N-gram-like Language Models Predict Naturalistic Reading Time Best

James A. Michaelov (jamic@mit.edu)[1,2] **& Roger P. Levy (rplevy@mit.edu)**[1]
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[2]MIT Libraries CREOS, Massachusetts Institute of Technology

Abstract

Recent work has found that contemporary language models such as transformers can become so good at next-word prediction that the probabilities they calculate become worse for predicting naturalistic reading time. In this paper, we propose that this can be explained by reading time being shaped by simple *n*-gram statistics rather than the more complex statistics learned by state-of-the-art transformer language models. We demonstrate that the neural language models whose predictions are most correlated with *n*-gram probability are also those that calculate probabilities that are the most correlated with eye-tracking-based metrics of reading time on naturalistic text.

Keywords: reading; language processing; language models

Introduction

Two decades ago, researchers discovered that a word's bigram probability—that is, the probability of the word occurring given a specific previous word—is significantly correlated with the time taken to read the word, with higher-probability words being read faster than lower-probability words (McDonald & Shillcock, 2003a, 2003b). In the years since, evidence has continued to mount that bigram surprisal (negative log-probability; see Hale, 2001) is a reliable predictor of reading time, as is surprisal from higher-order *n*-grams where probability is calculated based on the $n-1$ previous words rather than the just the previous word (Boston et al., 2008; Brothers & Kuperberg, 2021; Demberg & Keller, 2008; Fossum & Levy, 2012; Frank & Bod, 2011; Levy, 2008; Mitchell et al., 2010; Oh et al., 2022; Smith & Levy, 2008, 2011, 2013). As natural language technology has advanced, researchers have increasingly turned to more complex language models to calculate statistical probability, including recurrent neural networks (Aurnhammer & Frank, 2019a, 2019b; Brothers & Kuperberg, 2021; Eisape et al., 2020; Fossum & Levy, 2012; Frank & Bod, 2011; Goodkind & Bicknell, 2018; Hao et al., 2020; Kuribayashi et al., 2021; Merkx & Frank, 2021; Monsalve et al., 2012; Oh et al., 2022; Wilcox et al., 2020) and more recently, transformers (Eisape et al., 2020; Hao et al., 2020; Kuribayashi et al., 2021; Meister et al., 2021; Merkx & Frank, 2021; Oh & Schuler, 2023a, 2023b, 2025; Oh et al., 2022, 2024, 2025; Opedal et al., 2024; Shain, 2024; Shain et al., 2024; Wilcox, Meister, et al., 2023; Wilcox, Pimentel, Meister, Cotterell, & Levy, 2023; Wilcox et al., 2020).

An initial trend reported in this line of work was that more powerful models—those with more parameters, those with more advanced or scalable architectures, those trained on larger corpora, and those with better next-word prediction capabilities—produce predictions that show the best fit to human reading times for naturalistic text (Goodkind & Bicknell, 2018; Wilcox, Meister, et al., 2023; Wilcox et al., 2020). For example, Goodkind and Bicknell (2018) observed a linear relationship between a language model's perplexity (a measure of next-word prediction capability; Jelinek et al., 1977) and the fit between its predictions and reading time data.

However, this pattern has not remained robust (Eisape et al., 2020; Hao et al., 2020; Merkx & Frank, 2021), and the evidence has begun to accumulate for the existence of a limit to the extent to which this is the case: contemporary transformer language models can become so good at next-word prediction that the surprisals they assign to words become less correlated with naturalistic reading time (Kuribayashi et al., 2021; Oh & Schuler, 2023a, 2023b; Oh et al., 2022, 2024; Shain et al., 2024). This *inverse scaling* effect does not appear to be determined by any specific factor that makes a given model better at next-word prediction; it applies to cases where larger models (i.e., those with a greater number of parameters) are better at next-word prediction than smaller models (Oh et al., 2022, 2024), cases where models trained on larger datasets are better at next-word prediction than models trained on smaller datasets (Oh & Schuler, 2023b), and possibly cases where some architectures tend to perform better at next-word prediction than others (Eisape et al., 2020).

Other limitations of relying solely on neural language models to calculate surprisal have also emerged. Unigram surprisal (i.e., negative log-frequency) is often a significant predictor of reading time above and beyond language model surprisal (Shain, 2024), despite the latter being statistical models of language probability. The same may also be true for bigram and trigram surprisal (Goodkind & Bicknell, 2021). In fact, the extent to which adding unigram and bigram surprisal as predictors improves the fit of a regression predicting reading time increases as models become better next-word predictors, which has been interpreted as 'compensation' for the increasingly worse fit of neural language model surprisal (Oh & Schuler, 2025). In addition, much of the effect of surprisal on reading time appears to be explainable by frequency alone (Opedal et al., 2024). Thus, the evidence suggests that contemporary language model surprisal does not fully capture the effects of statistical probability on reading time, and as models become larger and more powerful, their surprisals do so less.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

Where does the inverse scaling come from?

Oh and Schuler (2023b) find that using surprisal from language models trained on up to 2 billion words leads to predictions that better fit naturalistic reading time data, but surprisal from models trained beyond this point shows an increasingly worse fit to the data. This pattern is also observable when comparing models according to their perplexity—models appear to become 'too good' at next-word prediction for their surprisals to fit reading time well (Oh & Schuler, 2023b). One specific observed phenomenon is that more powerful models assign relatively lower surprisals to low-frequency words (especially nouns and adjectives; Oh and Schuler, 2023a) than do less powerful models, and this in turn can lead to a systematic under-estimation of these words' reading times (Oh et al., 2024). This may also partially explain why frequency has been found to explain variance in reading time above and beyond surprisal (Oh & Schuler, 2025; Shain, 2024). However, exactly what underlies the decreasing fit between language model predictions and human language processing as indexed by reading time is still not fully understood.

A general explanation for these patterns is that during the earlier phases of training, language models make increasingly human-like predictions as they initially begin to learn language statistics, but as they continue to train, their predictions increasingly reflect properties of the data and their own inductive biases. One version of this is the aforementioned possibility that models simply become 'too good' at next-word prediction overall due to being trained on orders of magnitude more data than humans are exposed to, thereby being able to predict low-frequency words when humans may not be (Oh et al., 2024). But data type could also matter. For example, during early language acquisition, our input is primarily made up of speech, and often child-directed speech. by contrast, language models are trained on 'high-quality' text data, selected based on specific criteria that are not aligned with this (see, e.g., Wilcox et al., 2025). It is also possible that even if the models were trained on exactly the same data that the average human were exposed to, the predictions of humans and language models would begin to diverge as the models better learned to reflect the empirical distribution—previous work suggests that human predictions may not always exactly reflect empirical probabilities, potentially due to our own inductive biases or the augmentation of predictions with other kinds of knowledge (see, e.g., Brothers & Kuperberg, 2021; Michaelov & Bergen, 2024; Michaelov et al., 2022; Smith & Levy, 2011). Finally, different model architectures are more or less cognitively plausible in a range of ways, which could in principle impact the extent to which their predictions match our own (Merkx & Frank, 2021; Michaelov et al., 2021; Ryu & Lewis, 2021), though no impact of this on the inverse scaling effect has been observed (Michaelov et al., 2024).

The n-gram hypothesis

While the aforementioned possibilities are plausible and likely worthwhile to explore further, in this paper we propose a different explanation, which we empirically evaluate on eye-tracking data: **naturalistic reading time is shaped by lower-order n-gram statistics, and thus, language models that make the most n-gram-like predictions show the best fit to the data**.

This would account for the aforementioned inverse scaling patterns observed in previous work, aligns with older work using n-gram surprisal to predict reading time, and is supported by several additional findings. First, researchers have often found that operations that reduce the role of context—including restricting the context window (Kuribayashi et al., 2022), reducing memory capacity (Timkey & Linzen, 2023), or adding a recency bias (C. Clark et al., 2025; De Varda & Marelli, 2024; Goodkind & Bicknell, 2021; Vaidya et al., 2023)—improve language model predictions' fit to reading time. In addition, language models of all architectures tend to show a consistent training trajectory where their predictions closely match unigram probability, followed by bigram probability, trigram probability, and so on (Chang et al., 2024; Karpathy et al., 2016; Michaelov et al., 2025). When Oh and Schuler (2023b) looked at the Pythia suite of language models (Biderman et al., 2023), they found that fit to reading time increases until step 1000 of training (2 billion training tokens), at which point the fit begins to decrease. More recently, Michaelov et al. (2025) found that step 1000 is around the point during training at which the Pythia models' predictions' correlations to bigram and trigram probabilities peak; after which their correlation drops, as does the correlation with unigram probability. This is exactly what one would expect if reading time aligns with these n-gram statistics.

Why n-grams? Our proposal is derived from considering the time course of the neurocognitive mechanisms at play during naturalistic reading. Current research on the neuroscience of language comprehension suggests that the N400 component of the event-related brain potential (ERP), which reliably occurs 300-500ms after a stimulus is first encountered, indexes the activation of a word's representations in the brain (Aurnhammer et al., 2021; Brouwer et al., 2012; Federmeier, 2021; Kuperberg et al., 2020; Kutas & Federmeier, 2011), and is followed by a number of positive ERP components that have been associated with integration, reanalysis, or anomaly detection and correction (Aurnhammer et al., 2021; Brouwer et al., 2012; DeLong & Kutas, 2020; Kuperberg et al., 2020). But in naturalistic reading, the average time spent on a word during the first pass (i.e., first pass/gaze duration) is generally under 250ms (Berzak et al., 2025; Cop et al., 2017; Kennedy et al., 2013). Taking the lower bound of estimates about how long is needed to program the motor actions of a saccade (110-150ms; Becker and Jürgens, 1979; Reichle et al., 2012) and the upper bound of estimates of how long a saccade takes to complete on average during natural reading (20-40ms; Rayner and Morrison, 1981; Reichle et al., 2012), if the first pass duration of a word is 250ms, this would put the initiation of the programming of the saccade from word w_i to w_{i+1} at a maximum of 450ms after word w_{i-1} was first

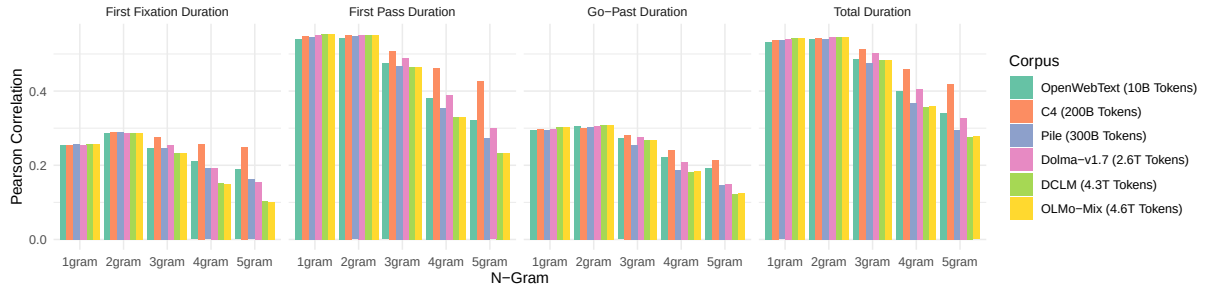

Figure 1: The correlation between n-gram surprisal and reading time in the Provo Corpus.

fixated; which is during the N400, and well before any of the later processes have even begun. Thus, it is hard to see how the first pass duration of the average word w_i could be shaped by a fully-integrated representation of $w_1 \ldots w_{i-1}$. To the extent that reading time is shaped by predictions about upcoming words, these predictions must, at least some of the time, be based on shallower, under-specified, or otherwise incomplete representations of their context.

One strong potential candidate for this is n-gram prediction. Prediction based on n-grams relies only on having identified the previous words in the sequence. Under a hierarchical predictive coding model of language comprehension (e.g., Bornkessel-Schlesewsky & Schlesewsky, 2019; Kuperberg et al., 2020; Lewis & Bastiaansen, 2015; Nour Eddine et al., 2024), for example, this could be accomplished relatively early by the lower levels of the predictive system that are associated with word form at the same time as the N400 and later ERP components begin to unfold at higher levels. And the fact that recurrent neural networks, transformers, and other modern language models, all of which are explicitly optimized for next-word prediction also all learn n-gram relations (Akyürek et al., 2024; Bietti et al., 2023; Chang & Bergen, 2025; Chang et al., 2024; Choshen et al., 2022; Karpathy et al., 2016; Michaelov et al., 2025; Sun & Lu, 2022; Voita et al., 2024) may be taken to suggest that it is rational for any predictive system to learn such patterns. It is therefore plausible that n-gram and potentially other shallow statistics could account for many of the predictability effects observed in naturalistic reading.

It is worth noting that the n-gram hypothesis is not intended to be an account of *all* contextual or probabilistic reading time effects. Longer-range dependencies are known to shape processing (Staub & Clifton, 2006); and indeed, our account does not preclude the possible impact of rich contextual information from $w_{<i-1}$. But since natural texts—which are the context in which inverse scaling effects have been observed—are generally written to be coherent rather than explicitly oriented around known loci of processing difficulty (as in the case of psycholinguistic stimuli), it is likely that this does not have as much of a measurable impact above and beyond surface-level statistics operationalized by n-grams. It also worth noting that our account is primarily focused on first pass duration, but the inverse scaling effect has also been observed for go-past duration (Oh & Schuler, 2023a, 2023b, 2025; Oh et al., 2022, 2024, 2025). But since go-past duration is largely made up of first pass duration, it would be surprising if the former showed a different scaling pattern unless there were some specific reason to believe that surprisal from more powerful language models would be sufficiently better at predicting the time spent on regressions out of words to counteract the worse prediction of first pass duration. We are not aware of any direct evidence for such a prediction; and in fact, what causes regressions is still an open research question (T. H. Clark et al., 2026; Rego et al., 2026; Wilcox et al., 2024).

Experiment 1

If the n-gram hypothesis holds, we should expect the relationship between n-gram surprisal and reading time to show a different pattern to that of neural language models. Crucially, if reading time reflects lower-order n-gram relations, then these should not show the same inverse scaling. That is, the correlation between lower-order n-gram surprisal and reading time should not get worse no matter how large the corpus—if anything, it should improve. Meanwhile, we might expect to see a decrease in correlation as n-gram order increases. This is not the pattern observed in past work (Goodkind & Bicknell, 2018; Hao et al., 2020), but it is worth noting that the n-grams in these studies relied on corpora of roughly 1B tokens, which is before inverse scaling emerges in transformers (Oh & Schuler, 2023b). We carry out this analysis with contemporary corpora ranging in size from tens of billions to trillions of tokens.

Method

n-grams We calculate the surprisal (negative log-probability) of words based on their n-gram probability in six corpora: OpenWebText (Gokaslan & Cohen, 2019), C4 (Raffel et al., 2020), the Pile (Gao et al., 2020), Dolma (Soldaini et al., 2024), DCLM (Li et al., 2024), and OLMo-Mix (OLMo Team et al., 2025). Following Michaelov et al. (2025), we calculate n-gram probability using the raw counts of word sequences in the corpus as calculated by *infini-gram* (Liu et al., 2024), and use the Stupid Backoff (Brants et al., 2007) procedure to convert these counts into probabilities. For Open-

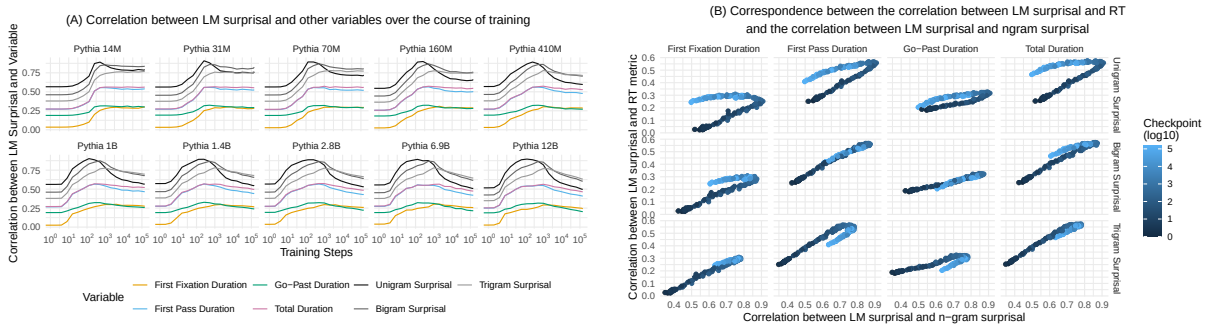

Figure 2: (A) The relationship between language model surprisal over the course of training and both n-gram surprisal and reading time in the Provo Corpus. (B) The relationship between the two sets of correlations.

WebText n-grams, we build an infini-gram index locally; for the other corpora, we rely on the infini-gram API. We provide all data, code and analysis scripts at https://osf.io/msuwk.

Reading Time We carry out our analysis on the Provo Corpus (Luke & Christianson, 2018), an eye-tracking dataset comprised of data from 470 participants reading 55 passages. We consider four reading time measures: First Fixation Duration, First Pass Duration (duration from when a word is first fixated to when it is first left), Go-Past Duration (duration from when a word is first fixated to the first fixation *past* the word), and Total Duration (the sum of all fixations on the word). We exclude words that were skipped on the first pass, as well as words at the beginning of end of a sentence (see, e.g., Oh and Schuler, 2023b; Shain et al., 2024). We then calculate the mean of each metric for each word in the corpus (2,449 after the aforementioned exclusions). Of the aforementioned metrics, we are only aware of one previous study investigating scaling effects on the prediction of first fixation duration and total duration (Hao et al., 2020).

Results

The results of this experiment are presented in Figure 1. The first clear pattern we observe is that unigram and bigram surprisal tend to be most highly correlated with reading time with little difference between the two (the one exception to this is First Fixation Duration, where bigram surprisal has a substantially higher correlation). Each n-gram beyond this (i.e., for $n \geq 3$) shows a progressively lower correlation to each reading time metric, and in addition, for these n-grams, there is a general tendency for n-grams based on larger corpora to show a lower correlation. Finally, we see corpus-specific effects. Specifically, we see that C4 n-grams, and to a lesser extent Dolma n-grams, are more highly correlated with reading time than the two aforementioned patterns would predict.

Discussion

We find that, across the board, lower-order n-grams ($n \in \{1, 2\}$) show a stronger correlation to reading time metrics than higher-order n-grams ($n \in \{3, 4, 5\}$), consistent with our predictions. While we do observe some inverse scaling as a function of dataset size, this only occurs for the higher-order n-grams, and so may simply be due to these larger datasets allowing for better estimation of these higher-order n-grams. For the lower-order n-grams, if anything, there is a very small but noticeable degree of positive scaling as a function of corpus size.

Additionally, we observe that there is substantial and apparently idiosyncratic variation between n-grams calculated on different corpora, even after accounting for corpus size. One possible explanation for this is text domain or genre. For example, the fact that C4 was constructed from cleaned Common Crawl data, which includes data from the whole internet may mean that it has a more naturalistic distribution than the others, which are designed explicitly to be 'high-quality' data that lead to good language model performance—this includes the intentional inclusion of books and scientific articles, only collecting webpages that have been shared by users on websites such as Reddit, and aggressive data filtering procedures.

Experiment 2

We next turn to our core question of whether the extent to which scaling effects observed in the relationship between language models and reading time can be explained by the extent to which language model surprisal resembles that of lower-order n-grams, as would be expected if reading time is sensitive to n-gram probability. First, we attempt to replicate the results of Oh and Schuler (2023b) and Michaelov et al. (2025), testing whether it is indeed the case that the correlation between language model surprisal and reading time peaks when the model's predictions most reflect n-grams.

Method

We use the same reading time metrics as in Experiment 1. Given our aim to explain results such as those in Oh and Schuler (2023b), we use the Pythia models (Biderman et al., 2023) to calculate neural language model surprisal. These are a set of 10 autoregressive transformer language models ranging in size from 14 million parameters to 12 billion parameters, and trained on the 300 billion tokens of the aforementioned Pile text corpus, with checkpoints provided at various points

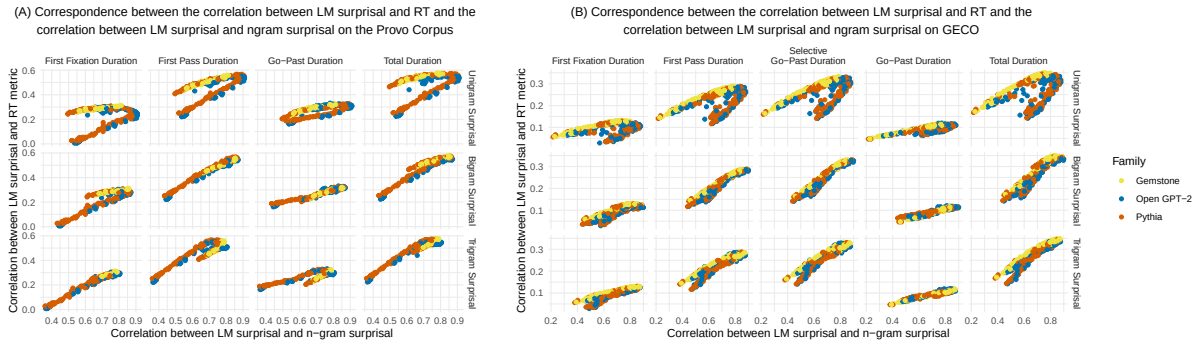

Figure 3: The relationship between the correlation between language model surprisal and n-gram surprisal and the correlation between language model surprisal and each measure of reading time in the (A) Provo and (B) GECO datasets.

over the course of training. Thus, it is possible to investigate how language models surprisal correlates with reading time both as a function of parameters and training tokens. Since the models are trained on the Pile, we use the unigram, bigram, and trigram surprisals calculated on the Pile in Experiment 1.

Results

First, we compare the trajectory of the correlation between Pythia surprisal and reading time to that of the correlation between Pythia surprisal and n-gram surprisal, which we present in Figure 2A. The degree to which First Fixation Duration is correlated with Pythia surprisal shows a highly similar time course to trigram surprisal, while the time course for the other metrics of reading time more closely resemble those of unigram and bigram surprisal.

To further compare the synchronicity between correlation to these variables, we look at the correlation between these correlations (Figure 2B). As can be seen, in almost all cases, the highest correlation to reading time occurs at the point at which surprisal is most correlated with n-gram surprisal. The nonlinearities in the correlations (i.e., the "V" shapes) show that as models continue to train, their surprisal's correlation with reading time drops more slowly than the correlation with unigram surprisal (relative to the initial increase), a pattern that reduces and even reverses for higher-order n-gram surprisal. Thus, relationship between model surprisal's correlation to bigram surprisal and to First Pass Duration and Go-Past Duration is relatively linear, and the same is true for First Fixation Duration and Total Duration in their relation to trigram surprisal. This pattern is borne out when we calculate the correlation between these correlations (see Table 1).

Discussion

Overall, we see a general trend where the correlation between language model surprisal and First Pass Duration and Go-Past Duration is highly correlated with the correlation between language model surprisal and bigram surprisal. Meanwhile, for First Fixation Duration and Total Duration, we see a closer relationship with trigram surprisal. The former is surprising in light of the pattern that in the case of neural indices of processing, later responses appear to correlate with more complex relations between words and their context (DeLong & Kutas, 2020; Kuperberg et al., 2020; Van Petten & Luka, 2012); though it is worth noting the generally low overall correlation between all surprisal metrics and First Fixation Duration.

Experiment 3

Our results show an extremely high correlation between the extent to which language model surprisal correlates with reading time and the extent to which it correlates with n-gram surprisal. However, in the previous experiment, we only look at one set of language models (Pythia) and one reading time dataset (the Provo Corpus). In this experiment, we test the reliability of the results of Experiment 2 by replicating the analysis with more models on more reading time data.

Method

We include all the reading time data from Experiments 1–2, as well as the Ghent Eye-Tracking Corpus (GECO; Cop et al., 2017), a dataset made up of 14 experimental participants reading a novel. This dataset includes an additional metric known as Selective Go-Past Duration (SGPD), a variant of go-past duration that only includes the time spent fixated on the word itself. After carrying out the same exclusions and calculating mean reading times, we carried out our analysis on data from 43,251 words.

In addition to calculating neural language model surprisal with the Pythia models (as in Experiment 2), we calculate the surprisal of all words in both reading time datasets using the Open GPT-2 (Karamcheti et al., 2021) and Gemstone (McLeish et al., 2025) models. The Open GPT-2 models are trained on the OpenWebText corpus, and are an open replication of the small (124M parameter) and medium (355M parameter) GPT-2 models, with checkpoints taken over the course of training during 5 training runs (with different random seeds). Due to an error with one of these seeds (Hawkins, 2023), we use the remaining 4 seeds of each model. The Gemstone models are a set of 22 models ranging from 50 million to 2 billion parameters, designed to for testing whether the width and depth of language models have an impact on performance

Table 1: The Pearson correlation coefficient r between the correlation between language model surprisal and n-gram surprisal and the correlation between language model surprisal and First Fixation Duration (FFD), First Pass Duration (FPD), Go-Past Duration (GPD), and Total Duration (TD); with the highest bolded.

		GECO					Provo			
Family	**N-gram**	**FFD**	**FPD**	**SGPD**	**GPD**	**TD**	**FFD**	**FPD**	**GPD**	**TD**
Gemstone	Unigram Surprisal	0.964	0.981	**0.984**	**0.990**	0.980	0.884	0.980	0.955	0.949
Gemstone	Bigram Surprisal	0.942	**0.986**	**0.984**	0.970	0.966	**0.896**	**0.986**	**0.972**	**0.966**
Gemstone	Trigram Surprisal	**0.976**	0.970	0.974	0.982	**0.982**	0.886	0.959	0.969	0.962
Open GPT-2	Unigram Surprisal	0.564	0.643	0.662	0.771	0.646	0.511	0.772	0.723	0.677
Open GPT-2	Bigram Surprisal	0.918	**0.974**	**0.976**	0.938	0.964	0.955	**0.989**	**0.973**	**0.988**
Open GPT-2	Trigram Surprisal	**0.95**	0.945	0.953	**0.969**	**0.966**	**0.991**	0.918	0.908	0.963
Pythia	Unigram Surprisal	0.238	0.377	0.409	0.671	0.355	0.478	0.743	0.808	0.622
Pythia	Bigram Surprisal	0.878	**0.967**	**0.971**	0.933	0.945	0.925	**0.993**	**0.984**	0.973
Pythia	Trigram Surprisal	**0.931**	0.957	0.964	**0.949**	**0.967**	**0.993**	0.952	0.899	**0.986**

above and beyond their contribution to total parameter count. They are trained on a subset of the Dolma corpus, with checkpoints taken over the course of training.

Results

The results are presented in Figure 3. Across models, metrics, and datasets, we see a general trend: a language model that makes predictions that are more correlated with n-grams also makes predictions that are more correlated with reading time. As in Experiment 2, this is generally true to the greatest extent for bigram and trigram surprisal (see Table 1). Also as in Experiments 1–2, we see the lowest overall correlation between surprisal (neural language model and n-gram) and First Fixation Duration and Go-Past Duration—in GECO, this is more pronounced for the latter. A notable detail in Figure 3 is that we see mostly the same pattern across model families—the relationship between the correlation to n-grams and to reading time appears stable.

Discussion

We observe the same general pattern as in Experiment 2, that language models that make more n-gram-like predictions also make predictions that are more strongly correlated with each of the reading time metrics.

General Discussion

Our results provide clear evidence that the extent to which a neural language model's predictions reflect low-order n-grams is correlated with the extent to which the surprisal of words calculated using it correlates with reading time. This is true across reading time metric, dataset, and language model family. Thus, the evidence is consistent with our original hypothesis—more n-gram-like language models show a better correlation to naturalistic reading time.

What does this tell us about human language processing? As previously noted, we do not argue for n-gram-based prediction as a complete account of how context shapes reading time. Nor do we argue that humans do not rely on more complex statistical relationships between words—indeed, recent work has shown that more powerful language models trained on more data other indices of language processing better than models trained on less data (Michaelov & Bergen, 2026; Michaelov et al., 2024). Nonetheless, we clearly observe that more n-gram-like language models make predictions that correlate better with reading time, and note that a substantial amount of previous empirical work has demonstrated a reliable link between reading time and n-gram surprisal, as previously discussed. Thus, taken together, our results suggest that to the extent that language statistics play a role in the reading of naturalistic text, these are best captured by n-grams.

The idea that first pass duration can explain the pattern in other metrics is also supported by the data analyzed in that most words are only fixated once in the first pass (Provo: 82.5%; GECO: 88.7%), and most words are not refixated after the first pass (Provo: 82.5%; GECO: 83.4%). The lower overall correlation for First Fixation Duration is likely due to the fact that in cases where there is a difference between it and First Pass Duration, it is smaller, and thus, proportionally more impacted by spillover effects from previous words. Similarly, Go-Past Duration includes the time taken to read other words in regressions from the current word.

Another question is why we see slightly higher correlations between neural language models and reading time (at their best) than between raw n-grams and reading time. One possibility is that neural language models simply learn a better smoothing function (recall that we use a very simple backoff scheme). Alternatively, it is possible that eye movements are sensitive to other basic properties of words that are captured by weaker neural language models but not n-grams, such as word-level semantic relatedness (see Michaelov et al., 2025). Disentangling these possibilities is a question for future work.

Acknowledgments

For their discussion and helpful comments, we would like to thank the members of the Computational Psycholinguistics Laboratory at MIT, Benjamin Bergen, and the attendees of the 39th Annual Conference on Human Sentence Processing (HSP 2026). We would also like to thank Catherine Arnett for feedback on an earlier version of this manuscript, as well as the anonymous reviewers at HSP and CogSci 2026. James Michaelov was supported by a grant from the Andrew W. Mellon foundation (#2210-13947) during the writing of this paper.

References

Akyürek, E., Wang, B., Kim, Y., & Andreas, J. (2024). In-Context Language Learning: Architectures and Algorithms. *Proceedings of the 41st International Conference on Machine Learning*, 787–812. Retrieved October 21, 2025, from https://proceedings.mlr.press/v235/akyurek24a.html

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., & Crocker, M. W. (2021). Retrieval (N400) and integration (P600) in expectation-based comprehension. *PLOS ONE*, *16*(9), e0257430. https://doi.org/10.1371/journal.pone.0257430

Aurnhammer, C., & Frank, S. L. (2019a). Comparing Gated and Simple Recurrent Neural Network Architectures as Models of Human Sentence Processing [tex.howpublished: https://escholarship.org/uc/item/0br7f339]. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci 2019)*. Retrieved June 2, 2022, from https://escholarship.org/uc/item/0br7f339

Aurnhammer, C., & Frank, S. L. (2019b). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, *134*, 107198. https://doi.org/10.1016/j.neuropsychologia.2019.107198

Becker, W., & Jürgens, R. (1979). An analysis of the saccadic system by means of double step stimuli. *Vision Research*, *19*(9), 967–983. https://doi.org/10.1016/0042-6989(79)90222-0

Berzak, Y., Malmaud, J., Shubi, O., Meiri, Y., Lion, E., & Levy, R. (2025). OneStop: A 360-Participant English Eye Tracking Dataset with Different Reading Regimes. *Scientific Data*, *12*(1), 1995. https://doi.org/10.1038/s41597-025-06272-2

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., & Wal, O. V. D. (2023). Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *Proceedings of the 40th International Conference on Machine Learning*, 2397–2430. Retrieved October 19, 2023, from https://proceedings.mlr.press/v202/biderman23a.html

Bietti, A., Cabannes, V., Bouchacourt, D., Jegou, H., & Bottou, L. (2023). Birth of a Transformer: A Memory Viewpoint. *Advances in Neural Information Processing Systems*, *36*, 1560–1588. Retrieved July 29, 2025, from https://papers.nips.cc/paper_files/paper/2023/hash/0561738a239a995c8cd2ef0e50cfa4fd-Abstract-Conference.html

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2019). Toward a Neurobiologically Plausible Model of Language-Related, Negative Event-Related Potentials. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00298

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1). https://doi.org/10.16910/jemr.2.1.1

Brants, T., Popat, A. C., Xu, P., Och, F. J., & Dean, J. (2007, June). Large Language Models in Machine Translation. In J. Eisner (Ed.), *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 858–867). Association for Computational Linguistics. Retrieved May 13, 2024, from https://aclanthology.org/D07-1090

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, *116*, 104174. https://doi.org/10.1016/j.jml.2020.104174

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143. https://doi.org/10.1016/j.brainres.2012.01.055

Chang, T. A., & Bergen, B. (2025). Bigram Subnetworks: Mapping to Next Tokens in Transformer Language Models. Retrieved May 1, 2026, from https://openreview.net/forum?id=0TD3eO46gk

Chang, T. A., Tu, Z., & Bergen, B. K. (2024). Characterizing Learning Curves During Language Model Pre-Training: Learning, Forgetting, and Stability [Place: Cambridge, MA]. *Transactions of the Association for Computational Linguistics*, *12*, 1346–1362. https://doi.org/10.1162/tacl_a_00708

Choshen, L., Hacohen, G., Weinshall, D., & Abend, O. (2022, May). The Grammar-Learning Trajectories of Neural Language Models. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8281–8297). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.568

Clark, C., Oh, B.-D., & Schuler, W. (2025, January). Linear Recency Bias During Training Improves Transformers' Fit to Reading Times. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 7735–7747). Association for Computational Linguistics. Retrieved January 26, 2026, from https://aclanthology.org/2025.coling-main.517/

Clark, T. H., Levy, R. P., & Gibson, E. (2026). Readers make targeted regressions to plausible errors in reanalysis of "noisy-channel garden-path" sentences. *Proceedings of the 30th Conference on Computational Natural Language Learning*.

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, *49*(2), 602–615. https://doi.org/10.3758/s13428-016-0734-0

De Varda, A., & Marelli, M. (2024, August). Locally Biased Transformers Better Align with Human Reading Times. In T. Kuribayashi, G. Rambelli, E. Takmaz, P. Wicke, & Y. Oseki (Eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 30–36). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.cmcl-1.3

DeLong, K. A., & Kutas, M. (2020). Comprehending surprising sentences: Sensitivity of post-N400 positivities to contextual congruity and semantic relatedness. *Language, Cognition and Neuroscience*, *35*(0), 1044–1063. https://doi.org/10.1080/23273798.2019.1708960

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. https://doi.org/10.1016/j.cognition.2008.07.008

Eisape, T., Zaslavsky, N., & Levy, R. (2020). Cloze Distillation: Improving Neural Language Models with Human Next-Word Prediction. *Proceedings of the 24th Conference on Computational Natural Language Learning*, 609–619. https://doi.org/10.18653/v1/2020.conll-1.49

Federmeier, K. D. (2021). Connecting and considering: Electrophysiology provides insights into comprehension [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/psyp.13940]. *Psychophysiology*, *59*, e13940. https://doi.org/10.1111/psyp.13940

Fossum, V., & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics (CMCL 2012)*, 61–69.

Frank, S. L., & Bod, R. (2011). Insensitivity of the Human Sentence-Processing System to Hierarchical Structure. *Psychological Science*, *22*(6), 829–834. https://doi.org/10.1177/0956797611409589

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2020). The Pile: An 800GB Dataset of Diverse Text for Language Modeling [arXiv:2101.00027 [cs]]. Retrieved February 3, 2022, from https://arxiv.org/abs/2101.00027

Gokaslan, A., & Cohen, V. (2019). OpenWebText corpus [http://Skylion007.github.io/OpenWebTextCorpus]. http://Skylion007.github.io/OpenWebTextCorpus

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, 10–18. https://doi.org/10.18653/v1/W18-0102

Goodkind, A., & Bicknell, K. (2021, March). Local word statistics affect reading times independently of surprisal [arXiv:2103.04469 [cs]]. https://doi.org/10.48550/arXiv.2103.04469

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 - NAACL '01*, 1–8. https://doi.org/10.3115/1073336.1073357

Hao, Y., Mendelsohn, S., Sterneck, R., Martinez, R., & Frank, R. (2020). Probabilistic Predictions of People Perusing: Evaluating Metrics of Language Model Performance for Psycholinguistic Modeling. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 75–86. https://doi.org/10.18653/v1/2020.cmcl-1.10

Hawkins, R. D. (2023, July). Arwen is checkpoint progression outlier? [https://github.com/stanford-crfm/mistral/issues/199]. Retrieved January 19, 2025, from https://github.com/stanford-crfm/mistral/issues/199

Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, *62*(S1), S63. https://doi.org/10.1121/1.2016299

Karamcheti, S., Orr, L., Bolton, J., Zhang, T., Goel, K., Narayan, A., Bommasani, R., Narayanan, D., Hashimoto, T., Jurafsky, D., et al. (2021). Mistral — a journey towards reproducible language model training [Stanford Center for Research on Foundation Models: https://crfm.stanford.edu/2021/08/26/mistral.html]. https://crfm.stanford.edu/2021/08/26/mistral.html

Karpathy, A., Johnson, J., & Fei-Fei, L. (2016). Visualizing and Understanding Recurrent Networks. *International Conference on Learning Representations (Workshop Track)*. Retrieved January 19, 2025, from https://openreview.net/forum?id=71BmK0m6qfAE8VvKUQWB

Kennedy, A., Pynte, J., Murray, W. S., & Paul, S.-A. (2013). Frequency and predictability effects in the Dundee Corpus: An eye movement analysis. *Quarterly Journal of Experimental Psychology*, *66*(3), 601–618. https://doi.org/10.1080/17470218.2012.676054

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A Tale of Two Positivities and the N400: Distinct Neural Signatures Are Evoked by Confirmed and Violated Predictions at Different Levels of Representation. *Journal of Cognitive Neuroscience*, *32*(1), 12–35. https://doi.org/10.1162/jocn_a_01465

Kuribayashi, T., Oseki, Y., Brassard, A., & Inui, K. (2022, December). Context Limitations Make Neural Language Models More Human-Like. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 10421–10436). Association for Computational Lin-

guistics. https://doi.org/10.18653/v1/2022.emnlp-main.712

Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., & Inui, K. (2021). Lower Perplexity is Not Always Human-Like. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5203–5217. https://doi.org/10.18653/v1/2021.acl-long.405

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Levy, R. P. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Lewis, A. G., & Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex*, *68*, 155–168. https://doi.org/10.1016/j.cortex.2015.02.014

Li, J., Fang, A., Smyrnis, G., Ivgi, M., Jordan, M., Gadre, S. Y., Bansal, H., Guha, E. K., Keh, S., Arora, K., Garg, S., Xin, R., Muennighoff, N., Heckel, R., Mercat, J., Chen, M. F., Gururangan, S., Wortsman, M., Albalak, A., … Shankar, V. (2024). DataComp-LM: In search of the next generation of training sets for language models. Retrieved February 2, 2026, from https://openreview.net/forum?id=CNWdWn47IE#discussion

Liu, J., Min, S., Zettlemoyer, L., Choi, Y., & Hajishirzi, H. (2024). Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens. *First Conference on Language Modeling*. Retrieved October 23, 2025, from https://openreview.net/forum?id=u2vAyMeLMm

Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, *50*(2), 826–833. https://doi.org/10.3758/s13428-017-0908-4

McDonald, S. A., & Shillcock, R. C. (2003a). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, *43*(16), 1735–1751. https://doi.org/10.1016/S0042-6989(03)00237-2

McDonald, S. A., & Shillcock, R. C. (2003b). Eye Movements Reveal the On-Line Computation of Lexical Probabilities During Reading. *Psychological Science*, *14*(6), 648–652. https://doi.org/10.1046/j.0956-7976.2003.psci_1480.x

McLeish, S. M., Kirchenbauer, J., Miller, D. Y., Singh, S., Bhatele, A., Goldblum, M., Panda, A., & Goldstein, T. (2025). Gemstones: A Model Suite for Multi-Faceted Scaling Laws. Retrieved February 2, 2026, from https://openreview.net/forum?id=iZk78dZ1Ap

Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021). Revisiting the Uniform Information Density Hypothesis. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 963–980. https://doi.org/10.18653/v1/2021.emnlp-main.74

Merkx, D., & Frank, S. L. (2021). Human Sentence Processing: Recurrence or Attention? *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 12–22. https://doi.org/10.18653/v1/2021.cmcl-1.2

Michaelov, J. A., Arnett, C., & Bergen, B. K. (2024). Revenge of the Fallen? Recurrent Models Match Transformers at Predicting Human Language Comprehension Metrics. Retrieved September 12, 2025, from https://openreview.net/forum?id=amhPBLFYWv

Michaelov, J. A., Bardolph, M. D., Coulson, S., & Bergen, B. K. (2021). Different kinds of cognitive plausibility: Why are transformers better than RNNs at predicting N400 amplitude? *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, 300–306.

Michaelov, J. A., & Bergen, B. K. (2024). On the Mathematical Relationship Between Contextual Probability and N400 Amplitude. *Open Mind*, *8*, 859–897. https://doi.org/10.1162/opmi_a_00150

Michaelov, J. A., & Bergen, B. K. (2026). *Better Language Models Better Model the N400, but not Reading Time* [Manuscript accepted with minor revisions at the Journal of Memory and Language].

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2022). So Cloze yet so Far: N400 Amplitude is Better Predicted by Distributional Information than Human Predictability Judgements [Conference Name: IEEE Transactions on Cognitive and Developmental Systems]. *IEEE Transactions on Cognitive and Developmental Systems*. https://doi.org/10.1109/TCDS.2022.3176783

Michaelov, J. A., Levy, R. P., & Bergen, B. K. (2025). Language Model Behavioral Phases are Consistent Across Architecture, Training Data, and Scale. *Advances in Neural Information Processing Systems*, *38*.

Mitchell, J., Lapata, M., Demberg, V., & Keller, F. (2010). Syntactic and semantic factors in processing difficulty: An integrated measure. *Proceedings of the 48th annual meeting of the association for computational linguistics*, 196–206.

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 398–408.

Nour Eddine, S., Brothers, T., Wang, L., Spratling, M., & Kuperberg, G. R. (2024). A predictive coding model of the N400. *Cognition*, *246*, 105755. https://doi.org/10.1016/j.cognition.2024.105755

Oh, B.-D., Clark, C., & Schuler, W. (2022). Comparison of Structural Parsers and Neural Language Models as Surprisal Estimators. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.777963

Oh, B.-D., & Schuler, W. (2023a). 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. https://doi.org/10.1162/tacl_a_00548

Oh, B.-D., & Schuler, W. (2023b, December). Transformer-Based Language Model Surprisal Predicts Human Reading Times Best with About Two Billion Training Tokens. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1915–1921). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.128

Oh, B.-D., & Schuler, W. (2025). Dissociable frequency effects attenuate as large language model surprisal predictors improve. *Journal of Memory and Language*, *143*, 104645. https://doi.org/10.1016/j.jml.2025.104645

Oh, B.-D., Yue, S., & Schuler, W. (2024, March). Frequency Explains the Inverse Correlation of Large Language Models' Size, Training Data Amount, and Surprisal's Fit to Reading Times. In Y. Graham & M. Purver (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2644–2663). Association for Computational Linguistics. Retrieved May 31, 2024, from https://aclanthology.org/2024.eacl-long.162

Oh, B.-D., Zhu, H., & Schuler, W. (2025, July). The Inverse Scaling Effect of Pre-Trained Language Model Surprisal Is Not Due to Data Leakage. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 1820–1827). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.91

OLMo Team, Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., Lambert, N., Schwenk, D., Tafjord, O., Anderson, T., Atkinson, D., Brahman, F., Clark, C., Dasigi, P., Dziri, N., … Hajishirzi, H. (2025, January). 2 OLMo 2 Furious [arXiv:2501.00656 [cs]]. https://doi.org/10.48550/arXiv.2501.00656

Opedal, A., Chodroff, E., Cotterell, R., & Wilcox, E. (2024, November). On the Role of Context in Reading Time Prediction. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 3042–3058). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.179

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, *21*(140), 1–67. Retrieved November 9, 2021, from http://jmlr.org/papers/v21/20-074.html

Rayner, K., & Morrison, R. E. (1981). Eye movements and identifying words in parafoveal vision. *Bulletin of the Psychonomic Society*, *17*(3), 135–138. https://doi.org/10.3758/BF03333690

Rego, A. T. L., Melo, A. N., Snell, J., & Meeter, M. (2026). What drives regressions in reading? Insights from surprisal and saliency from language models. *Cognition*, *273*, 106535. https://doi.org/10.1016/j.cognition.2026.106535

Reichle, E. D., Rayner, K., & Pollatsek, A. (2012). Eye movements in reading versus nonreading tasks: Using E-Z Reader to understand the role of word/stimulus familiarity [_eprint: https://doi.org/10.1080/13506285.2012.667006]. *Visual Cognition*, *20*(4-5), 360–390. https://doi.org/10.1080/13506285.2012.667006

Ryu, S. H., & Lewis, R. (2021). Accounting for Agreement Phenomena in Sentence Comprehension with Transformer Language Models: Effects of Similarity-based Interference on Surprisal and Attention. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 61–71. https://doi.org/10.18653/v1/2021.cmcl-1.6

Shain, C. (2024). Word Frequency and Predictability Dissociate in Naturalistic Reading. *Open Mind*, *8*, 177–201. https://doi.org/10.1162/opmi_a_00119

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), e2307876121. https://doi.org/10.1073/pnas.2307876121

Smith, N. J., & Levy, R. P. (2008). Optimal Processing Times in Reading: A Formal Model and Empirical Investigation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*. Retrieved June 8, 2020, from https://escholarship.org/uc/item/3mr8m3rf

Smith, N. J., & Levy, R. P. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. *Proceedings of the Annual Meeting of the Cognitive Science Society, 33*, 7.

Smith, N. J., & Levy, R. P. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., Hofmann, V., Jha, A., Kumar, S., Lucy, L., Lyu, X., Lambert, N., Magnusson, I., Morrison, J., Muennighoff, N., … Lo, K. (2024, August). Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15725–15788). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.840

Staub, A., & Clifton, C. (2006). Syntactic prediction in language comprehension: Evidence from either...or. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 425–436. https://doi.org/10.1037/0278-7393.32.2.425

Sun, X., & Lu, W. (2022, July). Implicit n-grams Induced by Recurrence. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies (pp. 1624–1639). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.117

Timkey, W., & Linzen, T. (2023, December). A Language Model with Limited Memory Capacity Captures Interference in Human Sentence Processing. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 8705–8720). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.582

Vaidya, A., Turek, J., & Huth, A. (2023, December). Humans and language models diverge when predicting repeating text. In J. Jiang, D. Reitter, & S. Deng (Eds.), *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)* (pp. 58–69). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-1.5

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

Voita, E., Ferrando, J., & Nalmpantis, C. (2024, August). Neurons in Large Language Models: Dead, N-gram, Positional. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 1288–1301). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.75

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. P. (2020). On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci 2020)*.

Wilcox, E. G., Meister, C., Cotterell, R., & Pimentel, T. (2023, December). Language Model Quality Correlates with Psychometric Predictive Power in Multiple Languages. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7503–7511). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.466

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470. https://doi.org/10.1162/tacl_a_00612

Wilcox, E. G., Hu, M. Y., Mueller, A., Warstadt, A., Choshen, L., Zhuang, C., Williams, A., Cotterell, R., & Linzen, T. (2025). Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, *144*, 104650. https://doi.org/10.1016/j.jml.2025.104650

Wilcox, E. G., Pimentel, T., Meister, C., & Cotterell, R. (2024). An information-theoretic analysis of targeted regressions during reading. *Cognition*, *249*, 105765. https://doi.org/10.1016/j.cognition.2024.105765