Meister, 2024) for Pythia-70m. Our selection of Pythia-70m follows Oh and Schuler (2023), where larger LMs are worse predictors of human behavioral data.

Scanpath-conditioned scoring. We compute scanpath-conditioned surprisal with both autoregressive models and masked language models. For Pythia-70m, each sampled scanpath history is scored with the standard autoregressive word probability using the previously fixated words in temporal order. For $M_{0;\text{dense}}$, the context is given by $(w_1, \ldots, w_i, \ldots, w_j)$ where w_i is the target word. To avoid target leakage, we remove w_i from the context and compute $p(w_i \mid w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_j)$. Word probabilities are computed over the full tokenisation of the word following Pimentel and Meister (2024).

We note that autoregressive language models are sensitive to out-of-order token sequences (Ford et al., 2018). We therefore treat Pythia-70m as an auxiliary comparison and use MLMs as the main implementation, since MLMs are more robust to degraded and permuted word order (Sinha et al., 2021), while autoregressive model quality can vary substantially. Further, MLMs are naturally adapted to scoring for $M_{0;\text{dense}}$ by simply masking the target word. For both autoregressive and MLM models, when computing BU and IG we truncate to the top-K tokens plus residual mass.

MLM baselines via pseudo-log-likelihood. As MLMs do not define a joint probability over sequences via the chain rule, we cannot compute standard surprisal. Instead, we score following Salazar et al. (2020) and compute pseudo-log-likelihood (PLL), which approximate word log-probability for masked language models. To maintain consistency with the scanpath-conditioned models, in which future words are absent from the context entirely, we compute PLL using PLL-SENTENCE-L2R variant by Kauf and Ivanova (2023), thus leaving only the left context visible. We compute both PLL and scanpath-conditioning using ModernBERT (Warner et al., 2024).

Repeated Reading. For repeated reading, conditioning directly on the complete first reading can make conventional LM surprisal artificially small as the model has access to the earlier presentation of the same text. Our scanpath-conditioned models instead treat the first reading as part of the reader's memory state, subject to stochastic retention. At fixation t in the repeated reading, the context is constructed as

$$\big(x_1^{(1)}, \ldots, x_n^{(1)}, \underbrace{\texttt{[GAP]}, \ldots, \texttt{[GAP]}}_{g}, x_1^{(2)}, \ldots, x_{t-1}^{(2)}\big)$$

where $x^{(1)}$ denotes first-reading fixations, $x^{(2)}$ denotes prior fixations in the repeated reading, and g is the number of intervening fixations represented as phantom temporal events. This preserves the temporal separation between readings without treating the intervening period as lexical input. When the full gap would exceed ModernBERT's context window, we cap only the phantom sequence so that both readings remain available which affects less than 1% of paragraph pairs. We report two main variants, `[GAP]`, implemented with a reserved unused ModernBERT vocabulary item, and `[PAD]`.

6 Evaluation

We evaluate predictive power across four eye-tracking measures spanning different stages of processing: first fixation duration (FFD), gaze duration (GD), forward reading time (FRT), pre-regression gaze duration (RGD), and regressive go-past time (RGo-Past). FFD and GD index early and aggregate first-pass processing cost respectively. FRT and RGD partition GD by exit direction, where RGD captures the processing load that triggers a regression. RGo-Past extends RGD to include subsequent rereading time, capturing the full cost of reanalysis. Full definitions are given in Appendix A.

Scanpath-conditioned surprisal is defined at the fixation level, hence a word may receive multiple fixations and multiple surprisal values each reflecting distinct information states. We therefore sample each predictor at the fixation that naturally anchors each measure: FFD as the first fixation trivially, and all other measures at the exit fixation, the point at which first-pass processing has concluded and the reader commits to leaving the word. Averaging across fixations would be inappropriate as each surprisal value is computed under a different accessible history, reflecting a qualitatively distinct point in the reader's evolving belief state rather than a repeated estimate of the same quantity.

We estimate the relationship between predictors and reading times using Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1986). Following Klein et al. (2024) and Wilcox et al. (2023), we include word frequency $f_i = -\log p(w_i)$ (Speer et al., 2018) and character length l_i as base controls, modeling their interaction via tensor-product smooths $te(\cdot, \cdot)$. To account for spillover effects (Rayner, 1998), we include the corresponding previous predictors where for standard surprisal it is given by w_{i-1} and for scanpath-conditioned at time $t-1$. Consistent with prior work showing the surprisal to RT relationship to be linear (Klein et al., 2024; Wilcox et al., 2023; Smith and Levy, 2008; Shain et al., 2024; Smith and Levy, 2013), surprisal and scanpath predictors enter the model as linear terms. We ablate for non-linear terms in Appendix D.

Models. To isolate the unique predictive contribution of scanpath-conditioned surprisal beyond standard LM surprisal, we fit GAMs[1] under a nested design. A baseline of control variables is augmented first with standard LM surprisal (m_{surp}), then additionally with scanpath-conditioned surprisal (m_{combo}):

$$\begin{aligned} m_{\text{base}}: &\quad RT_i \sim te(f_i, l_i) + te(f_{i-1}, l_{i-1}) \\ m_{\text{surp}}: &\quad RT_i \sim m_{\text{base}} + s_i^{\text{LM}} + s_{i-1}^{\text{LM}} \\ m_{\text{combo}}: &\quad RT_i \sim m_{\text{surp}} + s_i^{M} + s_{\text{prev}}^{M} \end{aligned}$$

All reported $\Delta\mathcal{L}$ is the improvement of m_{combo} over m_{surp}, measuring how much variance scanpath-conditioned surprisal explains over standard LM surprisal. For M^+, m_{combo} additionally includes linear terms for information gain IG_i and belief update magnitude BU_i.

Predictive Power. Following Goodkind and Bicknell (2018) and Wilcox et al. (2020), we measure predictive power using delta log-likelihood:

$$\Delta\mathcal{L} = \log \mathcal{L}^{\text{combo}}(RT_i) - \log \mathcal{L}^{\text{surp}}(RT_i)$$

where,

$$\mathcal{L}^{M}(RT_i) = f_{\text{norm}}\Big(RT_i \mid \mu = \widehat{RT}_i^{M},\ \sigma^2 = \sigma_\varepsilon^2\Big)$$

is the Gaussian likelihood of observation i under model M, with fitted mean $\widehat{RT}_i^{M}$ and residual variance σ_ε^2. We report $\Delta\mathcal{L}$, the per-word mean of $\Delta\mathcal{L}_i$, in units of $100 \times \Delta\mathcal{L}$, measured on held-out data using 10-fold cross-validation. A positive $\Delta\mathcal{L}$ indicates improved predictive performance on unseen data. Significance is assessed via paired permutation tests (Wilcox et al., 2023). We control the false discovery rate across all reported tests within each model family using the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995).

Hyperparameters We select hyperparameters via grid search over $\tau \in \{5, 10, 20\}$ for M_1 and $d \in \{0.3, 0.5, 0.8\}$, $\eta \in \{0.5, 1.0, 2.0\}$ for M_2, evaluated on Ordinary Reading of OneStop with $M = 10$ and top-$K = 100$. We tuned on a single regime as τ, d and η are cognitive parameters describing the reader rather than the language model and should therefore generalize. The best configuration as measured by held-out $\Delta\mathcal{L}$ on gaze duration, was $\tau = 10$ for M_1 and $d = 0.8$, $\eta = 0.5$ for M_2. All reported results use these parameters with $M = 30$ and top-$K = 500$.

7 Results

We present all of our results using either Pythia-70m as the autoregressive model or ModernBERT as the MLM model over their respective basline scoring methods. All $\Delta\mathcal{L}$ values reflect incremental predictive gain over a fixed surprisal baseline using the same GAM structure throughout. Values are directly comparable across models within the same baseline (Pythia-70m or ModernBERT) but not across baselines. In addition significance is measured using the BH procedure. Lastly, for readability we report all values in the scale of $100 \times \Delta\mathcal{L}$.

7.1 Deterministic Models

Tables 1 and 2 report results for $M_{0;\text{dense}}$ and $M_{0;\text{sparse}}$. Both models give positive incremental $\Delta\mathcal{L}$ across nearly all metrics and corpora, confirming that scanpath structure carries information beyond surprisal alone. FFD has the smallest gains, consistent with prior work (Berzak and Levy, 2023), and RGo-Past effects are negligible across OneStop corpora. Repeated-reading regimes give smaller $\Delta\mathcal{L}$ than their first-pass counterparts, consistent with Klein et al. (2024) thus giving evidence that a fixed surprisal baseline does not fully account for prior exposure, leaving a memory-dependent component unmodelled. CELER diverges notably in RGD and RGo-Past, likely reflecting its sentence-level rather than paragraph-level structure.

[1] Following Klein et al. (2024) all models use `mgcv` with cubic regression splines (`bs="cr"`, `k=6`); no random effects are included due to convergence issues.

	M_0 Dense					M_0 Sparse				
Metric	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**
FFD	0.01^{**}	0.01^{**}	0.01^{*}	0.02^{**}	0.01^{**}	0.01^{**}	0.02^{**}	$\mathbf{0.06}^{**}$	0.01^{**}	0.03^{**}
GD	0.04^{**}	0.12^{**}	0.01	0.22^{**}	0.10^{**}	0.17^{**}	0.12^{**}	0.13^{**}	$\mathbf{0.23}^{**}$	0.18^{**}
FRT	0.03^{**}	0.04^{**}	-0.00	0.10^{**}	0.04^{**}	0.16^{**}	0.11^{**}	0.16^{**}	0.14^{**}	$\mathbf{0.17}^{**}$
RGD	-0.26^{**}	0.50^{**}	0.15^{**}	$\mathbf{0.55}^{**}$	0.49^{**}	-0.50^{**}	0.26^{**}	0.08	0.46^{**}	0.32^{**}
RGo-Past	0.52^{**}	-0.01	-0.04	-0.03^{**}	-0.00	$\mathbf{1.24}^{**}$	0.03^{**}	-0.06^{**}	0.06^{**}	0.14^{**}

Table 1: $100 \times \Delta\mathcal{L}$ for M_0 Dense and Sparse against a Pythia-70m surprisal baseline. Bold marks the highest value per metric row. p-values are Benjamini–Hochberg corrected across all tests. $^{*}p_{\text{BH}} < .05$, $^{**}p_{\text{BH}} < .01$, $^{***}p_{\text{BH}} < .001$.

	M_0 Dense					M_0 Sparse				
Metric	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**
FFD	0.00^{**}	0.09^{**}	$\mathbf{0.15}^{**}$	0.03^{**}	0.05^{**}	0.01^{**}	0.06^{**}	0.13^{**}	0.05^{**}	0.05^{**}
GD	0.06^{**}	0.24^{**}	0.22^{**}	0.35^{**}	0.24^{**}	0.50^{**}	0.39^{**}	0.27^{**}	$\mathbf{0.75}^{**}$	0.33^{**}
FRT	0.04^{**}	0.23^{**}	0.27^{**}	0.22^{**}	0.19^{**}	0.48^{**}	0.42^{**}	0.35^{**}	$\mathbf{0.78}^{**}$	0.36^{**}
RGD	$\mathbf{1.81}^{**}$	0.40^{**}	0.28^{**}	0.63^{**}	0.52^{**}	-0.65^{**}	0.27^{**}	0.08^{*}	0.52^{**}	0.28^{**}
RGo-Past	0.07^{**}	0.09^{**}	-0.00	0.10^{**}	$\mathbf{0.23}^{**}$	-1.45^{**}	0.03^{**}	-0.05^{**}	0.04^{**}	0.06^{**}

Table 2: $100 \times \Delta\mathcal{L}$ for M_0 Dense and Sparse against a ModernBERT PLL surprisal baseline. Bold marks the highest value per metric row. p-values are Benjamini–Hochberg corrected across all tests. $^{*}p_{\text{BH}} < .05$, $^{**}p_{\text{BH}} < .01$, $^{***}p_{\text{BH}} < .001$.

ModernBERT has larger $\Delta\mathcal{L}$ than Pythia-70m throughout. The non-linear, non-sequential structure of scanpath-conditioned context is less naturally accommodated by autoregressive chain-rule factorization, whereas bidirectional masking handles these dependencies more directly. The dense to sparse ordering further interacts with each specific model where $M_{0;\text{dense}}$ leads under Pythia, where its broader context outweighs $M_{0;\text{sparse}}$'s literal scanpath, but $M_{0;\text{sparse}}$ leads under ModernBERT, where $M_{0;\text{dense}}$ overlaps substantially with the PLL-style baseline.

7.2 Stochastic Models

Tables 3 and 4 report results for M_1 and M_2. Both models give larger $\Delta\mathcal{L}$ than the M_0 models, confirming that explicitly modeling memory accessibility carries incremental signal. Following M_0, ModernBERT has larger gains than Pythia throughout, unsurprisingly given both models build upon $M_{0;\text{sparse}}$. Notably, FFD and RGo-Past across OneStop remain almost identical to M_0 despite the additional memory structure, indicating that modeling reader state gives no incremental gain for these metrics, consistent with their informativeness being limited at the level of surprisal itself rather than the scanpath formulation.

M_2 has larger $\Delta\mathcal{L}$ than M_1 across most conditions, suggesting scanpath-dependent encoding strength provides signal beyond recency-based decay alone. M_1 treats accessibility as a function of temporal lag only, M_2 additionally weights by fixation frequency, meaning repeated attention directly increases retention probability.

Ordinary Reading consistently has larger gains than Information Seeking across both models and baselines, most clearly in GD and FRT under ModernBERT (e.g., M_2 GD: Ordinary $= 2.31$ vs. Information Seeking $= 1.46$). Repeated-reading regimes remain attenuated relative to their first-pass counterparts, though the attenuation narrows under M_1 and M_2 compared to M_0, suggesting explicit memory modeling partially but not fully accounts for prior exposure.

CELER diverges from Ordinary Reading across both models. GD and FRT are noticeably smaller, while RGD and RGo-Past show the most divergence. Higher positive values under Pythia such as under M_1 RGo-Past $= 0.78$ and with sign reversals under ModernBERT M_1 gives RGo-Past$= -1.75$. This divergence is consistent across M_0, M_1, and M_2, suggesting it reflects a structural property of CELER's sentence-level scanpaths rather than any particular model.

7.3 Repeated Reading

Table 5 reports carry-over results against the independent condition (Table 4). For Ord. Rep., including the first reading in the memory context has substantial increased effects over the independent condition on GD, FRT, and RGD, confirming that prior passage exposure is captured by the scanpath-conditioned memory state in a way that independent surprisal cannot. For Info. Rep., carry-over gains are more modest and closely approximate the independent condition, suggesting that goal-directed reading leaves little memory trace that the scanpath can exploit beyond what surprisal already

	M_1					M_2				
Metric	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**
FFD	0.02^{**}	0.03^{**}	$\mathbf{0.07}^{**}$	0.01^{**}	0.05^{**}	0.01^{**}	0.03^{**}	0.06^{**}	0.01^{**}	0.05^{**}
GD	0.19^{**}	0.30^{**}	0.15^{**}	0.23^{**}	$\mathbf{0.34}^{**}$	0.22^{**}	0.24^{**}	0.17^{**}	0.22^{**}	0.31^{**}
FRT	0.19^{**}	0.27^{**}	0.18^{**}	0.15^{**}	$\mathbf{0.34}^{**}$	0.21^{**}	0.23^{**}	0.22^{**}	0.16^{**}	0.31^{**}
RGD	-0.31^{**}	0.45^{**}	0.15	0.40^{**}	$\mathbf{0.58}^{**}$	0.38^{**}	0.37^{**}	0.13	0.37^{**}	0.45^{**}
RGo-Past	$\mathbf{0.72}^{**}$	0.07^{**}	-0.02	0.06^{**}	0.14^{**}	0.58^{**}	0.05^{**}	-0.04	0.03^{**}	0.09^{**}

Table 3: $100 \times \Delta\mathcal{L}$ for M_1 and M_2 against a Pythia-70m surprisal baseline. Bold marks the highest value per metric row. p-values are Benjamini-Hochberg corrected across all reported tests. $^{*}p_{\text{BH}} < .05$, $^{**}p_{\text{BH}} < .01$, $^{***}p_{\text{BH}} < .001$.

	M_1					M_2				
Metric	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**
FFD	0.02^{**}	0.07^{**}	0.11^{**}	0.06^{**}	0.11^{**}	0.02^{**}	0.08^{**}	$\mathbf{0.14}^{**}$	0.06^{**}	0.12^{**}
GD	0.56^{**}	0.71^{**}	0.46^{**}	1.16^{**}	0.69^{**}	0.73^{**}	1.46^{**}	0.67^{**}	$\mathbf{2.31}^{**}$	1.18^{**}
FRT	0.55^{**}	0.66^{**}	0.48^{**}	1.02^{**}	0.66^{**}	-1.07^{**}	1.51^{**}	0.74^{**}	$\mathbf{2.37}^{**}$	1.22^{**}
RGD	0.75^{**}	0.90^{**}	0.65^{**}	1.27^{**}	1.04^{**}	$\mathbf{2.65}^{**}$	1.16^{**}	0.65^{**}	1.61^{**}	1.11^{**}
RGo-Past	-1.75^{**}	0.09^{**}	-0.03	0.12^{**}	0.14^{**}	0.00	0.08^{**}	-0.04	0.16^{**}	$\mathbf{0.17}^{**}$

Table 4: $100 \times \Delta\mathcal{L}$ for M_1 and M_2 against a ModernBERT PLL surprisal baseline. Bold marks the highest value per metric row. p-values are Benjamini–Hochberg corrected across all reported tests. $^{*}p_{\text{BH}} < .05$, $^{**}p_{\text{BH}} < .01$, $^{***}p_{\text{BH}} < .001$.

	Gap context		**Pad context**	
Metric	M_1	M_2	M_1	M_2
Ord. Rep.				
FFD	0.08^{**}	0.07^{**}	$\mathbf{0.10}^{**}$	0.08^{**}
GD	0.82^{**}	$\mathbf{1.44}^{**}$	0.74^{**}	1.41^{**}
FRT	0.73^{**}	$\mathbf{1.35}^{**}$	0.69^{**}	1.29^{**}
RGD	1.20^{**}	$\mathbf{1.58}^{**}$	1.00^{**}	1.50^{**}
RGo-Past	0.13^{**}	0.01	$\mathbf{0.14}^{**}$	0.02
Info Rep.				
FFD	0.08^{**}	0.04^{**}	$\mathbf{0.09}^{**}$	0.04^{**}
GD	0.49^{**}	$\mathbf{0.69}^{**}$	0.48^{**}	0.65^{**}
FRT	0.46^{**}	$\mathbf{0.68}^{**}$	0.48^{**}	0.64^{**}
RGD	0.70^{**}	$\mathbf{0.72}^{**}$	0.60^{**}	0.56^{**}
RGo-Past	-0.02	$\mathbf{-0.02}$	-0.03	-0.02

Table 5: $100 \times \Delta\mathcal{L}$ for M_1 and M_2 under Gap and Pad carry-over contexts, evaluated against a ModernBERT PLL surprisal baseline. p-values are Benjamini-Hochberg corrected across all carry-over tests. $^{*}p_{\text{BH}} < .05$, $^{**}p_{\text{BH}} < .01$, $^{***}p_{\text{BH}} < .001$.

captures. [GAP] has larger gains than [PAD] for Ord. Rep., most likely reflecting that [PAD] is in-distribution for ModernBERT whereas [GAP] is a reserved unseen token, making it a cleaner separator. M_2 consistently gives larger $\Delta\mathcal{L}$ than M_1 on GD, FRT, and RGD for Ord. Rep., mirroring the main results. For Info. Rep. this advantage is largely absent and reverses on RGD under [PAD], suggesting that the processing load triggering regressions in goal-directed repeated reading is not well captured by fixation-frequency weighted retrieval, but rather by recency alone. RGo-Past shows no significant carry-over effect in either corpus, indicating that the full reanalysis cost on regression-exit trials is not modulated by prior passage exposure.

7.4 Distributional Uncertainty

Table 6 reports GD results for the predictive state dynamics ablation. Extending M_1 and M_2 with KL-based belief update and entropy-based information gain, following established accounts of predictive state in reading (Hale, 2006; Levy et al., 2009), both M_1^+ and M_2^+ have increased effects over their base counterparts. The contribution of IG relative to BU is substantially larger under M_2 than M_1 where for Ordinary Reading, M_1 $IG = 1.50$ against BU $= 2.95$, whereas M_2 $IG = 2.68$ against $BU = 3.05$. This suggests that entropy change is more informative under retrieval-based accessibility, and may partly account for the $M_2 > M_1$ advantage. M_2^+ has the largest $\Delta\mathcal{L}$ across all corpora, and the corpus ordering mirrors main results throughout.

8 Discussion

Across all models and baselines, conditioning on scanpath structure consistently improves predictive power over standard surprisal, addressing the limitation that text-prefix conditioning is a poor proxy for the context actually available during reading (Rayner, 1998; Hahn and Keller, 2023). A consistent dissociation emerges however in that FFD and RGo-Past show little to no incremental gain while GD, FRT, and RGD improve consistently. FFD reflects early lexical processing prior to full contextual integration, and the absence of predictive increase gives further indication that surprisal-based predictors have limited predictive power on early fixation behavior. RGo-Past is more theoreti-

	M_1			M_2		
Dataset	**BU**	**IG**	M_1^+	**BU**	**IG**	M_2^+
CELER	1.00^{***}	0.65^{***}	1.03^{***}	1.00^{***}	0.79^{***}	$\mathbf{1.03}^{***}$
Info Seek	1.96^{***}	0.97^{***}	2.13^{***}	2.02^{***}	1.76^{***}	$\mathbf{2.25}^{***}$
Info Rep.	0.98^{***}	0.49^{***}	1.01^{***}	0.92^{***}	0.75^{***}	$\mathbf{1.02}^{***}$
Ordinary	2.95^{***}	1.50^{***}	3.15^{***}	3.05^{***}	2.68^{***}	$\mathbf{3.29}^{***}$
Ord. Rep.	1.63^{***}	0.84^{***}	1.69^{***}	1.65^{***}	1.36^{***}	$\mathbf{1.78}^{***}$

Table 6: GD results for the belief update (BU), information gain (IG), and combined ablation for M_1 and M_2. $^{*}p_{\text{BH}} < .05$, $^{**}p_{\text{BH}} < .01$, $^{***}p_{\text{BH}} < .001$. Full ablation can be found in Appendix C

cally notable as our models improve RGD, capturing the predictive uncertainty that triggers a regression, yet fail to predict reanalysis integration once a regression has occurred. This was in part the motivation for scanpath-conditioned modeling. When a regression occurs the reader's information state changes and yet the improvement does not extend to RGo-Past, further cementing that reintegration costs lie beyond expectation-based accounts altogether (Timkey et al., 2025; Rabe et al., 2024). Our models thus capture *when* readers regress but not *how long* recovery takes, suggesting that repair and reintegration reflect processes that no surprisal formulation, scanpath-conditioned or otherwise, is likely to fully account for.

Explicitly modeling memory accessibility over the scanpath is central to our framework and the results show this as M_1 and M_2 consistently outperform M_0, confirming that naive context conditioning leaves substantial predictive signal unaccounted for. That retrieval-based M_2 outperforms recency-based M_1 suggests that fixation frequency, not just temporal distance, is informative for accessibility, consistent with retrieval-based accounts of working memory in reading (Lewis and Vasishth, 2005; Van Dyke and Lewis, 2003). The addition of BU and IG provides further incremental gains beyond the base memory models. Interestingly, IG contributes proportionally more under M_2 than M_1. This suggests that the advantage of M_2 over M_1 reflects not only better retention modeling but a more structured predictive state.

The repeated reading results further support the importance of memory modeling. Conditioning on the first pass within the same memory context substantially increases predictive power for Ord. Rep. across GD, FRT, and RGD, demonstrating that prior exposure leaves a memory trace which independent surprisal does not account for when computed using modern LMs (Klein et al., 2024).

CELER shows markedly more volatile behavior throughout, most likely reflecting its sentence-level stimuli producing far denser scanpaths relative to text length than in OneStop (Berzak et al., 2025, 2022). Specifically, with rereading frequently spanning a larger proportion of the entire stimuli and with fixed parameters over both CELER and OneStop, it therefore constrains what the memory models can exploit.

The interaction between model formulation and base LM architecture is a finding with direct practical implications. Under autoregressive models, $M_{0;\text{dense}}$ yields larger $\Delta\mathcal{L}$ than either M_1 or M_2 in Ordinary Reading, whereas under ModernBERT the memory models dominate throughout. The broader pattern of substantially larger gains against ModernBERT is consistent with the non-linear, non-sequential structure of scanpath-conditioned context being more naturally handled by bidirectional masked models than by autoregressive chain-rule factorisation. Together these results suggest that the choice of base LM architecture is not incidental to scanpath surprisal modelling but should be treated as a design decision in its own right.

9 Conclusion

We introduced a suite of scanpath-conditioned surprisal models that consistently improve over text-prefix surprisal across reading time measures and datasets. Explicitly modeling memory accessibility further improves over naive scanpath conditioning, and both base LM architecture and memory formulation matter for the magnitude of gains obtained. FFD and RGo-Past have very little increased effects throughout, suggesting that early lexical processing and reintegration costs lie beyond the reach of expectation-based accounts regardless of conditioning scheme.

10 Limitations

We report results for a single masked language model (ModernBERT) and a single autoregressive baseline (Pythia-70m). While these represent strong models at their respective scales, it remains to be established whether the gains from

scanpath conditioning generalize across model families, architectures, and scales. In particular, the relationship between model quality and the benefit of scanpath conditioning remains an open question, analogous to the scaling results of Oh and Schuler (2023).

Model parameters were optimized on OneStop and held fixed across all evaluations. While this provides a consistent comparison, it likely contributes to the volatile and in some cases collapsed behavior observed for CELER, whose sentence-level structure and denser scanpaths differ substantially from the paragraph-level regime the parameters were tuned for. Corpus-specific parameter optimization may recover more stable estimates for sentence-level datasets.

The BU/IG extension, while effective, is computationally expensive. Each fixation requires Monte Carlo sampling over accessible histories and computing full-vocabulary predictive distributions for KL divergence and entropy calculations, limiting practical scalability. A typical batch size requires tens of gigabytes of GPU memory, making the approach feasible for analysis but challenging for real-time applications or large-scale studies without substantial computational resources.

Finally, we do not control for low-level oculomotor factors (landing position, launch site, parafoveal preview benefit) beyond word length, which may partly explain weaker effects on FFD.

Acknowledgments

We gratefully acknowledge the Linguistic Data Consortium (LDC) for providing access to the CELER dataset by supporting Michael Mooney with their Data Scholarship. We thank our anonymous reviewers for their insightful feedback and helpful suggestions.

References

Ahmed M. Ahmed, A. Feder Cooper, Oluwasanmi Koyejo, and Percy Liang. 2026. Extracting books from production language models. *ArXiv*, abs/2601.02671.

Brian Bartek, Richard L. Lewis, Shravan Vasishth, and Mason R. Smith. 2011. In search of on-line locality effects in sentence comprehension. *Journal of experimental psychology. Learning, memory, and cognition*, 37 5:1178–98.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

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

Yevgeni Berzak, Jonathan Malmaud, Omer Shubi, Yoav Meiri, Ella Lion, and Roger Levy. 2025. Onestop: A 360-participant english eye tracking dataset with different reading regimes. *Scientific Data*, 12.

Yevgeni Berzak, Chie Nakamura, Amelia Smith, Emily Weng, Boris Katz, Suzanne Flynn, and Roger Levy. 2022. Celer: A 365-participant corpus of eye movements in l1 and l2 english reading. *Open Mind*, 6:41–50.

Klinton Bicknell and Roger Levy. 2009. A model of local coherence effects in human sentence processing as consequences of updates from bottom-up prior to posterior beliefs. In *North American Chapter of the Association for Computational Linguistics*.

Thomas M. Cover and Joy A. Thomas. 2005. Elements of information theory.

Ralf Engbert, Antje Nuthmann, Eike Richter, and Reinhold Kliegl. 2005. Swift: a dynamical model of saccade generation during reading. *Psychological review*, 112 4:777–813.

Janet D. Fodor and Fernanda Ferreira. 1998. *Reanalysis in Sentence Processing*. Springer Netherlands.

Nicolas Ford, Daniel Duckworth, Mohammad Norouzi, and George Dahl. 2018. The importance of generation order in language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2942–2946, Brussels, Belgium. Association for Computational Linguistics.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814.

Mario Giulianelli, Sarenne Wallbridge, and Raquel Fern'andez. 2023. Information value: Measuring utterance predictability as distance from plausible alternatives. *ArXiv*, abs/2310.13676.

Adam Goodkind and K. Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Workshop on Cognitive Modeling and Computational Linguistics*.

Michael Hahn and Frank Keller. 2023. Modeling task effects in human reading with neural network-based attention. *Cognition*, 230:105289.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *North American Chapter of the Association for Computational Linguistics*.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672.

Trevor Hastie and Robert Tibshirani. 1986. Generalized additive models. *Statistical Science*, 1(3):297–310.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Jukka Hyönä and Pekka Niemi. 1990. Eye movements during repeated reading of a text. *Acta psychologica*, 73 3:259–80.

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

Lena A. Jäger, Felix Engelmann, and Shravan Vasishth. 2017. Similarity-based interference in sentence comprehension: Literature review and bayesian meta-analysis. *Journal of Memory and Language*, 94:316–339.

Carina Kauf and Anna Ivanova. 2023. A better way to do masked language model scoring. *Preprint*, arXiv:2305.10588.

Keren Gruteke Klein, Yoav Meiri, Omer Shubi, and Yevgeni Berzak. 2024. The effect of surprisal on reading times in information seeking and repeated reading. *ArXiv*, abs/2410.08162.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106 3:1126–77.

Roger Levy, Klinton Bicknell, Tim Slattery, and Keith Rayner. 2009. Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50):21086–21090.

Richard L. Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science*, 29 3:375–419.

Richard L. Lewis, Shravan Vasishth, and Julie A. Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in cognitive sciences*, 10 10:447–54.

Brian McElree. 2000. Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, 29(2):111–123.

Clara Meister, Mario Giulianelli, and Tiago Pimentel. 2024. Towards a similarity-adjusted surprisal theory. In *Conference on Empirical Methods in Natural Language Processing*.

Akira Miyake, Marcel A. Just, and Patricia A. Carpenter. 1994. Working memory constraints on the resolution of lexical ambiguity: Maintaining multiple interpretations in neutral contexts. *Journal of Memory and Language*, 33(2):175–202.

Byung-Doh Oh and Tal Linzen. 2025. To model human linguistic prediction, make llms less superhuman. *Preprint*, arXiv:2510.05141.

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.

Byung-Doh Oh and William Schuler. 2025. The impact of token granularity on the predictive power of language model surprisal. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4150–4162, Vienna, Austria. Association for Computational Linguistics.

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.

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

Gary E. Raney and Keith Rayner. 1995. Word frequency effects and eye movements during two readings of a text. *Canadian journal of experimental psychology = Revue canadienne de psychologie experimentale*, 49 2:151–72.

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

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

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.

Omer Shubi and Yevgeni Berzak. 2023. Eye movements in information-seeking reading. In *Annual Meeting of the Cognitive Science Society*.

Omer Shubi, Cfir Avraham Hadar, and Yevgeni Berzak. 2025. Decoding reading goals from eye movements. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5616–5637, Vienna, Austria. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nathaniel J. Smith and Roger Levy. 2008. Optimal processing times in reading: A formal model and empirical investigation. volume 30, pages 595–600.

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

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. Luminosoinsight/wordfreq: v2.2.

Adrian Staub. 2025. Predictability in language comprehension: Prospects and problems for surprisal. *Annual Review of Linguistics*, 11(Volume 11, 2025):17–34.

William Timkey, Kuan-Jung Huang, Byung-Doh Oh, Grusha Prasad, Suhas Arehalli, Tal Linzen, and Brian Dillon. 2025. Eye movements reveal a dissociation between prediction and structural processing in language comprehension. Preprint.

Aditya R. Vaidya, Javier Turek, and Alexander G. Huth. 2023. Humans and language models diverge when predicting repeating text. *ArXiv*, abs/2310.06408.

Julie A Van Dyke and Richard L Lewis. 2003. Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, 49(3):285–316.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

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 Philip Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *ArXiv*, abs/2006.01912.

A Eye-Tracking Measure Definitions

All measures are derived from the first-pass reading sequence. The first pass on word w_i is the contiguous sequence of fixations from when w_i is first fixated until the gaze exits to a different word. Words first reached via a regression from the right are excluded, as their first-pass metrics are undefined.

First Fixation Duration (FFD). The duration of the initial fixation on w_i, reflecting early lexical processing.

Gaze Duration (GD). The sum of all first-pass fixation durations on w_i, serving as an aggregate measure of first-pass processing cost.

Forward Reading Time (FRT). Equal to GD when the first-pass exit saccade is forward (i.e. to a word to the right of w_i), and undefined otherwise. FRT reflects processing cost during fluent, uninterrupted reading.

Pre-Regression Gaze Duration (RGD). Equal to GD when the first-pass exit saccade is regressive (i.e. to a word to the left of w_i), and undefined otherwise. RGD reflects the processing load that triggers a regression. FRT and RGD partition GD by exit direction and are therefore complementary.

Regressive Go-Past Time (RGo-Past). The sum of all fixation durations from first entry into w_i until the gaze first lands on any word to its right, defined only for regression-exit trials. RGo-Past extends RGD to include subsequent rereading of prior material, capturing the full cost of reanalysis.

B Distribution Ablation

To test whether the increase in effects of our stochastic scanpath-conditioned models depend specifically on the choice of linking function, we compare three alternatives under the same scanpath-induced uncertainty distribution $\mathcal{Q}_{t-1}$. The first is our main belief-integrated formulation,

$$-\log \mathbb{E}_{\tilde{H}\sim\mathcal{Q}_{t-1}}\big[p(x_t \mid \tilde{H})\big].$$

The second is expected surprisal,

$$\mathbb{E}_{\tilde{H}\sim\mathcal{Q}_{t-1}}\big[-\log p(x_t \mid \tilde{H})\big],$$

which corresponds to averaging surprisal over latent accessible histories. The third is a point-

	M_1			M_2		
Metric	$-\log \mathbb{E}[P]$	$\mathbb{E}[-\log P]$	**MAP**	$-\log \mathbb{E}[P]$	$\mathbb{E}[-\log P]$	**MAP**
CELER						
FFD	0.02^{**}	0.02^{***}	0.01^{***}	0.02^{**}	$\mathbf{0.02}^{***}$	0.02^{***}
GD	0.56^{**}	0.47^{***}	0.40^{***}	$\mathbf{0.73}^{**}$	0.71^{***}	0.57^{***}
FRT	0.55^{**}	0.46^{***}	0.38^{***}	-1.07^{**}	$\mathbf{0.72}^{***}$	0.55^{***}
RGD	0.75^{**}	-1.79^{***}	$\mathbf{1.80}^{***}$	2.65^{**}	-0.94^{***}	-2.14^{***}
RGo-Past	-1.75^{**}	0.14^{***}	0.11^{***}	0.00	-0.25^{***}	$\mathbf{0.14}^{***}$
Ordinary						
FFD	0.06^{**}	0.06^{***}	0.05^{***}	0.06^{**}	$\mathbf{0.07}^{***}$	0.06^{***}
GD	1.16^{**}	1.07^{***}	0.82^{***}	$\mathbf{2.31}^{**}$	2.08^{***}	2.11^{***}
FRT	1.02^{**}	0.91^{***}	0.65^{***}	$\mathbf{2.37}^{**}$	2.15^{***}	2.15^{***}
RGD	1.27^{**}	1.18^{***}	1.00^{***}	$\mathbf{1.61}^{**}$	1.47^{***}	1.54^{***}
RGo-Past	0.12^{**}	0.12^{***}	0.10^{***}	$\mathbf{0.16}^{**}$	0.15^{***}	0.13^{***}

Table 7: $100 \times \Delta\mathcal{L}$ for expectation variants of M_1 and M_2 against a ModernBERT PLL surprisal baseline using linear terms. p-values are Benjamini-Hochberg corrected across all tests. $^{*}p_{\text{BH}} < .05$, $^{**}p_{\text{BH}} < .01$, $^{***}p_{\text{BH}} < .001$.

	M_1			M_2		
Metric	$-\log \mathbb{E}[P]$	$\mathbb{E}[-\log P]$	**MAP**	$-\log \mathbb{E}[P]$	$\mathbb{E}[-\log P]$	**MAP**
CELER						
FFD	$\mathbf{0.03}^{**}$	$\mathbf{0.03}^{***}$	0.02^{***}	**0.03**	$\mathbf{0.03}^{***}$	0.02^{***}
GD	0.54^{**}	0.44^{***}	0.37^{***}	**0.72**	0.70^{***}	0.53^{***}
FRT	0.53^{**}	0.44^{***}	0.37^{***}	0.73	$\mathbf{0.73}^{***}$	0.54^{***}
RGD	1.38^{**}	-1.33^{***}	-1.31^{***}	**1.57**	-1.32^{***}	0.27^{***}
RGo-Past	$\mathbf{1.39}^{**}$	0.22^{***}	-1.54^{***}	**1.39**	0.23^{***}	-9.73^{***}
Ordinary						
FFD	**0.04**	0.03^{***}	0.02^{***}	**0.04**	$\mathbf{0.04}^{***}$	0.03^{***}
GD	1.32	1.26^{***}	0.92^{***}	**2.20**	1.91^{***}	2.02^{***}
FRT	1.23	1.16^{***}	0.79^{***}	**2.25**	1.99^{***}	2.02^{***}
RGD	1.27	1.19^{***}	0.96^{***}	**1.62**	1.40^{***}	1.52^{***}
RGo-Past	0.06	0.06^{***}	0.04^{***}	**0.07**	$\mathbf{0.07}^{***}$	0.05^{***}

Table 8: $100 \times \Delta\mathcal{L}$ for expectation variants of M_1 and M_2 against a ModernBERT PLL surprisal baseline using non-linear terms. Bold marks the highest value per metric within each dataset. p-values are Benjamini-Hochberg corrected across all tests. $^{*}p_{\text{BH}} < .05$, $^{**}p_{\text{BH}} < .01$, $^{***}p_{\text{BH}} < .001$.

estimate baseline based on the maximum a posteriori (MAP) accessible history,

$$\tilde{H}_t^{\text{MAP}} = \arg\max_{\tilde{H}} \mathcal{Q}_{t-1}(\tilde{H}),$$

with surprisal computed as

$$-\log p(x_t \mid \tilde{H}_t^{\text{MAP}}).$$

All three variants are evaluated using the same underlying accessibility model (M_1 or M_2), differing only in how uncertainty over accessible histories is linked to processing difficulty. This ablation therefore isolates whether predictive gains arise from the scanpath-induced uncertainty itself or from our specific choice to integrate over predictive probabilities before taking surprisal.

We use ModernBERT as the base LM as it significantly outperformed Pythia-70m. In addition, we tested this using both linear and non-linear terms. Linear models are computed as described in Section 6, with surprisal and scanpath predictors entering as plain linear terms. For the non-linear variants, these terms are instead modeled via penalized cubic regression splines $s(\cdot, \texttt{bs="cr"}, k = 6)$, following Wilcox et al. (2023) and Klein et al. (2024). All other aspects of the specification, baseline controls, and cross-validation remain identical across both variants.

Table 7 compares three estimators of scanpath-conditioned surprisal: $-\log \mathbb{E}[P]$, $\mathbb{E}[-\log P]$, and MAP. Across both corpora and models, $-\log \mathbb{E}[P]$ yields the largest or joint-largest $\Delta\mathcal{L}$ for the majority of metrics, most clearly for GD, FRT, and RGD for OneStop. $\mathbb{E}[-\log P]$ and MAP perform comparably to each other but consistently fall below $-\log \mathbb{E}[P]$, with MAP showing the weakest results overall. Table 8 reports the same comparison under non-linear model terms. The ordering across estimators is largely preserved with $-\log \mathbb{E}[P]$ giving the largest or joint-largest $\Delta\mathcal{L}$ across metrics

and corpora, with MAP consistently weakest. Together, the linear and non-linear results empirically motivate $-\log \mathbb{E}[P]$ as the primary estimator.

C BU/IG Ablation

Figure 1 reports full metric results for the BU, IG, and combined ablation across all reading time measures, extending the GD results reported in Table 6 of the main paper.

D Non-Linear Terms

Tables 9 to 12 report full results under non-linear GAM terms for all models and baselines. Effect sizes are modestly larger in some conditions relative to the linear specification, most noticeably for GD and FRT under ModernBERT, but the pattern of results is consistent throughout. The ordering across models holds in both directions: M_2 yields larger $\Delta\mathcal{L}$ than M_1, ModernBERT yields larger gains than Pythia, and the corpus ordering mirrors the linear results. FFD and RGo-Past remain small across all conditions, and all substantive conclusions from the main paper are unchanged. This further is in alignment with the effects seen by (Klein et al., 2024) on OneStop when comparing linear and non-linear fitted GAMs.

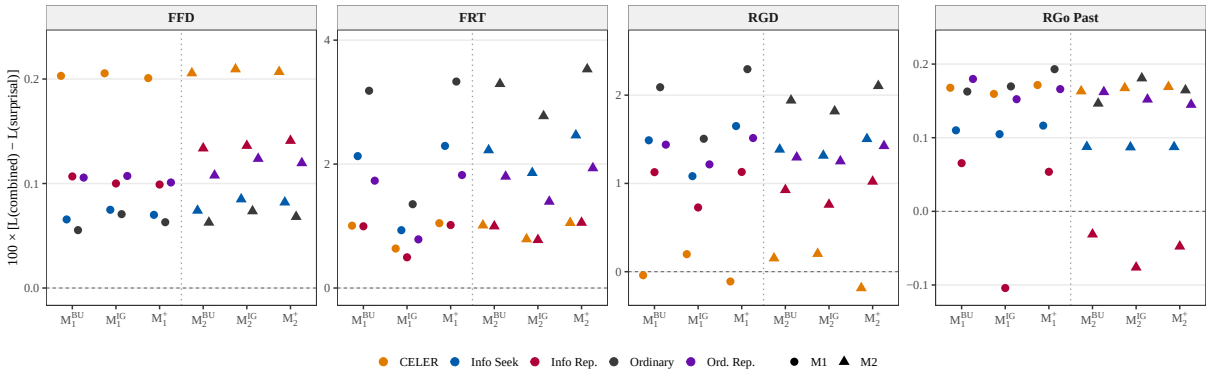

Figure 1: $100 \times \Delta\mathcal{L}$ for BU, IG, and Full (M_1^+/M_2^+) variants against ModernBERT PLL. Dotted line separates M_1 and M_2 groups.

	M_0 Dense					M_0 Sparse				
Metric	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**
FFD	0.01^{**}	0.01^{**}	0.02^{**}	0.02^{**}	0.02^{**}	0.01^{**}	0.02^{**}	$\mathbf{0.05}^{**}$	0.00	0.02
GD	0.05^{**}	0.20^{**}	0.09^{**}	0.28^{**}	0.20^{**}	0.19^{**}	0.31^{**}	0.22^{**}	$\mathbf{0.43}^{**}$	0.29^{**}
FRT	0.04^{**}	0.09^{**}	0.06^{**}	0.12^{**}	0.09^{**}	0.17^{**}	0.28^{**}	0.21^{**}	$\mathbf{0.33}^{**}$	0.30^{**}
RGD	0.63^{**}	0.63^{**}	0.16^{**}	0.64^{**}	0.61^{**}	$\mathbf{0.75}^{**}$	0.38^{**}	0.17^{**}	0.56^{**}	0.38^{**}
RGo-Past	$\mathbf{0.57}^{**}$	-0.01	-0.05	-0.02^{*}	-0.01	-0.21^{*}	0.00	-0.07^{*}	0.03^{**}	0.13^{**}

Table 9: $100 \times \Delta\mathcal{L}$ for M_0 Dense and Sparse vs. Pythia-70m. p_{BH}: $^{*}<.05$, $^{**}<.01$, $^{***}<.001$.

	M_0 Dense					M_0 Sparse				
Metric	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**
FFD	0.01^{**}	0.09^{**}	$\mathbf{0.19}^{**}$	0.04^{**}	0.05^{**}	0.00^{**}	0.03^{**}	0.10^{**}	-0.01^{**}	0.00
GD	0.08^{**}	0.18^{**}	0.18^{**}	0.32^{**}	0.21^{**}	0.11^{**}	0.37^{**}	0.14^{**}	0.42^{**}	0.21^{**}
FRT	0.06^{**}	0.20^{**}	0.25^{**}	0.21^{**}	0.21^{**}	-0.08^{**}	0.57^{**}	0.25^{**}	$\mathbf{0.70}^{**}$	0.37^{**}
RGD	-2.19^{**}	0.26^{**}	0.16	$\mathbf{0.55}^{**}$	0.44^{**}	-2.63^{**}	0.13^{**}	-0.00	0.18^{**}	0.09^{**}
RGo-Past	$\mathbf{1.26}^{**}$	-0.08^{**}	-0.03	-0.06^{**}	0.06	-8.12^{**}	0.04^{**}	-0.05^{**}	0.02^{*}	0.10^{**}

Table 10: $100 \times \Delta\mathcal{L}$ for M_0 Dense and Sparse vs. ModernBERT. p_{BH}: $^{*}<.05$, $^{**}<.01$, $^{***}<.001$.

	M_1					M_2				
Metric	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**
FFD	0.02^{**}	0.03^{**}	0.05^{**}	0.00	0.05^{**}	0.02^{**}	0.03^{**}	0.04^{**}	0.00	$\mathbf{0.06}^{**}$
GD	0.17^{**}	$\mathbf{0.56}^{**}$	0.25^{**}	0.28^{**}	0.51^{**}	0.20^{**}	0.46^{**}	0.30^{**}	0.29^{**}	0.50^{**}
FRT	0.16^{**}	$\mathbf{0.56}^{**}$	0.26^{**}	0.24^{**}	0.50^{**}	0.19^{**}	0.46^{**}	0.31^{**}	0.26^{**}	0.53^{**}
RGD	-0.30^{**}	0.58^{**}	0.12	0.48^{**}	$\mathbf{0.67}^{**}$	-0.54^{**}	0.52^{**}	0.21^{*}	0.46^{**}	0.51^{**}
RGo-Past	$\mathbf{0.99}^{**}$	0.04^{**}	-0.03	0.02	0.17^{**}	-2.26^{**}	0.01	-0.05	0.02^{*}	0.10^{**}

Table 11: $100 \times \Delta\mathcal{L}$ for M_1 and M_2 against a Pythia-70m surprisal baseline. $^{*}p_{\text{BH}} < .05$, $^{**}p_{\text{BH}} < .01$, $^{***}p_{\text{BH}} < .001$.

	M_1					M_2				
Metric	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**	**CELER**	**Info Seek**	**Info Rep.**	**Ordinary**	**Ord. Rep.**
FFD	0.03^{**}	0.06^{**}	0.04	0.04^{**}	0.09^{**}	0.03^{**}	0.07^{**}	0.09^{**}	0.04^{**}	$\mathbf{0.10}^{**}$
GD	0.54^{**}	0.87^{**}	0.53^{**}	1.32^{**}	0.72^{**}	0.72^{**}	1.39^{**}	0.62^{**}	$\mathbf{2.20}^{**}$	1.16^{**}
FRT	0.53^{**}	0.86^{**}	0.52^{**}	1.23^{**}	0.68^{**}	0.73^{**}	1.47^{**}	0.65^{**}	$\mathbf{2.25}^{**}$	1.20^{**}
RGD	1.38^{**}	0.92^{**}	0.70^{**}	1.27^{**}	1.02^{**}	1.57^{**}	1.09^{**}	0.59^{**}	$\mathbf{1.62}^{**}$	1.09^{**}
RGo-Past	1.39^{**}	-0.02	-0.03	0.06^{**}	0.05	$\mathbf{1.39}^{**}$	-0.02	-0.03	0.07^{**}	0.09^{*}

Table 12: $100 \times \Delta\mathcal{L}$ for M_1 and M_2 against a ModernBERT surprisal baseline. $^{*}p_{\text{BH}} < .05$, $^{**}p_{\text{BH}} < .01$, $^{***}p_{\text{BH}} < .001$