The Effect of Word Predictability on Reading Times in Information Seeking and Repeated Reading

Keren Gruteke Klein[1], **Yoav Meiri**[1], **Omer Shubi**[1], **Yevgeni Berzak**[1,2]
[1]Faculty of Data and Decision Sciences
Technion - Israel Institute of Technology, Haifa, Israel
[2]Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology, Cambridge, USA
{gkeren,meiri.yoav,shubi}@campus.technion.ac.il, berzak@technion.ac.il

Abstract

The effect of word predictability on processing difficulty has been a central topic of investigation in psycholinguistics. Here, we use a broad coverage reading corpus in English to examine three language processing regimes that are common in daily life but have not been addressed with respect to this question: information seeking, repeated processing, and the combination of the two. Using standard, reading regime agnostic surprisal estimates, we find that the prediction of surprisal theory regarding a logarithmic relation between word predictability and processing time extends to these regimes. However, when examining surprisal estimates using regime-specific contexts, we also obtain two findings which are at odds with the predictions of surprisal theory. First, we find that in information seeking, such estimates do not improve the predictive power of processing times compared to surprisals from standard regime-agnostic contexts. Further, repeated presentation contexts yield near zero surprisal estimates with null effects on repeated processing times. These results suggest a misalignment of memory and context representation mechanisms between humans and current language models.[1]

1 Introduction

A key question in psycholinguistics concerns the cognitive processes that underlie the real-time integration of new linguistic material with previously processed linguistic context. A central framework for examining this question is surprisal theory (Hale, 2001; Levy, 2008). This theory ties word processing cost to the word's surprisal, and predicts a logarithmic relation between word predictability and processing difficulty. Due to its theoretical implications (see Shain et al. (2024b) for an extended discussion), multiple studies have tested this prediction empirically with native speakers of multiple languages (Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020; Brothers and Kuperberg, 2021; Berzak and Levy, 2023; Wilcox et al., 2023; Shain et al., 2024b; Xu et al., 2023). All these studies find significant word predictability effects, and most are consistent with a logarithmic relation between predictability and reading times.

However, thus far this relation has been examined only in one experimental paradigm, which can be referred to as *ordinary reading*. This paradigm presupposes that the participant did not have prior, or at least recent, exposure to the linguistic material. It further assumes that they have no specific goals beyond general comprehension of this material. These assumptions do not hold in many daily situations, where language comprehenders often have specific goals with respect to the linguistic input and/or process the same input multiple times. This limits the generality of the conclusions that can be drawn from prior studies.

In this work, we examine the effect of word predictability on reading times in English L1 in three common, but understudied language processing regimes: (1) information seeking, (2) repeated processing, and (3) the combination of the two. Prior work on information seeking (Hahn and Keller, 2023; Shubi and Berzak, 2023) and repeated reading (Hyönä and Niemi, 1990; Raney and Rayner, 1995; Meiri and Berzak, 2024) has shown substantial differences in eye movement patterns in these regimes compared to ordinary reading, and the extent to which the predictions of surprisal theory hold in these regimes is currently unknown.

We analyze and compare the functional form and predictive power of two types of contexts, standard regime-agnostic contexts that capture the general predictability of a word, and regime-specific contexts which include the task in information seeking and a prior appearance of the linguistic content in repeated reading. We examine two main hypotheses stemming from the predictions of surprisal theory: (1) the functional form of the re-

[1]Code and data will be made publicly available.

lation between standard word predictability and reading times should be logarithmic in any reading regime and (2) word predictability estimates from regime-specific contexts should yield higher predictive power for reading times in the respective regimes compared to regime-agnostic contexts, due to a more accurate representation of the context and the processing goals, which should lead to better alignment with subjective word probabilities.

Our main results are the following:

1. **Regime-agnostic contexts** yield robust logarithmic predictability effects in information seeking, repeated reading and their combination.

2. **Regime-specific contexts** do not improve the predictive power of word probabilities for reading times compared to standard regime-agnostic contexts.

    (a) In information seeking, providing the information seeking task in the context does not improve the model's predictive power for reading times.

    (b) In repeated processing, providing a prior appearance of the linguistic material leads to in-context memorization, with surprisal values close to zero and no predictive power for reading times.

2 Related Work

The first study to empirically examine the relation between word predictability and reading times was Smith and Levy (2013). They used broad coverage eye-tracking and self-paced reading data for English, and found evidence for a logarithmic relation. Following this work, several studies obtained similar results consistent with a logarithmic relation using additional corpora and a number of different methodologies for curve fitting and testing linearity, including Goodkind and Bicknell (2018), Wilcox et al. (2020) and Shain et al. (2024b). Wilcox et al. (2023) and Xu et al. (2023) obtained similar results across a variety of languages other than English. Brothers and Kuperberg (2021) found a linear relation using a controlled self-paced reading experiment and cloze estimates of word probabilities. Re-analysis of this data with language model probabilities resulted in a logarithmic relation (Shain et al., 2024a). Our study continues this line of work and extends it to different reading regimes.

Both information seeking and repeated reading have received limited attention in psycholinguistics. Work that examined information seeking (Hahn and Keller, 2023; Shubi and Berzak, 2023) found substantial differences in eye movement patterns compared to ordinary reading, both in fixation and in saccade patterns. The differences are especially driven by the division to task-relevant and task-irrelevant information in this regime. Different eye movement behavior was also found in repeated reading, where among others, shorter reading times and longer saccades were observed (Hyönä and Niemi, 1990; Raney and Rayner, 1995).

Vaidya et al. (2023) found that in a repeated reading cloze task, language models have substantially higher next word prediction accuracy compared to humans. They further identified a link between "induction heads", which are attention heads that recognize repeated token sequences and increase the probability of the previously observed continuation (Elhage et al., 2021), as a core contributor to this behavior in language models. Our findings for repeated reading are in line with these results.

3 Data

We use OneStop, a broad coverage eye movements in reading dataset in English with 360 L1 participants, collected by Malmaud et al. (2020) with an Eyelink 1000+ eye-tracker (SR Research). The experiment was conducted under an institutional IRB protocol, and all the participants provided written consent before participating in the study. The textual materials are taken from OneStopQA (Berzak et al., 2020) and comprise 30 articles from the Guardian with 4-7 paragraphs (162 paragraphs in total). Each paragraph in OneStopQA is accompanied by 3 multiple choice reading comprehension questions. The textual span in the paragraph which contains the essential information for answering the question correctly, called the critical span, is manually annotated in each paragraph for each question.

An experimental trial consists of reading a single paragraph on a page, followed by answering one reading comprehension question on a new page without the ability to go back to the paragraph.

Ordinary reading vs information seeking 180 participants are in an ordinary reading regime in which they see the question only after having read the paragraph. This setup is similar to eye-tracking corpora used in prior studies. The remaining 180 participants are in an information-seeking regime in which the question (but not the answers) is presented prior to reading the paragraph.

First vs repeated reading Each participant reads 10 articles in a random presentation order, followed by two articles that are presented for a second time with identical text but with a different question for each paragraph. The article in position 11 is a repeated presentation of the article in position 10. The article in position 12 is a repeated presentation of one of the articles in positions 1–9. Thus, OneStop contains both consecutive and non-consecutive repeated reading at the article level.

OneStop has 2,532,799 data points (i.e. word tokens over which eye-tracking data was collected). Following standard practice, we exclude words that were not fixated, words with a total reading time greater than 3,000 ms, words that start or end a paragraph, words with punctuation, and surprisal values greater than 20 bits. After these filtering steps, we remain with 1,157,609 data points: 541,875 in first ordinary reading, 474,674 first reading information seeking, 82,357 repeated ordinary reading, and 58,703 repeated reading information seeking.

4 Methodology

We examine four different reading regimes that take advantage of the experimental manipulations in OneStop and reflect different types of interactions with the text. The first is ordinary reading during the first presentation of the text. This regime corresponds to the standard experimental setup in reading studies. Additionally, new to this work, we examine information seeking during first reading, and both ordinary reading and information seeking during repeated text presentation.

We estimate the functional form of the relation between word predictability and reading times in each of the four reading regimes using Generalized Additive Models (GAMs, Hastie and Tibshirani, 1986), which can fit non-linear relations between predictors and responses. We predict word reading times from word predictability and two control variables that have been shown to be predictive of reading times above and beyond predictability: word frequency and word length (Kliegl et al., 2004; Clifton Jr et al., 2016). To account for spillover effects from the previous word (Rayner, 1998), our models also include the predictability, frequency and length of the previous word.

Following prior work (e.g. Wilcox et al., 2023) our primary reading time measure is **first pass Gaze Duration**; the time from first entering a word to first leaving it during first pass reading. This measure is associated with the processing difficulty of a word given left-only context and is thus especially suitable for benchmarking against surprisal. In the Appendix, we examine additional measures: Gaze Duration and Total Fixation Duration. For completeness, we also provide results for first pass First Fixation duration and First Fixation duration, which tend to have small surprisal effects and are associated with lexical processing (Clifton Jr et al., 2007; Berzak and Levy, 2023). The definitions of all the measures are provided in the Appendix.

We estimate word predictability using surprisal, defined as $-\log p(w_i|w_{<i})$, where w_i is the current word and $w_{<i}$ is the preceding context. Surprisal values are estimated using a language model (see Section 4.3 below). In cases where a given word is split by the language model to several tokens, we sum the token surprisals. Frequency is defined as $-\log p(w_i)$, using word counts from Wordfreq (Speer et al., 2018). Word length is measured in number of characters excluding punctuation.

We define three models of interest, all fitted using `mgcv` (v1.9.1) `gam` (Wood, 2004) function with cubic splines ("`cr`")[2]:

- **Baseline model** which predicts reading times of the current word from the control variables frequency and length and their interaction using tensor product terms `te`.[3]
- **Linear model** which includes the baseline model terms and linear terms for the surprisal of the current and the previous words.[4]
- **Non-linear model** which includes the baseline model terms and smooth terms `s` for the surprisal of the current and previous words.[5]

4.1 Analysis 1: GAM Visualization

In this analysis, we visualize the relationship between surprisal and reading times using the linear and non-linear models. If the less constrained non-linear fit is visually similar to the linear fit, this

[2]The models do not include random effects due to convergence issues.

[3]Model formula in R:
$RT \sim te(freq, len) + te(freq\_prev, len\_prev)$

[4]Model formula in R: $RT \sim surp + surp\_prev + te(freq, len) + te(freq\_prev, len\_prev)$

[5]Model formula in R:
$RT \sim s(surp, k = 6) + s(surp\_prev, k = 6) + te(freq, len) + te(freq\_prev, len\_prev)$

would provide initial evidence for a logarithmic relation between predictability and reading times. To this end, we fit each of the two models on the reading time data of each of the four reading regimes, and predict reading times for surprisal values in the range of 0-20 in 0.1 increments.

4.2 Analysis 2: Predictive Power

Complementary to analysis 1, we measure the increase in the model's log-likelihood relative to the baseline model, which includes only the control variables frequency and length, without surprisal, for both the linear and the non-linear models. A statistically significant difference in the predictive power of the non-linear and linear models would provide evidence against linearity. Following prior work (e.g. Wilcox et al., 2020; Oh and Schuler, 2022; Wilcox et al., 2023), we measure predictive power for data point i using delta log-likelihood:

$$\Delta LL_i = \log L^{target}(RT_i|x^{target}) - \log L^{baseline}(RT_i|x^{baseline})$$

where RT_i is the reading measure, $x^{baseline}$ are the control predictors and x^{target} are the target predictors, which include the control predictors and surprisal. L^M is the likelihood under the model M:

$$L^M(RT_i|x) = f_{norm}(RT_i|\mu = \hat{RT_i}, \sigma^2 = \sigma^2_{RT})$$

where $\hat{RT_i}$ is the RT prediction of the model M given the predictor set x, σ^2_{RT} is the standard deviation of the residuals of the fitted GAM model M and f_{norm} is the Gaussian density function.

We examine ΔLL, the per-word mean of ΔLL_i. To reduce the risk of overfitting, we measure ΔLL on held-out data that was not used to fit the models, using 10-fold cross validation. A positive ΔLL indicates that the addition of surprisal terms increases the predictive power of the model. We then compare the ΔLL of the linear and non-linear models. If there is no significant difference between the two, we do not reject the null hypothesis of a linear relation between surprisal and reading times. Following Wilcox et al. (2023), we test the significance of the differences in the ΔLL of the two models using a paired permutation test.

4.3 Language Models

An important methodological consideration for our study is the choice of the language model. Our selection criteria for the language model is predictive power, as measured by ΔLL. We measure the predictive power of 30 publicly available language models on the OneStop reading time data, and select the model with the highest predictive power across the four reading regimes and the linear and non-linear models.

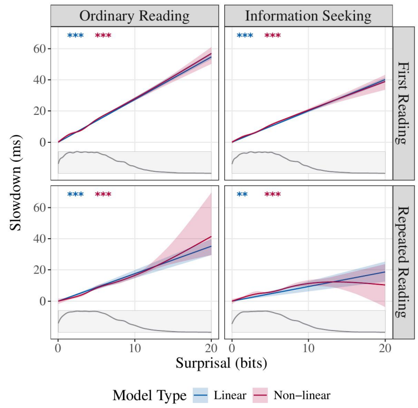
We examine models from the GPT-2 (Radford et al., 2019), GPT-J (Wang and Komatsuzaki, 2021), GPT-Neo (Black et al., 2021), Pythia (Biderman et al., 2023), OPT (Zhang et al., 2022), Mistral (Jiang et al., 2023), Gemma (Team, 2024) and Llama-2 (Touvron et al., 2023) families, ranging from 70 million to 70 billion parameters. We note that this list includes GPT-2-small, which was used in prior work for related analyses (Oh and Schuler, 2022; Shain et al., 2024b). Figure 1 in the Appendix presents model predictive power as a function of the model's log perplexity measured on the 30 articles of OneStopQA. This comparison yields **Pythia-70m** as the model with the highest predictive power.[6] Our main analyses therefore use surprisal estimates from this model. To test the robustness of the results to the choice of language model, in the Appendix we present additional analyses with the remaining 29 models.

4.4 Contexts

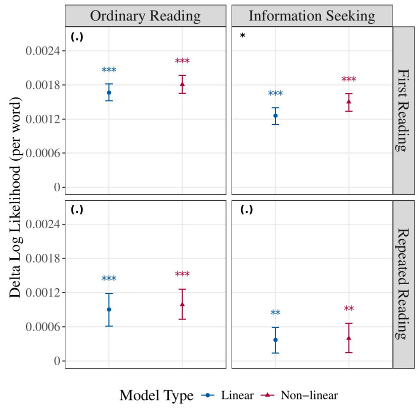
A second key methodological consideration concerns the context $w_{<i}$ that is provided to the language model for estimating the probability of the current word $w_{<i}$. We examine three approaches for the words included in context.

- **Standard Context**: In the first, regime-agnostic approach, which we take in Section 5, the context consists of the words preceding the current word in the paragraph.
- **Regime Context**: In the second, regime-specific approach, in Section 6, the context depends on the reading regime in that it includes the preceding question in information seeking and the paragraph in repeated reading.
- **Prompting+Regime Context**: An additional variant of the Regime Context in Section 6 which further includes textual prompts that emulate the instructions given to human readers.

[6] We note that this figure replicates the results of Oh and Schuler (2022) regarding the relation between perplexity and predictive power for recent language models, and extends them to non-ordinary reading regimes.



(a) GAM fits for the relation between surprisal and reading times, with bootstrapped 95% confidence intervals. Top left of each plot, the statistical significance of the s and linear terms of the current word's surprisal. At the bottom of each plot: a density plot of surprisal values.



(b) ΔLL means with 95% confidence intervals on held-out data using 10-fold cross validation. Above each confidence interval: the statistical significance of a permutation test that checks if the ΔLL is different from zero. Top left of each plot: the statistical significance of a permutation test for a difference between the ΔLL of the linear and non-linear models.

Figure 1: (a) GAM fits and (b) ΔLL for first pass Gaze Duration and Pythia-70m surprisals with standard context, using the linear and non-linear models. '***' $p < 0.001$, '**' $p < 0.01$. '*' $p < 0.05$, '(.)' $p > 0.05$. **Key results**: (a) Approximately linear curves for the non-linear models. (b) No statistically significant differences in the ΔLL of the linear and non-linear models, with the exception of information seeking in first reading. Smaller ΔLL in information seeking and repeated reading compared to first reading - ordinary reading for both models.

5 Surprisal with Standard Context

In our first set of analyses, we follow prior work on ordinary first reading, as well as information seeking and repeated reading (Shubi and Berzak, 2023; Meiri and Berzak, 2024), and use standard, reading regime-agnostic surprisal estimates, which are obtained by conditioning the model on the prior textual material in the paragraph.

5.1 GAM Visualization

Figure 1a presents the GAM surprisal curves for the linear and non-linear models. Visual inspection suggests that the non-linear model approximately tracks the linear fit. We further note that consistently with the findings of Shubi and Berzak (2023) and Meiri and Berzak (2024), surprisal effects, which can be inferred from the slope of the curves, are smaller in information seeking compared to ordinary reading, and smaller in repeated reading compared to first reading.

Figure 2a in the Appendix suggests that the results largely hold across different language models, although some of the models with the lowest perplexity also yield sublinear fits. Figure 3a in the Appendix examines additional reading measures for Pythia-70m, with linear fits for Gaze Duration and Total Fixation duration, and mixed results for first pass First Fixation and First Fixation where we observe sublinear curves in first reading. Overall, most curves of the non-linear models appear to approximate their linear counterparts.

In information seeking, Shubi and Berzak (2023) have shown different eye movement patterns over task critical information (the critical span) and task non-critical information. In repeated reading, Meiri and Berzak (2024) also showed differences in eye movements when repeated reading of an article happens immediately after the first presentation (article 11) compared to repeated reading with intervening articles (article 12). Figure 4 in the Appendix shows that linearity for first pass Gaze Duration holds both within and outside the critical span in information seeking, and also both with and without intervening articles during repeated reading.

5.2 Predictive Power

While visual inspection provides initial evidence for the linearity of reading times in surprisal across reading regimes, we further test this hypothesis by comparing the predictive power of the non-linear model relative to that of the linear model. Figure 1b presents the ΔLL of the linear and non-linear models for first pass Gaze Duration across the four reading regimes. We find that in three of the four regimes, there is no significant difference between the ΔLL of the two models. In information seeking - first reading, the difference is significant at

Regime	Standard Context	Regime Context	Description	Prompting + Regime Context	Prompt Text
First reading Ordinary reading	P	P	The preceding words in the paragraph.	pr1 + P	pr1: "You will now read a paragraph."
First reading Information seeking	P	Q + P	The question followed by the preceding words in the paragraph.	pr1 + Q + P	pr1: "You will now be given a question about a paragraph followed by the paragraph. You will need to answer the question."
Repeated reading Ordinary reading	P	P + P	The entire paragraph followed by the preceding words in the same paragraph.	pr1 + P + pr2 + P	pr1: "You will now read a paragraph." pr2: "You will now read the same paragraph again."
Repeated reading Information seeking	P	Q' + P + Q + P	The question for the first reading, followed by the paragraph, the question for the second reading and the preceding words in the same paragraph.	pr1 + Q' + P + pr2 + Q + P	pr1: "You will now be given a question about a paragraph followed by the paragraph. You will need to answer the question." pr2: "You will now read the same paragraph again with a different question before the paragraph. You will need to answer the question."

Table 1: Standard and regime-specific contexts provided to language models. 'pr' stands for prompt, Q and Q' for questions and P for paragraph.

$p < 0.05$. These results largely support our conclusion from the visual inspection of the GAM curves, that the surprisal - reading times relation is linear in all four regimes. We further note, that in line with the effect sizes, the predictive power of standard surprisal estimates is smaller in information seeking compared to ordinary reading, and smaller in repeated reading compared to first reading ($p < 0.05$ in all cases using a paired permutation test).

Figure 2b in the Appendix presents the results for first pass Gaze duration across different language models, suggesting that they are robust to the language model choice. Figure 3b in the Appendix presents additional reading measures and further shows that the results mostly extend to Gaze Duration and Total Fixation Duration, while mixed results are obtained for First Fixation measures, with larger ΔLL for the non-linear model in ordinary reading and information seeking during first reading. Figure 4 shows that the linearity of first pass Gaze Duration in surprisal holds both within and outside the critical span in information seeking, as well as for consecutive and non-consecutive article repeated reading. Overall, our analysis of ΔLL favors a linear relation between surprisal and reading times across all four reading regimes.

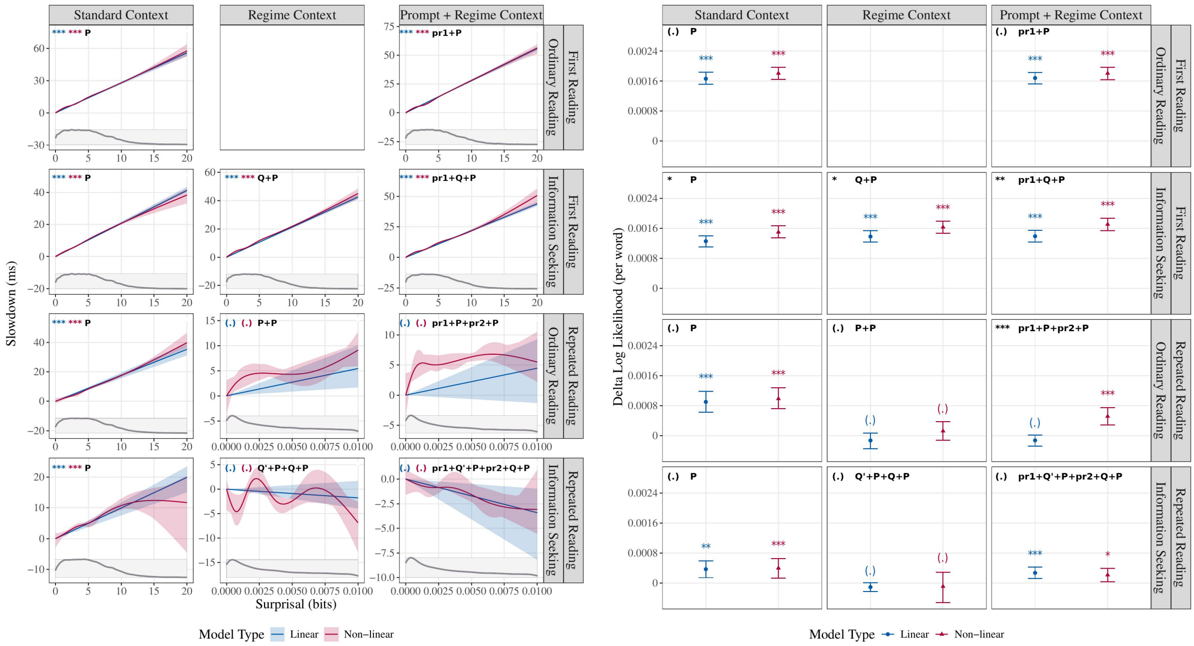
6 Comparison to Surprisal with Regime-Specific Context

Thus far, we used surprisal estimates based on the textual context in the paragraph. However, this context does not fully capture the reading task conditioning in the human data. Human participants in the first reading – information seeking regime receive a question prior to reading the paragraph. In repeated ordinary reading they have already read that paragraph. In repeated reading during information seeking they have previously read the paragraph and received a question prior to both the first and the second reading of the paragraph. These manipulations can alter linguistic expectations and were previously shown to influence reading times (Hyönä and Niemi, 1990; Malmaud et al., 2020; Shubi and Berzak, 2023; Meiri and Berzak, 2024). Furthermore, human participants receive explicit instructions regarding the different trial components in the reading experiment.

In the remainder of this work, we compare our results using the standard regime-agnostic surprisal estimates to surprisal estimates based on context types that more closely match the textual contexts and instructions presented to humans in each of the reading regimes. Our analyses focus on the following questions regarding the three regimes that are not ordinary first reading. (1) Does the functional form of the relation between surprisal and reading times remain linear under regime-conditioned surprisal estimates? (2) Do regime-conditioned surprisals lead to better predictive power for human reading times?

To address these questions, in addition to the standard context used in Section 5, we examine three **regime contexts** that correspond to each of the three reading regimes that involve information seeking and repeated reading. To further enhance the similarity to the experimental setup in the human data, we also examine a variant of the regime contexts in which the model additionally receives **prompts** that emulate the reading instructions received by human participants. The regime-specific contexts are presented in Table 1.

We note that although these contexts include the essential components of each reading regime, they do not fully match the eye-tracking experiment as



(a) GAM fits for the relation between surprisal and reading times across context types. Slowdown effects in *ms* for first pass Gaze Duration as a function of surprisal, with bootstrapped 95% confidence intervals. Top left of each plot, the significance of the s and linear terms of the current word's surprisal. At the bottom of each plot: a density plot of surprisal values. **Key results** for the Regime Context and Prompt + Regime Context: (a) in first reading - information seeking, approximately linear curves for the non-linear model. (b) In the two repeated reading conditions, surprisal values are close to zero and a null surprisal effect.

(b) ΔLL means with 95% confidence intervals on held-out data using 10-fold cross validation. Above each confidence interval: the statistical significance of a permutation test that checks if the ΔLL is different from zero. Top left of each plot: statistical significance of a permutation test for a difference between the ΔLL of the linear and non-linear models. **Key results** for Regime Context and Prompt + Regime Context: (1) In first reading - information seeking, no statistically significant differences in the ΔLL of the linear and non-linear models, and no increase in ΔLLs compared to the Standard Context. (2) In both repeated reading regimes, ΔLLs are *lower* compared to the Standard Context and not significantly above zero.

Figure 2: Comparison of GAM fits and ΔLL for first pass Gaze Duration with surprisal estimates of Pythia-70m from different context types. '***' $p < 0.001$, '**' $p < 0.01$. '*' $p < 0.05$, '(.)' $p > 0.05$.

they do not include intervening textual material between first and second presentations of a paragraph. The context window of our models is too small to include the text of an entire experimental session. However, in Table 1 in the Appendix we present a scheme for article-level analysis for articles 10 and 11, which fully matches the eye-tracking experiment for these articles, thus enabling to analyze cases with 3-6 intervening paragraphs between two readings of the same paragraph. We use this scheme with the Pythia-70m model, for which we employ a sliding window mechanism with an overlap size that ensures that each paragraph's first appearance is fully included in the context window of its repeated appearance.

6.1 GAM Visualization

In figure 2a we present GAM visualizations for the relationship between Pythia-70m surprisal and first pass Gaze Duration using the models presented in Section 4. We compare surprisals from conditioning on the standard paragraph context P to surprisals from reading regime contexts: Q+P for first reading - information seeking, P+P for repeated reading - ordinary reading, and Q'+P+Q+P for repeated reading - information seeking. We further present results for regime contexts with prompting.

For first reading - information seeking, surprisals from both regime-specific contexts yield linear curves. However, a very different behavior is observed in the repeated reading regimes. In these regimes, there is a collapse of the surprisals to values that are close to zero, as can be seen in the density plots. Moreover, we find null effects of surprisal on reading times. Thus, we obtain two different behaviors for information seeking and repeated reading. While the addition of the information seeking task does not substantially alter the predictive power of the model, conditioning twice on the paragraph leads to surprisal values that no longer maintain a significant relation to reading times.

6.2 Comparison of Predictive Power

In figure 2b we compare the ΔLL of the linear and non-linear models across standard and regime-specific surprisals with and without prompting. In first reading - information seeking, the regime context and the prompt + regime context provide weak evidence against linearity ($p = 0.04$ and $p = 0.01$ respectively). Interestingly, regime conditioning and prompting do not improve predictive power in this regime; the ΔLL of the regime context is not significantly higher compared to the standard context ($p = 0.25$ linear; $p = 0.27$ non-linear, using a paired permutation test). Adding prompting yields similar outcomes compared to the standard context ($p = 0.22$ linear; $p = 0.08$ non-linear).

In the repeated reading regimes we observe a different pattern. Importantly, the regime contexts in the ordinary reading condition lead to a *decrease* in the ΔLL compared to the standard context in both the linear ($p = 0.001$) and non-linear cases ($p = 0.009$). A similar pattern is observed when adding prompting with ($p = 0.001$) for the linear model and ($p = 0.038$) for the non-linear model. The regime contexts in the information seeking condition exhibit the same pattern of ΔLL decrease compared to the standard context, which is significant both without prompting ($p = 0.017$ linear; $p = 0.004$ non-linear) and with prompting ($p = 0.091$ linear; $p = 0.027$ non-linear). Furthermore, in nearly all cases the regime context ΔLL is not significantly above zero, suggesting that the corresponding surprisal estimates have no predictive power with respect to reading times. Taken together with the GAM visualizations in Figure 2a, we conclude that the examined language models are misaligned with human reading patterns in repeated reading, and do not provide useful surprisal estimates when conditioned for repeated reading.

These results are consistent across all the models examined, and specifically for the larger models, which could a-priori be expected to be more sensitive to context conditioning and prompting. In the Appendix, we present these results for GPT-2-small in Figure 5 and for the two largest models Llama 70b in Figure 6, and Mistral Instruct v0.3 in Figure 7. Furthermore, they generalize to repeated reading *with the intervening material* between paragraph readings as it appeared in the human eye-tracking experiment in articles 10 and 11. These results are presented for Pyhtia-70m in Appendix Figure 8.

7 Discussion and Conclusion

Surprisal theory predicts a logarithmic relationship between word predictability and word processing times. This prediction found support in studies with ordinary reading, but was not previously examined in information seeking and repeated reading. We find evidence that with standard surprisal estimates, the prediction of surprisal theory for a logarithmic effect of surprisal on reading times holds in these regimes. Effect size and predictive power of standard surprisal estimates are smaller in information seeking and repeated reading.

Our attempt to improve language model predictive power with regime-specific contexts yields two findings that are not in line with surprisal as a "causal bottleneck" for observed behavior (Levy, 2008), in that better representation of the context does not lead to better predictions of processing difficulty. First, we find that regime-specific surprisal estimates in first reading - information seeking do not improve the fit to human reading times. A more severe case of estimation collapse is observed in repeated reading, where in-context memorization leads to near zero surprisal estimates with no predictive power for reading times.

These findings highlight two different aspects of misalignment of context representation in language models and humans. The first case demonstrates a misalignment in the representation of task information. The second suggests very different memory and retrieval abilities in humans and current language models. These misalignments question not only the suitability of current language models as cognitive models, but also the psycholinguistic relevance of quantities extracted from such models.

We propose two possible explanations for the discrepancies in the real-time processing and memory mechanisms of humans and language models. The first explanation is that this mismatch stems from architectural aspects of current language models, and can be alleviated or even completely resolved with architectural or training procedure changes to said models. The second explanation, which poses a more substantial challenge to surprisal theory, is that they are related to factors that come into play in non-ordinary processing regimes that cannot be encoded in word probabilities irrespective of the language model used. Future work is required to adjudicate between these two explanations.

References

Yevgeni Berzak and Roger Levy. 2023. Eye movement traces of linguistic knowledge in native and non-native reading. *Open Mind*, pages 1–18.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. Starc: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Trevor Brothers and Gina R Kuperberg. 2021. Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116:104174.

Charles Clifton Jr, Fernanda Ferreira, John M Henderson, Albrecht W Inhoff, Simon P Liversedge, Erik D Reichle, and Elizabeth R Schotter. 2016. Eye movements in reading and information processing: Keith rayner's 40 year legacy. *Journal of Memory and Language*, 86:1–19.

Charles Clifton Jr, Adrian Staub, and Keith Rayner. 2007. Eye movements in reading words and sentences. *Eye movements*, pages 341–371.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

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

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*.

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

Jukka Hyönä and Pekka Niemi. 1990. Eye movements during repeated reading of a text. *Acta psychologica*, 73(3):259–280.

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

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

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

Jonathan Malmaud, Roger Levy, and Yevgeni Berzak. 2020. Bridging information-seeking human gaze and machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 142–152.

Yoav Meiri and Yevgeni Berzak. 2024. Déjà vu: Eye movements in repeated reading. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Byung-Doh Oh and William Schuler. 2022. 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.

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

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 expérimentale*, 49(2):151.

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

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024a. Are word predictability effects really linear? a critical reanalysis of key evidence. In *37th Annual Conference on Human Sentence Processing*.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024b. 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 *Proceedings of the Annual Meeting of the Cognitive Science Society*, Online.

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.

Gemma Team. 2024. Gemma.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Aditya Vaidya, Javier Turek, and Alexander Huth. 2023. Humans and language models diverge when predicting repeating text. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 58–69.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. 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*, page 1707–1713.

Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *arXiv preprint arXiv:2307.03667*.

Simon N Wood. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686.

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

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.