Eye movements reveal a dissociation between prediction and structural processing in language comprehension

William Timkey[1]*, Kuan-Jung Huang[2], Byung-Doh Oh[1], Grusha Prasad[3], Suhas Arehalli[4], Tal Linzen[1], Brian Dillon[5]

[1]New York University, [2]University of Maryland, [3]Colgate University, [4]Macalester College, [5]University of Massachusetts Amherst

Abstract

Reading seems smooth and effortless, but this appearance is deceiving: In the process of extracting a meaning from a text, our eyes linger much more on some words than others, and we often reread earlier portions of the text. These disruptions are particularly common in syntactically challenging sentences. What explains this complex pattern of eye movements? One prominent hypothesis explains these patterns as a special case of the impact of a word's predictability (its surprisal) on the difficulty of recognizing the word. An alternative hypothesis attributes these disruptions to syntactic structure-building operations that apply after word recognition. Earlier attempts to address this debate have been inconclusive because of small numbers of participants, coarse measurements that are ill-suited to disentangling these competing views, and a limited range of predictability estimates. Here, we conduct a large-scale study ($n = 368$) examining eye movements during the reading of syntactically challenging sentences, using 407 types of predictability estimates from large language models with multiple architectures and training settings. We find a stark dissociation: Early reading measures are well approximated by language model surprisal, but syntactic disambiguation incurs a significant additional cost, reflected in an increase in rereading that is not explained by surprisal. We further show that when rereading parts of the sentence, readers strategically target the words

*To whom correspondence should be addressed. E-mail: wpt2011@nyu.edu.

W.T., K-J.H., B-D.O., G.P., S.A., T.L., and B.D. designed research; K-J.H. collected data; W.T. and B-D.O. analyzed data; and W.T., K-J.H., B-D.O., G.P., S.A., T.L., and B.D. wrote the paper. The authors declare no competing interest.

most useful to amend the structure of the sentence. We conclude that linguistic knowledge guides moment-by-moment reading in two dissociable ways: Forward reading is driven by the word's surprisal, and backward reading reflects syntactic structure building operations.

1 Introduction

Humans possess a nearly universal ability to communicate complex thoughts to one another through language. We are able to do so because our languages are governed by syntactic rules, which determine how an utterance must be structured in order to convey a given meaning: We know, for example, that the sentence "The dog found the hiker" has a different meaning from "The hiker found the dog" because English syntax associates word order with grammatical role. We can rapidly and incrementally apply this grammatical knowledge to reconstruct the intended meaning of a sentence from speech or text (Dahan and Ferreira, 2019; Frazier and Rayner, 1982; Friederici, 2011; Levy, 2008a; Ryskin and Nieuwland, 2023; Tanenhaus et al., 1995). While language comprehension often seems effortless, decades of research have shown that a great deal of complexity underlies this process (Wagers and Dillon, 2025). As a sentence becomes more syntactically challenging, we observe greater processing difficulty, as reflected in a range of measurements, including acceptability judgments and electrophysiological measurements. In natural reading, in particular, readers will linger on a particular word, or return to earlier points in the sentence and reread them. Such comprehension difficulty can be seen when syntactic ambiguity is resolved (Frazier and Rayner, 1982; Stowe, 1986; Tabor et al., 2004; Traxler et al., 1998), at points of high syntactic complexity (Bartek et al., 2011; Gibson, 2000; Lewis and Vasishth, 2005; Staub et al., 2017), when processing rare syntactic structures (Christiansen and MacDonald, 2009; Vasishth et al., 2010), or when processing outright anomaly (Dillon et al., 2013; Molinaro et al., 2011; Warren and McConnell, 2007). Investigations of these sources of comprehension difficulty have been critical in developing our understanding of the mental processes that support language comprehension.

Could all these disruptions to the language comprehension process be explained by a single cognitive mechanism? One candidate for such a unified account of reading difficulty, surprisal theory (Hale, 2001; Levy, 2008a; Shain et al., 2024; Smith and Levy, 2013; Vigly et al., 2025), casts language comprehension as a pervasively *predictive* process. Words that are more predictable from their context are in general processed faster (Ehrlich and Rayner, 1981); processing difficulty, reflected for example in the amount of time spent reading a word, increases

roughly linearly with the word's surprisal, i.e. its negative log probability in context (Smith and Levy, 2013). Surprisal theory contends that, because syntactic rules serve as one of the source of information used to predict upcoming words, the principle of next-word prediction is sufficient to explain the difficulty readers experience in both simple and syntactically complex sentences (Hale, 2001; Levy, 2008a). This theory makes a number of predictions that have been empirically validated in a range of studies (Staub, 2024).

Central to such empirical tests of surprisal theory are precise estimates of a word's contextual probability. In recent years, this role has increasingly been served by large language models (LLMs) based on neural networks. LLMs are machine learning systems that learn to predict upcoming words based on earlier ones; to the extent that these predictions align with those made by humans, these systems can be used to test the predictions of surprisal theory. Indeed, LLM-based estimates of word-by-word contextual probability have been shown to predict human reading behavior on naturalistic sentences from short stories, novels, and news articles (Oh and Schuler, 2023b; Shain et al., 2024; Wilcox et al., 2020). While these results suggest that prediction is an important aspect of comprehension, they do not yet provide a unified account of comprehension across different levels of syntactic complexity, because the stimuli from these studies generally consist of syntactically simple sentences that are not explicitly designed to probe different types of structural processing difficulty. To establish that LLM surprisal can unify the range of structural processing difficulty effects seen in language comprehension, we need to determine if LLM surprisal can explain the full magnitude of the difficulty observed in human experiments, a magnitude that can sometimes be quite substantial (van Schijndel and Linzen, 2021).

A number of recent studies have responded to this challenge with targeted tests of surprisal's ability to predict processing difficulty in syntactically challenging sentences. In one such class of sentences, which is the focus of the present study, syntactic ambiguity early on in the sentence is resolved towards an unexpected structure. Suppose, for example, that a sentence that begins with "The hiker found the dog ..." continues with "was a delightful companion." Here, readers' realization that *the dog* is the subject of a new clause, rather than the direct object of *found*, is associated with substantial processing difficulty (Frazier and Rayner, 1982; Garnsey et al., 1997; Sturt et al., 1999). These recent studies have found that while surprisal theory successfully predicts the existence of some amount of processing difficulty at such disambiguation points (Futrell et al., 2019; Hu et al., 2020; van Schijndel and Linzen, 2018, 2021;

Wilcox et al., 2021), it dramatically under-predicts the *magnitude* of the slowdown associated with such disambiguation, sometimes by orders of magnitude (Arehalli et al., 2022; Huang et al., 2024; Kobzeva and Kush, 2024; van Schijndel and Linzen, 2021; Wilcox et al., 2021). To summarize, surprisal theory has made a range of successful—and non-obvious—predictions, but has persistently failed to account for central aspects of structural processing (Staub, 2024).

What underlies the gap between the predictions of LLM surprisal and empirical human reading times? One potential answer to this question is that surprisal estimates from the handful of LLMs used in previous studies are not well calibrated to a human's probabilistic syntactic knowledge, and that a more human-like probability model may be able to fully capture all reading behavior in syntactically challenging sentences (Oh and Linzen, 2025; Wilcox et al., 2020). But another, theoretically significant possibility is that structural processing and surprisal make distinct contributions to reading difficulty, and as such surprisal cannot serve as a unified theory explaining processing difficulty in reading, even if we had a hypothetical LLM that perfectly matched human next-word predictions.

Previous targeted studies of the gap between surprisal's predictions and human reading times (Kobzeva and Kush, 2024; van Schijndel and Linzen, 2021; Wilcox et al., 2021) have used coarse behavioral measures—primarily the latency of button presses in paradigms where the sentence is revealed word by word when the participant presses a button (*self-paced reading*)—which make it difficult to isolate the predictive processes from those involved in structural processing. Here, we use the much more sensitive eye-tracking-while-reading paradigm (Rayner, 1998), where we present participants with an entire sentence at once, and record their eye movements as they read it naturally. As we describe below, this detailed eye-movement record makes it possible to derive multiple measures of difficulty associated with different processing stages (Clifton Jr. et al., 2007; Reichle et al., 2009; Schotter and Dillon, 2025; Staub, 2011). In particular, we can distinguish the amount of time readers spend on a word when first encountering it, a stage that reflects the predictability of the word (Reichle et al., 2003, 2009; Richter et al., 2006; Smith and Levy, 2013; Staub, 2015, 2024), from later processes reflected in regressive eye movements to earlier points in the sentence, followed by rereading of certain parts of the sentence (Frazier and Rayner, 1982; von der Malsburg and Vasishth, 2011; Reichle et al., 2009; Warren et al., 2011).

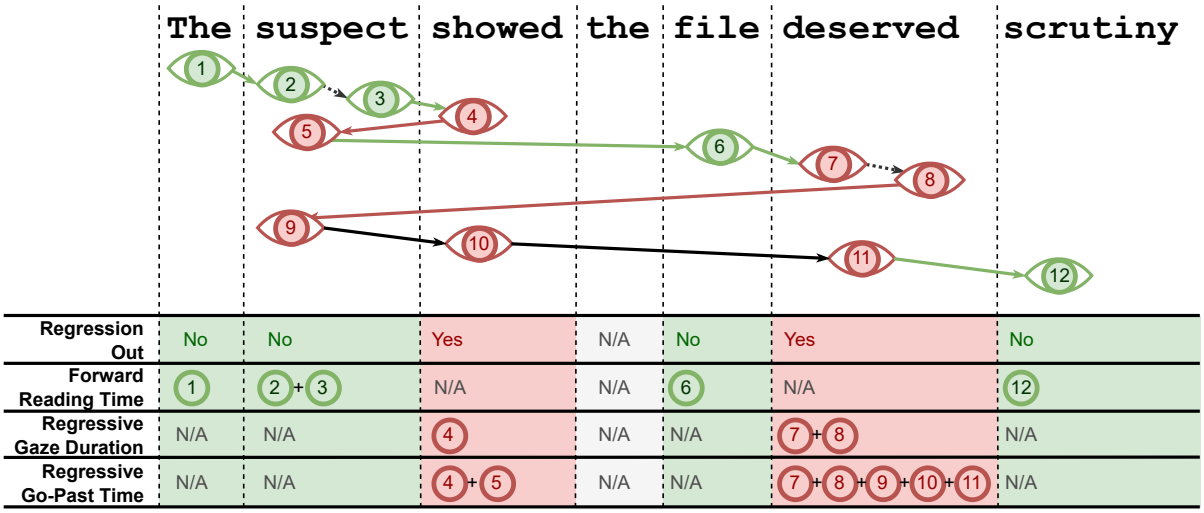
We collect a large-scale dataset with eye-movement recordings from 368 participants who read a diverse set of syntactically challenging constructions. Our materials are drawn from

Huang et al. (2024), a study which used the simpler self-paced reading methodology. We then use this dataset to evaluate the predictions of 407 different LLM-based surprisal estimates, derived from a range of types of LLMs, for four different measures derived from the eye-movement record. Overall, we find that LLM surprisal is largely able to explain the magnitude of human processing difficulty in the earliest stages of reading—that is, the amount of time a comprehender spends looking at a word for the first time—but it fails to explain later measures, including regressions and rereading patterns, which are plausibly associated with the structural processing necessary to integrate a word into a revised sentential context.

The LLMs' success in predicting early word recognition processes suggests that the misalignment between surprisal's predictions and human processing difficulty reported in earlier studies, which used coarser behavioral measures, cannot be attributed primarily to a general failure of LLMs to approximate the next-word probability distributions that guides human readers, or to the particular LLMs used in previous studies—indeed, we find that a much wider range of LLMs than explored in prior work are able to successfully predict early eye movement measures in syntactically challenging sentences. Instead, surprisal theory specifically fails to capture later stages of structural processing difficulty, where the word, after it has been recognized, needs to be integrated with the larger sentential context. We further show that rereading is informed by syntactic structure: Readers strategically reread the words that are most useful for amending the structure of the sentence. Overall, this pattern suggests that the structural processing associated with overcoming syntactic ambiguity is reflected in eye movements in a way that is dissociable from any effects of structure on word predictability, a finding that is unexpected on the view that surprisal reflects the sum total of the work that readers need to do to update their beliefs about the structure of linguistic input.

2 Measuring processing difficulty from eye movements

When we read, our eyes do not move smoothly from the start to the end of each line. Instead, they move through a text through a series of *fixations*, where the eyes remain stably focused on a single target point, and *saccades*, where the eyes rapidly move from one fixation target to another. The eyes primarily take up information during fixations. These fixations have a very narrow perceptual resolution: Readers can only process text within a few degrees of the fixation point, and must initiate a saccade to bring text occupying their peripheral vision into view. This makes the target of fixations a fine-grained indicator of what information is perceptually



	The	suspect	showed	the	file	deserved	scrutiny
Regression Out	No	No	Yes	N/A	No	Yes	No
Forward Reading Time	1	2 + 3	N/A	N/A	6	N/A	12
Regressive Gaze Duration	N/A	N/A	4	N/A	N/A	7 + 8	N/A
Regressive Go-Past Time	N/A	N/A	4 + 5	N/A	N/A	7 + 8 + 9 + 10 + 11	N/A

Figure 1: A visualization of how each word-level eye-tracking measure is computed from a sequence of fixations in a single trial. Each circle represents a fixation of some duration in a particular region of the sentence, and arrows represent the saccades between these fixations. Forward (first pass) interword saccades are colored green, regressive interword saccades are colored red, all other saccades are colored black. Fixations associated with forward reading are colored green, and fixations associated with regressive measures are colored red.

available to a reader at a given time. If we assume that the allocation of a reader's attention at a given time is reflected by the location of their fixations (Just and Carpenter, 1980), we can make inferences about word-by-word processing difficulty from the eye movement record. This record can be mapped onto word-level measures of reading difficulty based on the *number* and *duration* of fixations—where increased difficulty is generally associated with multiple fixations on the same word and longer durations of each fixation—and based on the *direction* of saccades between words. Around 80% to 95% of saccades between words are *forward* (or *progressive*), where a reader moves from the current word to a following word (Rayner, 1998). The remaining 5% to 20% of saccades between words are *backward* (or *regressive*), where a reader moves from the current word to a preceding word, such as when they reread. Regressive saccades have been shown to index interruptions and errors in processing (Altmann et al., 1992; Frazier and Rayner, 1982; Mitchell et al., 2008; Staub, 2011; Wilcox et al., 2024).

We select four reading measures for analysis in this work. Unlike some other measures used in the eye-tracking literature, all four measures only refer to the current word (w_i) and previous words ($w_{<i}$), which makes it possible to relate these measures to the surprisal of w_i (which naturally cannot take into account the words that follow w_i). These definitions all make reference to *first-pass* fixations—any fixations on w_i from when it is fixated for the first time before a saccade exits this word, but only in the first pass through the word, that is, given

that all prior fixations were on only w_i or words to its left $w_{<i}$. The four measures are as follows (see Figure 1 for a graphic illustration):

1. **FORWARD READING TIME**: The total duration of all first-pass fixations on word w_i in trials where the saccade which exits w_i is a *forward* saccade to $w_{>i}$ (i.e. to the right; other trials are excluded when computing this measure). This measure is typically taken to capture the cost of processing w_i when the process of integrating it into its context proceeds without errors (Engbert et al., 2005; Reichle et al., 2003, 2009).

2. **(FIRST-PASS) REGRESSIONS OUT**: A binary measure indicating the direction—left (coded as 1) or right (coded as 0)—of the first saccade which exits word w_i after all first-pass fixations on w_i. We aggregate this measure across trials as the proportion of saccades out of the word that were regressive (i.e. to the left). Models of eye movements in reading associate regressions with difficulty and errors related to integrating w_i with its context (Engbert et al., 2005; Reichle et al., 2009).

3. **REGRESSIVE GAZE DURATION**: The total duration of all first-pass fixations on word w_i in trials where the saccade which exits w_i is a *regressive* saccade (i.e. to the left). This measure reflects the partial cost of processing w_i before an error-driven regressive saccade is executed (Reichle et al., 2009).

4. **REGRESSIVE GO-PAST TIME**: This measure, computed only for trials where the reader's first-pass at w_i ends in a regressive saccade, is the sum of the duration of all first-pass fixations on word w_i, plus the duration of all subsequent fixations on the preceding words $w_{\leq i}$ until any word $w_{>i}$ is fixated. This measure is generally taken to reflect the cost of integrating w_i with its context when this process is interrupted or runs into errors (Clifton Jr. et al., 2007; Staub, 2011).

These four measures can be distinguished along two related axes. The first is time course: We refer to the measures that depend only on first pass fixations, FORWARD READING TIME and REGRESSIVE GAZE DURATION, as early measures, and to measures that incorporate rereading as late measures. The second axis is whether the measure indexes forward reading or rereading: FORWARD READING TIME is a forward measure, while the other measures are regressive measures.

If surprisal explains the full cost of recognizing and integrating a word with its context, as argued by surprisal theory, then we expect it to explain the full magnitude of syntactic processing effects in both early measures (e.g. FORWARD READING TIME) and late measures (e.g. REGRESSIVE GO-PAST TIME). If it only explains word recognition, we expect it to explain only early measures and not late measures.

The fixation time measures we analyze are *regression-contingent* (Altmann, 1994; Altmann et al., 1992; Rayner and Sereno, 1994; Vonk and Cozijn, 2003): Their definitions depend on the direction of the saccade that exits the word after the first-pass fixations on the word, either forward or regressive. This contrasts with more typical eye-tracking-while-reading measures, where fixation times are aggregated over both exit saccade directions. Our decision to analyze regression-contingent fixation measures has several empirical and theoretical motivations discussed in Appendix B; we present an analysis of more standard aggregate measures in Appendix B.1. The aggregate and regression-contingent analyses generally converge in their findings, but, as shown in the Appendix, many of the insights afforded by the regression-contingent analysis are obscured in aggregate analysis.

3 LLM surprisal explains forward reading but not rereading

Data collection and effects of interest. We analyze eye movement data from 368 participants (out of a total of 424 participants; the remainder are excluded based on the criteria described in *Materials and methods*). Participants read sentences from six types of syntactically challenging constructions that have been key to the development of psycholinguistic theories; each syntactically challenging construction is matched with a control construction that is maximally similar to it while avoiding the syntactic challenge in question (Figure 2, step 1; see *Materials and methods* for the list of constructions). These syntactically challenging ("experimental") and control sentences are interspersed with naturalistic *filler* sentences sampled from a corpus of news text and novels (Luke and Christianson, 2018). Five of the six types of syntactically challenging constructions contain a particular word where processing difficulty is predicted to arise and manifest in readers' eye movements; we refer to this word as the *critical word*. The sixth construction, relative clauses, has three different words where difficulty is predicted to differ between the conditions: the determiner, the noun, and the verb.

We calculate the empirical effects of interest as the difference in reading time or regression rates between the experimental and control sentences of each construction (Figure 2, step 5a).

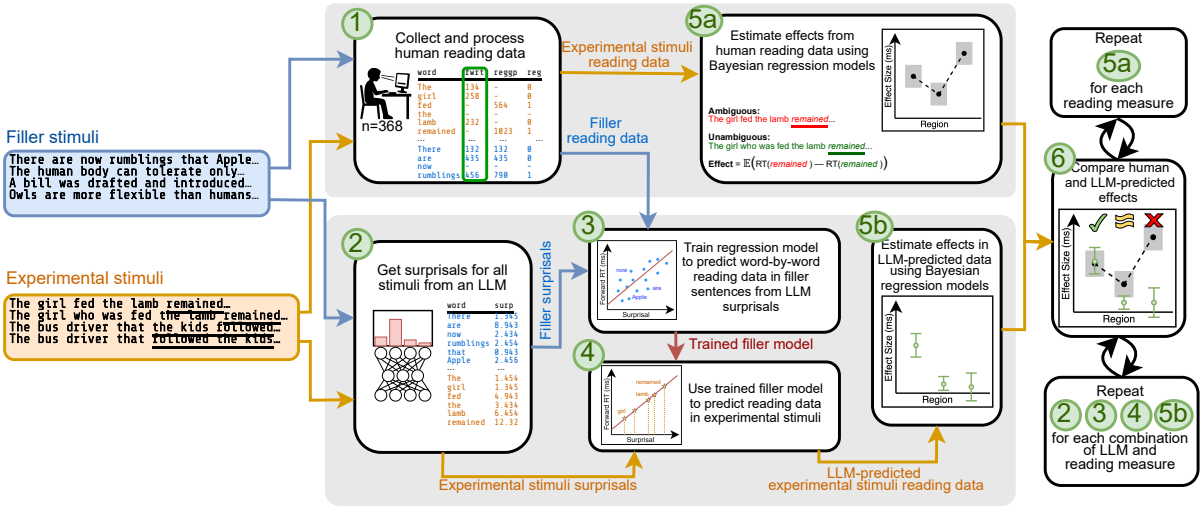


Figure 2: An overview of the methods employed in this work. The empirical effects due to syntactic disambiguation were first estimated from eye-tracking data (top), which were then compared against those predicted by LLM surprisal (bottom).

We measure this difference separately for the critical word and the "spillover word", that is, the word that immediately follows it. We do so because processing difficulty on word w_i is often reflected in the reading of w_{i+1}, a phenomenon referred to as "spillover". We estimate the effect sizes in each of the four reading measures using Bayesian regression models (see *Materials and methods*).

The clearest patterns of results across multiple eye movement measures—and therefore the best test beds for the surprisal hypothesis—is seen in two sets of constructions: The relative clauses and the three constructions where the beginning of the sentence has a preferred structural parse that then turns out to be inconsistent with the full sentence. We refer to these three constructions—Main Verb/Reduced Relative (MV/RR), Intransitive/Transitive (NP/Z) and Noun Phrase/Sentential Complement (NP/S)—as the "garden path constructions", following the traditional terminology in psycholinguistics. We focus here on these four constructions, and discuss results from the remaining two constructions, subject-verb agreement violations and attachment ambiguities, in Appendix D.

Predicting reading behavior from LLM surprisal. We follow the *generalization paradigm* used in previous studies (Arehalli et al., 2022; Huang et al., 2024; van Schijndel and Linzen, 2021); this paradigm is illustrated as steps 2–6 in the lower half of Figure 2. We first obtain LLM surprisal estimates for each of the words in our filler and experimental stimuli (Figure 2, step 2). We obtain 407 types of estimates derived from LLMs from six LLM families: the Pythia suite

(Biderman et al., 2023) of transformers (Vaswani et al., 2017); the GPT-2 family of transformers (Radford et al., 2019); the Mamba family of state-space models (Gu and Dao, 2023); a family of 125 Long Short-Term Memory (LSTM; Hochreiter and Schmidhuber, 1997) language models (van Schijndel et al., 2019), which we refer to as LSTM (vS); another LSTM language model trained on text from Wikipedia (Gulordava et al., 2018), which we refer to as LSTM (G); and a family of 10 Recurrent Neural Network Grammars (RNNGs) (Dyer et al., 2016; Noji and Oseki, 2021). The RNNGs are unique among the LLMs we test in that they are trained with explicit representations of the syntactic structure of the sentence. We extract surprisal estimates from the each RNNGs using 12 different beam widths, for a total of 120 sets of RNNG-based surprisal estimates. For details on the LLMs, see Appendix A.

For each combination of LLM-based surprisal estimate and reading measure, we train a regression model that predicts word-level reading measurements for each participant from the surprisal estimate and other word-level predictors (Figure 2, step 3). Each observation in these regression models is a single reading measurement at a single word from a single trial. These regression models, which we refer to as *filler models*, are crucially trained only on reading data from the filler sentences, not the experimental sentences. For each measure, we also train a *baseline* regression model, which predicts the same data from baseline predictors only (word frequency, length and position). We use linear regression for the continuous measures and logistic regression for REGRESSIONS OUT; see *Materials and methods* for details.

Before testing LLM predictions for the targeted experimental stimuli, we evaluate the power of surprisal from each LLM to predict reading data in the filler sentences on each reading measure. We quantify predictive power as the increase in the log-likelihood (ΔLL) of the filler reading measurements when surprisal from an LLM is included as a predictor, compared to a filler model with only baseline predictors, without surprisal (Wilcox et al., 2020). Because different filler models for different measures are fit to a different number of observations, we report the per-observation ΔLL to facilitate comparisons of predictive power across measures. Note that ΔLL only allows relative comparison of the goodness-of-fit of two regression models; it does not indicate what proportion of the variance is explained by the model. To address this, we compute an upper bound on the predictive power attainable by any word-level predictor of reading data, comparable to the noise ceiling measurements used in fMRI studies (Schrimpf et al., 2021). We then compare the ΔLL attained by LLM surprisal to this upper bound; see Appendix C.2 for details.

We then generate predicted reading measurements from the 407 filler models at the critical and spillover regions in the held-out experimental sentences (Figure 2, step 4). By maintaining this test-train split, we test a key prediction of surprisal theory: If surprisal can fully explain processing difficulty in complex and simple sentences alike under a single mechanism, then the relationship between surprisal and processing difficulty in the filler sentences should generalize to the syntactically challenging experimental sentences (van Schijndel and Linzen, 2021). From the LLM-predicted data, we derive predicted effects using regression models analogous to those used to analyze the empirical data. Both the human and LLM-predicted effect estimates take participant-level and item-level uncertainty into account. Due to computational resource limitations, we fit Bayesian regression models to the LLM-predicted reading data from only one LLM-derived surprisal estimate from each of the six LLM families. We selected the LLM in each LLM family that achieved the highest per-observation ΔLL averaged across all four measures: We reasoned that under surprisal theory the LLM which best explains reading behavior in the fillers should also best explain the held-out data. For the remaining models, we fit frequentist regression models with identical specifications to the Bayesian models. Detailed results across all LLMs from each family are provided in Appendix G.

Finally, we compare the LLM-predicted effect estimates with the empirical effects (Figure 2, step 6). We say that surprisal predicts or explains a measure if there is substantial overlap between the plausible range of model predictions and the plausible range of empirical effect sizes, where plausible ranges are based on the posterior estimates of the Bayesian models.

In filler sentences, LLM surprisal explains early measures better than late ones. In the models fit to the filler sentences, LLM surprisal is a significant predictor of all four measures in at least one model of each of the six LLM families we investigated (for model coefficients see Appendix C.1). How much variance is explained by surprisal? We consider this both in absolute terms, comparing absolute ΔLL across the four eye movement measures, and relative to the ceiling of explainable variance in each eye movement measure. Because we have a different number of observations for each measure (since they are conditioned on the direction of the saccade following the word), we divide ΔLL by the number of observations we have before comparing it across measures.

Among the four reading measures, LLM surprisal has greater predictive power in FORWARD READING TIME than any other measure (Figure 3), in both absolute and relative terms

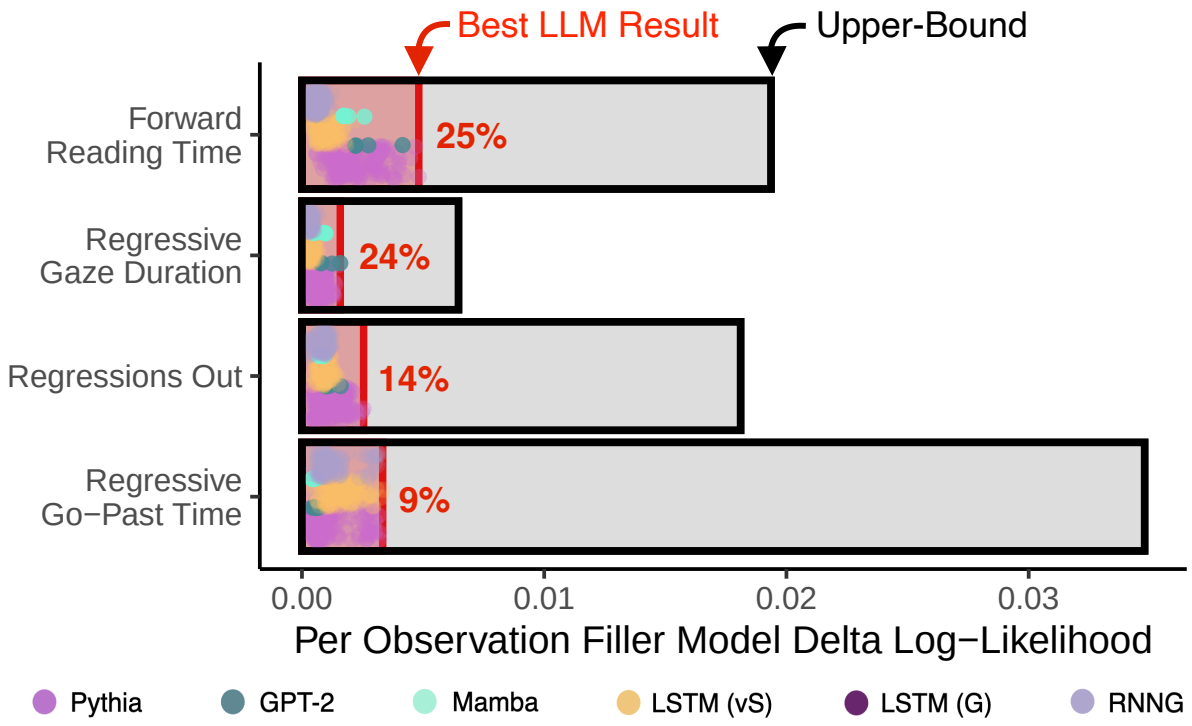


Figure 3: Predictive power of LLM surprisal in filler models for each of the four eyetracking measures, measured as delta log-likelihood (ΔLL) of the filler data over the no-surprisal baseline filler model. Because filler models for different measures are fit to a different number of observations (depending on the direction of the saccade exiting the word), we divide the total ΔLL by the number of observations each filler model is fit to, where each observation is a measurement for a single word from a single trial. Each dot corresponds to one of the 407 LLM-based surprisal estimates. A red line denotes the highest ΔLL attained by any LLM for a given measure, and the upper bound on ΔLL is denoted by gray bars.

(0.004 and 25% respectively), though in relative terms it is closely followed by REGRESSIVE GAZE DURATION (absolute: 0.0016; relative: 24%). In absolute terms, LLM surprisal attains the second largest ΔLL in REGRESSIVE GO-PAST TIME, but there is a great deal of explainable variance in this measure that was not captured by surprisal, with the best LLM attaining only 9% of the variance ceiling. Absolute ΔLL is lowest in REGRESSIVE GAZE DURATION, but there is high variability in this measure across participants, leading to a low variance ceiling.

Overall, surprisal is a relatively strong predictor of the early reading measures FORWARD READING TIME and REGRESSIVE GAZE DURATION in the filler sentences, but in the late reading measures it leaves a great deal of variance unexplained. To determine whether part of this unexplained variance should be attributed to syntactic processing difficulty, we next turn to a test of whether syntactic context drives reading behavior above and beyond surprisal and control predictors.

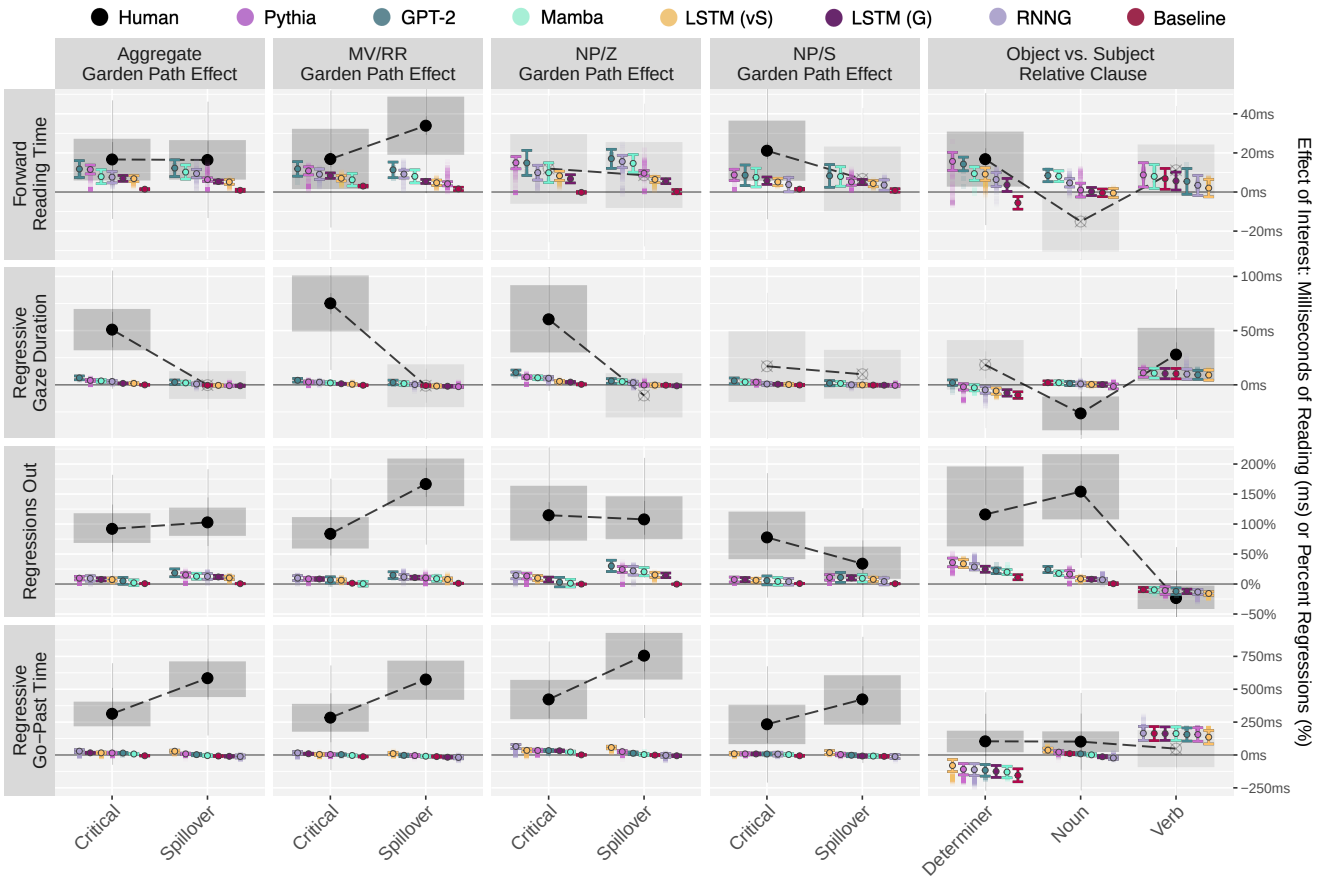


Figure 4: Empirical (Human) and predicted (LLM) effects of interest at the regions of interest in four selected constructions (columns) and four measures (rows). Human effect estimates are represented with black points for the mean, with shaded regions representing 95% credible intervals in the vertical dimension. Empirical null effects are denoted with lighter shading and an X on the mean estimate point. Colored points are predicted effect sizes from the LLM within each architecture whose surprisal values provided the best fit to reading data in the filler sentences, measured by filler model ΔLL. Within each plot, predicted LLM effects are sorted from largest to smallest, with error bars denoting 95% credible intervals, incorporating participant-level and item-level uncertainty. Mean effect estimates from all 407 LLM-derived surprisal estimates are displayed as square semi-transparent colored points without error bars.

LLM surprisal can explain FORWARD READING TIME in syntactically challenging sentences. Overall, there is substantial overlap between the predictions of LLM surprisal and the empirical FORWARD READING TIME measurements. In FORWARD READING in the relative clause construction, for example, the challenging sentences (ORC) are read 17 ms [3–31 ms] more slowly than the controls (SRC) at the determiner region; this is consistent with GPT-2 surprisal, which predicts an effect of 14 ms [11–18 ms] at this region. We also find evidence for an aggregate disambiguation effect, when aggregated across the three garden path constructions (MV/RR, NP/S and NP/Z), at both the critical (17 ms [6–27 ms]) and spillover word (16 ms [6–26 ms]). LLM surprisal from multiple LLM families is able to predict the full magnitude of these effects: The predicted effects either lie fully within the 95% credible interval of the empirical estimates of the human effects, or overlap substantially with it. Closest to the human effect magnitudes is GPT-2, which predicts an aggregate garden path effect of 12 ms [7–18 ms] at the critical word and 12 ms [8–16 ms] at the spillover word. More generally, the magnitude of the human effects overlap with the predictions of at least one LLM family in all constructions and regions, with the exception of the spillover word in the MV/RR construction, where the empirical effect of 33 ms [19–48 ms] is three times larger than the GPT-2 prediction of 11 ms [7–15 ms].

LLM surprisal fails to explain rereading in syntactically challenging sentences. We now turn to the measures associated with rereading: REGRESSIONS OUT, REGRESSIVE GAZE, and REGRESSIVE GO-PAST. In the three garden path constructions, the overall rate of regressions is 114% [72–164%] higher in the ambiguous sentences than the unambiguous controls at the critical word, and 108% [75–146%] higher at the spillover word. This is many times larger than the largest LLM-predicted effects of 10% [7–13%] at the critical word (from Pythia) and 19% [12–25%] at the spillover word (from GPT-2). A similar pattern of results holds for the regressive reading time measures. The mean reading time preceding a regression (REGRESSIVE GAZE) is 51 ms [32–70 ms] longer in ambiguous sentences than unambiguous control sentences. By contrast, the largest LLM-predicted effect in REGRESSIVE GAZE is 6 ms [5–8 ms] from GPT-2, nearly eight times smaller than the human effect. Perhaps the most drastic misalignment between human reading and LLM predictions is in REGRESSIVE GO-PAST, where we observe an empirical effect of 313 ms [218–406 ms] at the critical word and 584 ms [441–712 ms] at the spillover word. By contrast, the largest LLM-predicted effect is eleven

times smaller at the critical word (29 ms [18–40 ms] from the RNNG) and 20 times smaller at the spillover word (29 ms [20–37 ms] from LSTM (vS)). Furthermore, LSTM (G), RNNG, and Mamba all fail to predict not only the magnitude, but also the direction of the REGRESSIVE GO-PAST effect at the spillover word.

In the relative clause subset, LLMs likewise fail to predict effect magnitudes in the measures associated with rereading. In humans, we observe a 116% [63–196%] increase in regressions at the determiner in the object relative clause condition over the subject relative clause condition; this contrasts with the largest LLM-predicted effect of 36% [29–43%] from Pythia. The mismatch is more pronounced at the noun, where we observe a 154% [108–216%] increase at the noun compared to the largest LLM-predicted effect of 24% [19–29%] from GPT-2. In contrast to the empirical pattern at the determiner and noun, there is a small decrease in regressions in the ORC condition at the verb of 23% [2–42%]. While at this word the human and LLM-predicted effects overlap, this is also true of the no-surprisal baseline filler model, indicating that surprisal is not necessary to predict this effect. Empirical effects in REGRESSIVE GO-PAST are much smaller in the relative clause subset than the garden path subset. At the determiner, all LLM-predicted effects in effects in REGRESSIVE GO-PAST are in the opposite direction of those observed in humans at the determiner. There is also an effect of 102 ms [23–180 ms] at the noun, which overlaps with the prediction of 37 ms [22–52 ms] from LSTM (vS).

In summary, the misalignment between the empirical and LLM-predicted results is particularly clear in regressive measures in the three garden path construction, where LLM surprisal underpredicts both the frequency (REGRESSIONS OUT) and duration (REGRESSIVE GO-PAST TIME) of rereading by an order of magnitude. In the relative clause subset, the misalignment is clearer in REGRESSIONS OUT than the other regressive measures.

Discussion. Our data reveals syntactic disambiguation effects in both the relative clauses and garden path constructions, and in all of the reading measures we analyzed. LLM surprisal predicts similar effect magnitudes to those observed in humans in forward reading, but consistently fails to predict the rate at which subjects reread words, or the amount of time they spend rereading when encountering the critical word of a syntactically challenging sentence.

The similarity of the human and LLM-predicted effects in forward reading is not simply due to the fact that empirical effect magnitudes are generally smaller in FORWARD READING TIME than in the rereading measures; if anything, the LLM-predicted effects are often *larger* in

magnitude in FORWARD READING TIME than in the regressive measures. This reflects the fact that the surprisal coefficients in the filler models are larger in FORWARD READING TIME than REGRESSIVE GAZE, and were similar in magnitude to those in REGRESSIVE GO-PAST. This is likely driven by the fact that surprisal accounts for a greater proportion of variance and a larger absolute slowdown per bit of surprisal in forward reading time than it does in other measures (see Appendix A).

LLMs with more parameters and more training data—two properties that are strongly associated with better next-word-prediction accuracy (Biderman et al., 2023)—predicted *smaller* garden path effects (Appendix G), suggesting that further improving LLMs' next-word prediction accuracy is unlikely to lead to more accurate predictors of human rereading behavior. This finding parallels similar findings from an analysis of naturalistic reading (Oh and Schuler, 2023b). There was little difference in the magnitude of predicted effects between the RNNG models, which explicitly model syntactic structure, and the other LLM families, which do not. This suggests that the inability of LLM surprisal to predict the full costs of syntactic processing cannot be remedied solely by endowing LLMs with explicit syntactic representations (in line with Arehalli et al., 2022).

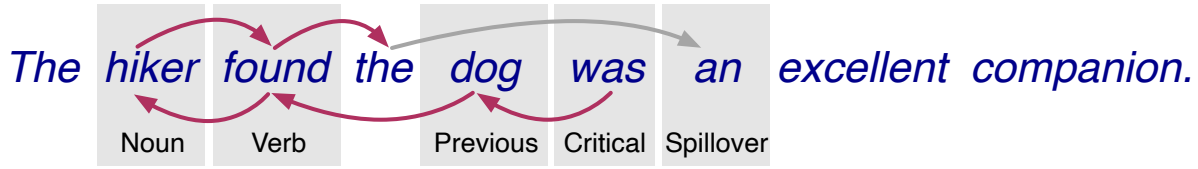
More generally, the effect sizes in REGRESSIVE GO-PAST cast doubt on the ability of any language model to produce surprisals high enough to predict the full effect magnitude. In our paradigm, LLM-predicted effects are approximately equal to the product of the difference in surprisal across conditions and the surprisal-to-reading-time regression coefficient estimated by the filler model. Assuming an optimistic surprisal-to-reading-time coefficient for REGRESSIVE GO-PAST of 12.4 ms per bit of surprisal (the upper 97.5% percentile of the largest coefficient estimate in Figure 7, which was from the RNNG), we would need to observe a surprisal *difference* of approximately 34 bits between the experimental and control condition in the garden path subset to predict even the lower bound of the empirical effect at the spillover word (which is 419 ms). But such a difference between the two conditions would require enormous surprisal values that are outside the range observed for our materials. The largest absolute surprisal value (i.e. not a difference across conditions) that we observe at any word in any of the experimental materials from the LLM with the largest coefficent was 25 bits (for the distribution of surprisal values across all words, see Appendix F). This means that even in the very unlikely case where the critical word were perfectly predictable in the unambiguous condition (i.e. its surprisal was zero), and were assigned the largest observed surprisal in the ambiguous

condition, we still would only predict an effect of 25 bits * 12.4 ms/bit = 310 ms, substantially less than the even the lower range of the empirical effect estimate. In summary, while it seems unlikely that LLM-based word surprisal from any LLM could explain the full magnitude of effects in the regressive measures, we find convergent evidence across multiple LLM families that surprisal can indeed explain effect magnitudes in uninterrupted forward reading.

4 Regressive eye movements are in the service of revising structure

The previous analyses suggest that LLM surprisal can predict the magnitude of the elevated FORWARD READING TIMES in syntactically challenging sentences compared to controls, but fails to account for the increased frequency of regressions to earlier words in those sentences. This dissociation suggests that the greater rate of regressions reflects a qualitatively distinct process from prediction. To shed light on this process, we next study which of the earlier words of the sentence are reread following regressive eye movements. We test the Selective Reanalysis Hypothesis, under which rereading is guided by the need to revise the structural representation of the sentence; as such, readers deploy more attention to the point of the sentence where they committed to a syntactic analysis that later turned out to be incorrect (Frazier and Rayner, 1982; Meseguer et al., 2002; Mitchell et al., 2008, for a related proposal, see Bicknell and Levy, 2010; Levy et al., 2009). To test this hypothesis, we focus on the three garden path constructions, which are associated with a significant increase in regressions that is not well explained by LLM surprisal (Figure 4). We contrast the Selective Reanalysis Hypothesis with the view that regressive eye movements do not serve to direct attention to parts of the sentence that require reanalysis, but instead are solely used to "buy time": They prevent the intake of new information while providing the reader with additional time to process the words that have already been read (Christianson et al., 2024; Inhoff and Weger, 2005; Mitchell et al., 2008). We refer to this view as the Time Out Hypothesis. Previous attempts to distinguish these two hypotheses have been inconclusive because of limitations due to sample sizes and analysis techniques (Frazier and Rayner, 1982).

We expect selective reanalysis to be most strongly associated with the ambiguous verb, *found* in the example in Figure 5. The syntactic frame (sometimes referred to as "subcategorization frame") associated with the verb *found* needs to be reanalyzed from one that includes a direct object (i.e. the hiker found something) to one with a sentential complement one (that is, the hiker found that something was the case). As such, the Selective Reanalysis Hypothesis



	Noun	Verb	Previous
Critical	1	2	1
Spillover	0	0	0

Figure 5: An example trial containing rereading initiated from the critical word *was* (top), and the six observations (three target regions by two source regions) included in the analysis from this trial (bottom). We count the number of fixations on three target regions (the noun *hiker*, the verb *found*, and the previous word *dog*) before the source region is exited to the right (illustrated by the gray arrow). The three 'Spillover' observations are zeroes as this trial does not any contain rereading initiated from the spillover word. If it did, fixations on *was* would be counted as those on the previous word.

predicts that this word will be the target of a larger number of regressions than other words.

Analyzing rereading behavior. We fit a Poisson regression model where the outcome variable is the number of fixations on each prior word, following first pass reading on either the critical or spillover word. We analyze fixation count, rather than fixation duration, for two reasons. First, a preliminary analysis showed that the different conditions are more clearly differentiated by the number of fixations rather than the durations of those fixations. Second, for trials that did not contain any rereading, the total fixation duration on prior words is zero, leading to a very skewed distribution over fixation duration; modeling the count of fixations on a previous word with a Poisson model allows these zeros to be explicitly accounted for.

Recall that due to spillover we expect the processing difficulty due to disambiguation to manifest on either the critical word or the following one (*was* and *an* respectively in Figure 5). As such, we analyze the fixations on earlier, pre-critical words following a regression launched from either of these two critical words. This includes the fixation on the first earlier word that immediately followed the regression, as well as any other subsequent fixations before the critical word or spillover word was eventually exited to the right, following all rereading fixations (the red arrows in Figure 5).

Within these post-regression fixations, we analyze the number of fixations on three target words. The first, which we refer to as the previous word, is the word immediately preceding either the critical, or spillover word (*dog* or *was* in Figure 5). Fixations on this word would be consistent with the Time Out Hypothesis, as this word is the closest to the site of regression and

as such regressing to it would require the least amount of effort (if, as the hypothesis argues, the content of the particular reread words is immaterial to rereading).

The second word we analyze is the earlier, ambiguous verb (*found* in Figure 5): This word is the source of syntactic ambiguity in the garden path sentences, and therefore a target that the Selective Reanalysis Hypothesis predicts will attract more fixations. The Selective Reanalysis Hypothesis also predicts more fixations on *dog* in Figure 5 as it also requires reanalysis (here, its role should be revised from the direct object *found* to the subject of the embedded clause), but because fixations on *dog* after a regression from *was* could also be driven by the need to buy time, as predicted by the Time Out Hypothesis, the ambiguous verb is the word for which rereading fixations most clearly adjudicate between the two hypotheses.

The last word of interest is the noun immediately preceding the earlier verb (*hiker* in Figure 5). We expect this noun to attract fixations due to two different reasons. First, this word may receive fixations as a result of overshooting a regressive saccade to the ambiguous verb, the target of rereading as predicted by the Selective Reanalysis Hypothesis. To account for this possibility, the Helmert coding scheme of our regression model (see *Materials and methods*) first contrasts the number of fixations on the previous word region from the average number of fixations on the verb and noun regions; we treat fixations on either of these two regions as consistent with the Selective Reanalysis Hypothesis. However, this noun may also attract fixations due to its physical position on the screen (e.g. as a byproduct of a scan path that starts at the very beginning of the sentence) rather than its structural role—a behavior more consistent with the Time Out Hypothesis than the Selective Reanalysis Hypothesis. Therefore, our coding scheme further contrasts the number of fixations on this noun region with that on the verb region; the Selective Reanalysis Hypothesis predicts a higher number of fixations on the verb region.

To summarize, the previous analyses show that more rereading takes place in the temporarily ambiguous garden path sentences than the controls. The Selective Reanalysis Hypothesis predicts that this increase is mainly due to increased fixations on the ambiguous verb itself rather than increased fixations on other word regions after a reader regresses from the critical or spillover word. As such, the Selective Reanalysis Hypothesis predicts an interaction of sentence ambiguity and region, such that the verb receives proportionally more fixations than either of the other regions in the analysis when the sentence is ambiguous.

All in all, we have six observations from each trial: fixations on each of the three target

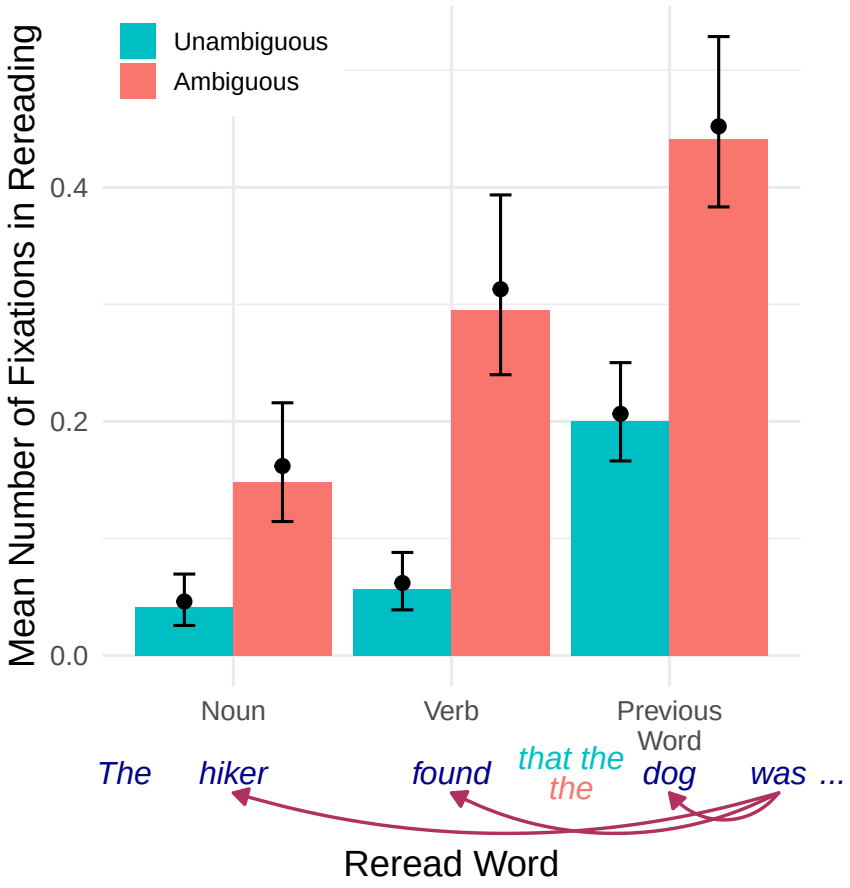


Figure 6: Number of fixations in rereading on each of the three target words by ambiguity condition, as observed in the human data (bars) and predicted by the regression model (dots and error bars). The numbers are lower than one because these words are not reread in most trials. The verb (*found*) attracts proportionally more fixations than the noun (*hiker*), and the verb (*found*) and noun (*hiker*) on average attracts proportionally more fixations than the previous word (*dog* or *was*), in ambiguous sentences (red) compared to unambiguous sentences (blue). The data were averaged across different constructions and source regions, and the error bars reflect highest-density 95% credible intervals of the posterior distribution conditioned on model parameters (see Appendix H).

regions following regressions initiated from the two source regions. We fit a Bayesian Poisson regression model (see *Materials and methods*) to examine the effect of ambiguity and target region on the number of fixations on earlier words. We aggregate the data across the three constructions (MV/RR, NP/S and NP/Z) to simplify the analysis, as neither the Selective Reanalysis Hypothesis nor the Time Out Hypothesis make fine-grained predictions that depend on the specific construction.

Readers reread the words that are most useful for amending the structure of the sentence. The empirical number of fixations on each target region, as well as posterior estimates from the regression model, are shown in Figure 6. When averaged across different constructions

and source regions, fixations on all three target words are more frequent in ambiguous trials than in unambiguous ones, more frequent on the previous word than on the noun or verb, and more frequent on the verb than on the noun. The two-way interactions between ambiguity and region indicate that the proportional increase in the number of fixations due to ambiguity is greater on the verb region than the noun region, and greater on the average of verb region and the noun region than the previous word region (see Appendix H for the posterior estimates of regression model parameters and posterior means of predictions for each target word and ambiguity condition). In other words, the verb attracts a larger share of rereading fixations in the syntactically challenging sentences—temporarily ambiguous sentences that need to be revised—than in the control sentences.

Discussion. The key finding of this analysis is that syntactic ambiguity drives a larger proportional increase in the number of fixations on the ambiguous verb than on either the noun that precedes this verb, or on the word that immediately precedes the word that triggered the regression. This partially selective pattern of fixations in syntactically challenging sentences is critical, since many of the preceding words are in positions that the Time Out Hypothesis predicts would be favored in re-reading just based on their physical location on the screen (close to the beginning of the sentence, or close to the word that triggered the regression). This is consistent with the predictions made by the Selective Reanalysis Hypothesis, according to which regressions are driven by the need to engage in additional structural processing. In this case, the re-reading may reflect an attempt to revise an initially incorrect interpretation or otherwise resolve uncertainty about specific previously encountered words (Bicknell and Levy, 2010; Frazier and Rayner, 1982; Levy et al., 2009).

Note that we have not attempted to characterize whether selective, structurally driven rereading behavior necessarily results in a more successful reanalysis. For example, Paape and Vasishth (2022) and Christianson et al. (2024) argue that people reread to confirm their initial, incorrect interpretation rather than to revise it, based on the lack of correlation between increased rereading and the resulting acceptability/accuracy of the final interpretation. However, an incorrect response to a comprehension question does not distinguish the lack of an attempt to reanalyze from a failed attempt to reanalyze: Comprehenders may attempt reanalysis upon reaching a point of syntactic disambiguation, but lack the means to carry out that reanalysis successfully (Christianson et al., 2001; Fodor and Inoue, 2000). In either case, the

finding that surprisal cannot account for this additional demand to reread further indicates a dissociation between prediction and structural processing during real-time comprehension.

5 General discussion

Prediction has been hypothesized to be central to cognition, including language. This hypothesis becomes even more alluring given the recent success of LLM-based AI systems, whose core language abilities emerge through a simple prediction-based objective. While it is clear that word-by-word language processing difficulty in humans is closely correlated with word predictability (especially in surprisal, the logarithm of a word's probability), the precise cognitive work indexed by surprisal in language processing remains an active topic of investigation (Hale, 2011; Hoover et al., 2023; Krieger et al., 2025; Shain et al., 2024). Understanding the limits of what can be explained by prediction is critical to the development of complete theories of language processing and cognition as a whole. In this work, we investigated whether a dissociation between prediction and other cognitive processes could be identified through the eye movements of readers in a large-scale benchmark dataset of reading behavior, containing both naturalistic and controlled experimental stimuli.

Our primary empirical finding is that word predictability can explain eye movements in forward reading, even in syntactically challenging sentences. But it cannot explain eye movements associated with interruptions to reading (i.e. rereading) in these syntactic contexts. This was based on surprisal estimates from over 400 language models, with a wide variety of architectures and training data. More specifically, surprisal can predict the full increase in a word's reading time when its context becomes more syntactically challenging, but only when fixations on the word are not followed by a regression (FORWARD READING TIME). By contrast, surprisal drastically underpredicts the rate at which critical words in syntactically challenging sentences trigger rereading (REGRESSIONS OUT), the reading time on a word preceding a regression (REGRESSIVE GAZE), and the amount of time spent rereading (REGRESSIVE GO-PAST). This dissociation was particularly clear in garden path constructions, where the surprisal-based predictions of effect magnitudes in FORWARD READING TIME generally overlapped with the actual effects observed in humans, but surprisal underpredicted effects in the rereading-based measures by orders of magnitude. This finding is corroborated by similar results in the dataset's naturalistic filler sentences. Controlling for baseline factors (word length, frequency, position), we saw that surprisal can better predict FORWARD READING TIME than

it can predict any of the measures associated with regressions or rereading.

Our results bring clarity to a puzzling empirical landscape. There is substantial empirical support for surprisal theory. Prior work gives evidence that surprisal can predict reading behavior in naturalistic sentences (Shain, 2024; Smith and Levy, 2013; Wilcox et al., 2020, 2023), correctly predicts that surprisal is linearly related to RTs even at the extremely low end of the probability scale (Shain et al., 2024; Smith and Levy, 2013; Wilcox et al., 2023). Despite this, there are clear limits to surprisal's predictive power. In particular, it drastically underestimates the full extent of the difficulty readers experience in during syntactic disambiguation (Huang et al., 2024; Kobzeva and Kush, 2024; van Schijndel and Linzen, 2021; Staub, 2024; Wilcox et al., 2021). There are many potential explanations for this misalignment, such as poor calibration of LLM predictions to those of humans (Oh and Linzen, 2025; Oh and Schuler, 2023b).

We conjectured instead that this now well-established finding arises because the sources of difficulty in reading times on naturalistic stimuli, which are well explained by surprisal, are distinct from the sources of difficulty on syntactically challenging sentences (e.g. garden path sentences), which are not. With the eye-tracking-while-reading paradigm, we were able to support this conjecture by showing that it is possible to trace distinct sources of difficulty to distinct reading measures. Indeed, our findings show that FORWARD READING TIME is well explained by surprisal in both naturalistic and syntactically challenging stimuli. Taken with independent converging evidence that variability in forward reading time is largely driven by lexical processing (i.e. word recognition difficulty), and variability in the other reading measures is primarily driven by difficulty integrating recognized words with their context (Clifton Jr. et al., 2007; Reichle and Sheridan, 2015; Reichle et al., 2009; Staub, 2011), a natural inference from our results is that surprisal explains (error-free) lexical processing, but not the post-lexical integration difficulty associated with structural processing.

Implications for theories of syntactic disambiguation. Sentence processing theories differ in whether they attribute syntactic disambiguation difficulty in garden path sentences to reanalysis (Frazier and Fodor, 1978; Lewis, 1998) or to predictability (Hale, 2001; Jurafsky, 1996; Levy, 2008a). Our results are consistent with an account wherein both surprisal and reanalysis contribute to the full cost of processing garden path sentences. While surprisal explains the added cost of recognizing (and error-free integration of) an unpredictable disambiguating word (indexed by FORWARD READING TIME), other mechanisms are needed to explain both

the rate at which integration of the disambiguating word fails (indexed by REGRESSIONS OUT), and the full cost of integration (indexed by REGRESSIVE GO-PAST TIME).

An open question for future research concerns exactly what drives the integration failures that result in regressive eye movements. One possibility is that garden path sentences lead comprehenders to create multiple, mutually incompatible interpretations of the stimulus over the course of comprehension, creating a representational conflict or competition that must be resolved (Botvinick, 2007; Botvinick et al., 2001). On this view, integration failures may result from the rapid detection of this conflict or incoherence, potentially cuing the comprehender to engage cognitive control processes to negotiate a determinate interpretation of the stimulus (for a recent review, see Ness et al., 2025). However, another approach would be to invoke mechanisms such as limited beam search over parses (Jurafsky, 1996) or particle filters (Levy, 2008b), where some of the parses drop out of consideration if they are less likely: The probability of a regression could reflect the probability that no parse remains under consideration. On this view, rapid 'integration failure' results from the recognition that there is no parse that is compatible with the current input.

Any full model would need to explain not only the probability of a regression, but also its target (Figure 6). This may reflect the work required to reconstruct the parse after all of the parses had been eliminated through beam search, or the work needed to resolve the competition between conflicting interpretations. Our analysis of the targets of fixations in rereading shows that these fixations may be in the service of reanalysis—future models of rereading should focus on how readers eventually update their syntactic (Christianson et al., 2024; Paape and Vasishth, 2022) or word representations (Bicknell and Levy, 2010; Levy et al., 2009) as a result of rereading, and how that influences subsequent reading behavior. Our large-scale eye-tracking dataset will serve as a useful benchmark for the development of such future models of garden-pathing.

Implications for surprisal theory. Various sentence processing theories predict a relationship between a word's surprisal and its processing difficulty. Our results suggest that the processing difficulty of a word can be decomposed into two parts, recognition and integration, and that surprisal only captures the former. This pattern is consistent with the Bayesian Reader (Levy, 2013; Norris, 2006, 2009), which casts word recognition as a problem of optimal visual discrimination, and derives that the optimal Bayesian solution to this problem predicts

a linear relationship between surprisal and word recognition times (Levy, 2013; Shain et al., 2024; Smith and Levy, 2013). Our finding is not compatible with with theories that posit that surprisal reflects the *full cost* of recognizing a word and integrating it with its context, which are the most common theories of the role surprisal in reading (Hale, 2001; Levy, 2008a; Smith and Levy, 2008; Vigly et al., 2025). Such theories, while more parsimonious than ones that explain recognition and integration through different mechanisms, are hard to reconcile with our results (see also Staub, 2024).

Our results dovetail with recent results that indicate a dissociation between prediction and structural processing in neurolinguistic measurements (Brennan et al., 2020; Hale et al., 2018; Stanojević et al., 2023), and establish that this same dissociation is evident in eye movements during free reading as well. However, it remains to be determined whether the factors driving regressive eye movements and re-reading in our studies reflects the same structure-building processes that these studies have identified.

The results of our large-scale eye-tracking dataset also potentially reveal limits on how much precision this type of data can provide when evaluating computational theories of sentence comprehension. While we found that surprisal may explain the general order of magnitude of forward reading time at the condition level—certainly a lot better than any of the other measures—its fit was still not perfect. Despite the large number of subjects, the uncertainty in the estimates of construction level effects remained quite large. Uncertainty in item-wise estimates was even larger, especially in early measures, making it infeasible to evaluate whether surprisal can explain item-wise variability (which was possible with self-paced reading, Huang et al., 2024). This may be a fundamental limitation of eye-tracking, as getting meaningfully more precise empirical estimates of effect size would likely require many times more subjects than we collected.

Implications for theories of oculomotor control. Finally, our results have implications both for the analysis of eye movements in reading, and theories of oculomotor control in reading. The close match between FORWARD READING TIME and LLM surprisal suggests the continued use of FORWARD READING TIME as a key measurement that could be used when evaluating the alignment between surprisal theory and eye-tracking data. If we are correct in our interpretation of our findings, then using standard gaze duration measures or later regression path measures, which by hypothesis index multiple, dissociable processes, could obscure the

relationship between theory and observation. Given the increasing theoretical importance of rigorous empirical evaluations of surprisal (Brothers and Kuperberg, 2021; Shain et al., 2024; Smith and Levy, 2013; Staub, 2024), the availability of an arguably more 'process pure' measurement of the role of predictability in reading could provide theoretically insightful findings.

6 Conclusion

Reading is characterized by a dynamic pattern of forwards and backwards eye movement, especially during the reading of syntactically challenging sentences, which are associated with longer reading times and a greater rate of rereading. Understanding this pattern can shed light on the cognitive processes of language comprehension. We test two hypotheses as to what drives the eye movement pattern in syntactically challenging sentences: One, that it can be explained as a consequence of word prediction, and another, that it is driven by two separate sets of processes, one involving predictability and another involving a later stage of revising prior structural commitments. Recent advances in language model technology make it possible to rigorously test the extent to which this process reflects word predictability. In a large-scale study that tracked eye movements from 368 human participants as they were reading syntactically challenging sentences, and that used a wide range of predictability estimates from hundreds of language models with different architectures and training setups, we confirm that the movement of the eyes during reading is guided closely by a reader's knowledge of her language. But we find that readers use this knowledge in two distinct, dissociable ways. In forward reading, this knowledge underpins sophisticated predictions of a word's probability in context, which in turn drives early processes of word recognition. But backward reading behavior appears to reflect linguistic knowledge in a different way, and is not just a reflection of this predictability. Instead, difficulty in the structural processes necessary to integrate a recognized word into syntactically challenging contexts causes regressions and targeted rereading behavior above and beyond the effect of predictability. While LLMs have enabled us to make substantial progress in modeling word predictability, our results suggest predictability estimates from those models cannot, on their own, explain human language comprehension, and that they need to be supplemented with equally accurate models of how words are combined into a structured meaning.

7 Materials and methods

7.1 Dataset construction

We collected eye-tracking data from human subjects that read items in four subsets (13 conditions in total)—the garden path subset, the relative clause subset, the attachment ambiguity subset, and the agreement violation subset—along with filler items. We describe all types of sentences in the remainder of this section. Each of the four subsets had 24 items, except for those in the agreement violation subset, which had 18 items. All items are identical to those used in Huang et al. (2024); a detailed list of the items can be found in Appendix I.

7.1.1 Stimuli subsets

Garden path subset. The garden path subset contains three constructions, the Main Verb/Reduced Relative (MV/RR), Noun Phrase/Sentential Complement (NP/S), and Transitive/Intransitive (NP/Z) constructions. The ambiguous condition of each construction has a temporary syntactic ambiguity that is resolved at the critical word (colored in red). The corresponding unambiguous condition includes additional material (colored in blue) that removes the syntactic ambiguity from the ambiguous condition.

1. **MV/RR Ambiguous**: The suspect sent the file deserved further investigation after the murder trial.
2. **MV/RR Unambiguous**: The suspect who was sent the file deserved further investigation after the murder trial.
3. **NP/Z Ambiguous**: Because the suspect changed the file deserved further investigation after the murder trial.
4. **NP/Z Unambiguous**: Because the suspect changed, the file deserved further investigation after the murder trial.
5. **NP/S Ambiguous**: The suspect showed the file deserved further investigation after the murder trial.
6. **NP/S Unambiguous**: The suspect showed that the file deserved further investigation after the murder trial.

In the MV/RR construction, when the verb *sent* is first read it is ambiguous between being the main verb of the sentence or an embedded verb in a relative clause modifying *the suspect*. When readers reach the actual main verb of the sentence, *deserved*, the sentence is disambiguated toward the relative clause interpretation. The addition of *who was* in the control condition makes it clear that the verb *sent* is unambiguously part of a relative clause.

In the NP/Z construction, the verb *changed* can either take a noun phrase complement (changed something) or no complement (zero complement: changed, meaning changed their clothes). Without the comma after *changed*, the following noun phrase *the file* is ambiguous between being a noun phrase complement of *changed* or the subject of another clause. The relativized verb *deserved* disambiguates the sentence in favor of the zero complement interpretation.

In the NP/S construction, the verb *showed* can either take a noun phrase complement (e.g., show something) or a sentential complement (e.g., show that something happened). Without the complementizer *that*, the complement *the file* is initially ambiguous between being a noun phrase complement, or the subject of a sentential complement. The second verb *deserved* disambiguates the sentence in favor of the sentential complement interpretation.

Relative clause subset. The relative clause subset contains sentence pairs with lexically matched object-extracted relative clauses (ORCs) and subject-extracted relative clauses (SRCs).

7. **ORC**: The bus driver who the kids followed wondered about the location of a hotel.

8. **SRC**: The bus driver who followed the kids wondered about the location of a hotel.

Both ORCs and SRCs consist of a determiner (*the*), a noun (*kids*), and a verb (*followed*), which are the three word regions of interest. ORCs are more difficult to process than SRCs in many languages (Lau and Tanaka, 2021).

Attachment ambiguity subset. The attachment ambiguity subset contains sentences that with two noun phrases and a relative clause. The relative clause can modify the nearby noun phrase (low attachment), the more distant noun phrase (high attachment), or either noun phrase (ambiguous), depending on the number of each noun in the noun phrase (colored in blue).

9. **Low Attachment**: In the lobby, Clyde bumped into the chauffeurs of the CEO who is reckless and very unpopular with the company.

10. **High Attachment**: In the lobby, Clyde bumped into the chauffeur of the CEOs who is reckless and very unpopular with the company.

11. **Ambiguous**: In the lobby, Clyde bumped into the chauffeur of the CEO who is reckless and very unpopular with the company.

In the ambiguous condition, the critical word *is* can agree with either noun of the preceding noun phrase (here, either *chauffeur* or *CEO*). Previous studies report faster processing relative clauses where attachment is ambiguous compared to either low attachment or high attachment relative clauses (van Gompel et al., 2005; Traxler et al., 2002, 1998). Our items were modified versions of the stimuli used in Dillon et al. (2019).

Agreement violation subset. The agreement violation subset contains sentences where the verb does not agree in number with the subject:

12. **Violation**: If the supervisor changes, the schedules deserves further inspection by the rest of the staff.

13. **No Violation**: If the supervisor changes, the schedules deserve further inspection by the rest of the staff.

Unlike the sentences in the previous subsets, sentences containing an agreement error are simply ungrammatical, and it is not possible to derive a well-formed syntactic analysis from them, for example, by reanalyzing the critical word *deserves*.

Filler items. The filler items are designed to elicit 'typical' reading in the absence of processing difficulty due to syntactically challenging constructions. To this end, we selected 40 sentences from the Provo corpus (Luke and Christianson, 2018), which contains text from news articles, magazines, and works of fiction.

7.1.2 Data collection and processing

A total of 424 native speakers of English, with normal or corrected to normal vision, participated in the experiment. Pre-registered exclusion criteria include accuracy on the comprehension questions for the filler items (we excluded participants with accuracy lower than 80%,

N=21) and excessive blinking rate on the target region of the critical trials (>25%, N=35), leaving 368 participants' data in the analysis.

Data were collected across four universities, using the same equipment and experimental setup. Participants were seated in a quiet room with a 20 inch LCD monitor positioned 60 cm in front of them. Each sentence was presented on a single line in 12-point monospaced font (1° visual angle fitting about 3–4 characters). Participants were instructed to read each sentence naturally and answer a question that followed, while maintaining their head on the SR Eyelink 1000 tower-mount eye tracker. The movement of one eye was recorded at a sampling rate of 1000 Hz. Three-horizontal-dot calibration was performed at the beginning of the experiment and re-conducted between trials as needed.

Following Huang et al. (2024), the four subsets were distributed into 18 experimental lists using a Latin square design. Each participant only read four unique items in each of the 13 experimental conditions (52 trials). Each list was combined with the same 40 filler items in a single experimental session. The items in the agreement subset always occurred in the last 18 trials of the session (embedded with 10 filler items); this was done because to make the reading of the remaining items as natural as possible, since these items were the only ones that contained grammatical errors. No items from two paired conditions (e.g. NP/S ambiguous and NP/S unambiguous) were presented consecutively.

Data were processed using the UMass eye-tracking Lab software. A fixation whose duration was shorter than 80 ms and within 1 character from the preceding or following fixation was combined with that fixation. We excluded from analysis trials with gaze duration on the critical region of 2000 ms or more.

7.2 Estimating effects due to syntactic disambiguation

7.2.1 Empirical effects from human data

The empirical effects were estimated as the difference in each reading measure at each word region of interest between the conditions in each subset. To this end, Bayesian mixed-effects regression models were fit separately for each combination of subset, reading measure, and word region. Full model formulae can be found in Appendix E.

For the garden path subset, ambiguity was coded using contrast coding (unambiguous: 0, ambiguous: 1). The three constructions were coded using treatment coding, where the first variable `MVRR_vs_NPS` contrasted the MV/RR construction with the NP/S construction (MV/RR: 0,

NP/S: 1), and the second variable `MVRR_vs_NPZ` contrasted the MV/RR construction with the NP/Z construction (MV/RR: 0, NP/Z: 1). The regression models included the main effects of ambiguity, `MVRR_vs_NPS`, and `MVRR_vs_NPZ`, as well as the two-way interactions `Ambiguity:MVRR_vs_NPS` and `Ambiguity:MVRR_vs_NPZ`. The models also included by-participant and by-item random slopes for all main and interaction effects, as well as by-participant and by-item random intercepts.

For the relative clause subset, clause type was coded using contrast coding (SRC: 0, ORC: 1). As the three word regions of interest (i.e. the determiner, noun, and verb) occurred at different linear positions across conditions, word position was first residualized out of each reading measure using a linear mixed-effects model fit to the filler sentences that includes the main effect of position, its by-participant random effect, and a by-participant random intercept. Subsequently, regression models that included the main effect of clause type, by-participant and by-item random slopes for clause type, and by-item random intercepts were fit to the residualized reading measures from each word region.

For the attachment ambiguity subset, ambiguity and attachment location were coded using sum coding (unambiguous: $-\frac{1}{3}$, ambiguous: $\frac{2}{3}$; low: $-\frac{1}{2}$, high: $\frac{1}{2}$). The regression models included the main effects of ambiguity and attachment location, by-participant and by-item random slopes for both main effects, and by-participant and by-item random intercepts.

Finally, for the agreement violation subset, violation was coded using contrast coding (no violation: 0, violation: 1). The regression models included the main effect of violation, by-participant and by-item random slopes for violation, and by-participant and by-item random intercepts.

All regression models were fit using the `brms` package in R (Bürkner, 2017). For the reading time measures (all measures except REGRESSIONS OUT), we use the same priors as Huang et al. (2024). For REGRESSIONS OUT, we set the prior on the intercept to $\mathcal{N}(5.7, 1.5)$, the coefficents to $\mathcal{N}(0, 1)$, and the standard deviation of random effects to $\mathcal{N}(0, 1.5)$. Subsequently, four independent Markov Chain Monte Carlo chains were used to draw 12,000 samples each from the posterior distribution. The first half of the samples were treated as warmup samples and discarded, resulting in a total of 24,000 post-warmup draws across the four chains. The $\hat{R}$ value of all estimated effects were 1.00, which indicates model convergence.

7.2.2 Predicted effects from LLM surprisal

The reading measures predicted by LLM surprisal were estimated using protocols established in previous work for self-paced reading (Arehalli et al., 2022; Huang et al., 2024; van Schijndel and Linzen, 2021), adapted for eye-movement data. First, for each combination of LLM and reading measure, we fit a frequentist (non-Bayesian) mixed-effects regression model to data collected using the filler items, using the `lme4` package in R (Bürkner, 2017). These *filler models* include LLM surprisal, word frequency, word length, and position of the word within the sentence as predictors. All predictors were scaled such that their mean was zero and their standard deviation one. To account for spillover effects, the filler models also include the surprisal, frequency and length of the previous word (where surprisal was estimated by the same LLM). They also included random slopes for surprisal and frequency by participant, and random intercepts for both participants and items. The full formulae for the baseline and surprisal filler models can be found in Appendix E. For filler models predicting REGRESSIONS OUT, a binary response variable, we use a logit link; we use linear link for all other filler models, where the predicted measure is continuous. Subsequently, these filler models were used to calculate the predicted trial-level reading measures for the words of interest in each condition. Finally, the difference in the predicted reading measures across conditions was estimated by fitting regression models with identical specifications to those that were used to estimate empirical effects. For REGRESSIONS OUT, the filler model predicted reading data was expressed as log odds of a regression, and the experimental effects were estimated on the logits using linear regression.

7.3 Poisson regression model of the number of fixations

We used a Poisson regression model to analyze the number of fixations on each word. In this model, the logarithm of the λ parameter of the Poisson distribution is determined by a linear combination of the following predictors of interest. First, ambiguity and source region were each coded using contrast coding (unambiguous: 0, ambiguous: 1; critical: 0, spillover: 1). Additionally, the three target regions were coded using Helmert coding, where the first variable `NV_vs_Prev` contrasted the previous word with the earlier target regions (previous: $-\frac{2}{3}$, verb: $\frac{1}{3}$, noun: $\frac{1}{3}$), and the second variable `N_vs_V` contrasted the verb with the noun that preceded it (previous: 0, verb: $-\frac{1}{2}$, noun: $\frac{1}{2}$). Finally, the three constructions were coded using treatment

coding, where the first variable MVRR_vs_NPS contrasted the Main Verb/Reduced Relative construction with the Direct Object/Sentential Complement construction (MV/RR: 0, NP/S: 1), and the second variable MVRR_vs_NPZ contrasted the Main Verb/Reduced Relative construction with the Transitive/Intransitive construction (MV/RR: 0, NP/Z: 1).

The regression model included the main effects of ambiguity, source region, NV_vs_Prev, N_vs_V, MVRR_vs_NPS, and MVRR_vs_NPZ, as well as the two-way interactions Ambiguity:NV_vs_Prev and Ambiguity:N_vs_V. The model also included by-participant and by-item random slopes for all main and interaction effects, as well as by-participant and by-item random intercepts. After filtering out two outlier observations that exceed 10, a total of 54,862 observations were used to fit this regression model.

We fit this model using the brms package in R (Bürkner, 2017). First, we set the priors on each effect as follows: $t(3, -2.3, 2.5)$ for the intercept, $\mathcal{N}(0, 1)$ for all main effects and interaction effects, $t(3, 0, 2.5)$ for the standard deviation of random effects, and LKJ(2) for the correlation between the random slopes and intercepts. Subsequently, four independent Markov Chain Monte Carlo chains were used to draw 7,500 samples each from the posterior distribution. The first one-third of the samples were treated as warmup samples and discarded, resulting in a total of 20,000 post-warmup draws across the four chains. The $\hat{R}$ value of all estimated effects were 1.00, which indicates model convergence.

Acknowledgments

This project is supported by the National Science Foundation (NSF) under grants BCS-2020914, BCS-2020945, IIS-2504953 and IIS-2504954, and in part through the NYU IT High Performance Computing resources, services, and staff expertise. We thank Adrian Staub and Shravan Vasishth for helpful comments. For assistance in collecting data, we are grateful to Katie Jordan, Samir Kassem, Emily Knick, Lakshi Dutta, Mari Kugemoto, Tomas Acuna, and Hanna Marie Tandle (UMass); Cara Leong, Soo-Hwan Lee, Simone Moeller Krogh, Kaustabh Ghoshal, Auromita Mitra, Nigel Flowers, Julia Cataldo, and Christine Gu (NYU); Seojin Lee and Georgie D'Sanson (JHU); and Anzi Wang and Edna Gutierrez (Colgate).

References

Altmann, G. T. M. (1994). Regression-contingent analyses of eye movements during sentence processing: Reply to Rayner and Sereno. *Memory & Cognition*, 22(3):286–290.

Altmann, G. T. M., Garnham, A., and Dennis, Y. (1992). Avoiding the garden path: Eye movements in context. *Journal of Memory and Language*, 31(5):685–712.

Arehalli, S., Dillon, B., and Linzen, T. (2022). Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning*, pages 301–313.

Bartek, B., Lewis, R. L., Vasishth, S., and Smith, M. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 37(5):1178–1198.

Bicknell, K. and Levy, R. (2010). Rational eye movements in reading combining uncertainty about previous words with contextual probability. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32, pages 1142–1147.

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., and van der Wal, O. (2023). Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 2397–2430.

Botvinick, M. (2007). Multilevel structure in behavior and in the brain: A computational model of Fuster's hierarchy. *Philosophical Transactions of the Royal Society, Series B: Biological Sciences*, 362:1615–1626.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., and Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108(3):624–652.

Brennan, J. R., Dyer, C., Kuncoro, A., and Hale, J. T. (2020). Localizing syntactic predictions using recurrent neural network grammars. *Neuropsychologia*, 146:107479.

Brothers, T. and 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.

Bürkner, P.-C. (2017). brms: An R Package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Charniak, E., Blaheta, D., Ge, N., Hall, K., Hale, J., and Johnson, M. (2000). BLLIP 1987-89 WSJ Corpus Release 1.

Christiansen, M. H. and MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning*, 59(s1):126–161.

Christianson, K., Dempsey, J., Tsiola, A., Deshaies, S.-E. M., and Kim, N. (2024). Retracing the garden-path: Nonselective rereading and no reanalysis. *Journal of Memory and Language*, 137:104515.

Christianson, K., Hollingworth, A., Halliwell, J. F., and Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42(4):368–407.

Clifton Jr., C., Staub, A., and Rayner, K. (2007). Chapter 15 - Eye movements in reading words and sentences. In van Gompel, R. P. G., Fischer, M. H., Murray, W. S., and Hill, R. L., editors, *Eye Movements: A Window on Mind and Brain*, pages 341–371. Elsevier.

Dahan, D. and Ferreira, F. (2019). Language comprehension: Insights from research on spoken language. In Hagoort, P., editor, *Human Language: From Genes and Brains to Behavior*, pages 21–33. MIT Press.

Dillon, B., Andrews, C., Rotello, C. M., and Wagers, M. (2019). A new argument for co-active parses during language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(7):1271–1286.

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2):85–103.

Dyer, C., Kuncoro, A., Ballesteros, M., and Smith, N. A. (2016). Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209.

Ehrlich, S. F. and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655.

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

Fodor, J. D. and Inoue, A. (2000). Garden path repair: Diagnosis and triage. *Language and Speech*, 43(3):261–271.

Frazier, L. and Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, 6:291–325.

Frazier, L. and Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2):178–210.

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, 91(4):1357–1392.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., and Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 32–42.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. (2020). The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint*, arXiv:2101.00027.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., and Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37(1):58–93.

Gibson, E. (2000). The Dependency Locality Theory: A distance-based theory of linguistic complexity. In *Image, Language, Brain: Papers From the First Mind Articulation Project Symposium*, pages 95–126. MIT Press.

van Gompel, R. P., Pickering, M. J., Pearson, J., and Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, 52(2):284–307.

Gu, A. and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint*, arXiv:2312.00752v2.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1195–1205.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Hale, J. (2011). What a rational parser would do. *Cognitive Science*, 35(3):399–443.

Hale, J., Dyer, C., Kuncoro, A., and Brennan, J. (2018). Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2727–2736.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hoover, J. L., Sonderegger, M., Piantadosi, S. T., and O'Donnell, T. J. (2023). The plausibility of sampling as an algorithmic theory of sentence processing. *Open Mind*, 7:350–391.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., and Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744.

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., and Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Inhoff, A. W. and Weger, U. W. (2005). Memory for word location during reading: Eye movements to previously read words are spatially selective but not precise. *Memory and Cognition*, 33(3):447–461.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20(2):137–194.

Just, M. A. and Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Kobzeva, A. and Kush, D. (2024). Grammar and expectation in active dependency resolution: Experimental and modeling evidence from Norwegian. *Cognitive Science*, 48(10):e13501.

Krieger, B., Brouwer, H., Aurnhammer, C., and Crocker, M. W. (2025). On the limits of LLM surprisal as a functional explanation of the N400 and P600. *Brain Research*, 1865:149841.

Lau, E. and Tanaka, N. (2021). The subject advantage in relative clauses: A review. *Glossa: A Journal of General Linguistics*, 6(1):34.

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

Levy, R. (2008b). Modeling the effects of memory on human online sentence processing with particle filters. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 937–944.

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In van Gompel, R. P. G., editor, *Sentence Processing*, pages 78–114. Psychology Press.

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

Lewis, R. L. (1998). Reanalysis and limited repair parsing: Leaping off the garden path. In Fodor, J. D. and Ferreira, F., editors, *Reanalysis in Sentence Processing*, pages 247–285. Springer Dordrecht.

Lewis, R. L. and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Luke, S. G. and Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

von der Malsburg, T. and Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, 65(2):109–127.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2016). Pointer sentinel mixture models. *arXiv preprint*, arXiv:1609.07843.

Meseguer, E., Carreiras, M., and Clifton Jr., C. (2002). Overt reanalysis strategies and eye movements during the reading of mild garden path sentences. *Memory & Cognition*, 30(4):551–561.

Mitchell, D., Shen, X., Green, M., and Hodgson, T. (2008). Accounting for regressive eye-movements in models of sentence processing: A reappraisal of the selective reanalysis hypothesis. *Journal of Memory and Language*, 59(3):266–293.

Molinaro, N., Barber, H. A., and Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *Cortex*, 47(8):908–930.

Ness, T., Langlois, V. J., Kim, A. E., and Novick, J. M. (2025). The state of cognitive control in language processing. *Perspectives on Psychological Science*, 20(2):219–240.

Noji, H. and Oseki, Y. (2021). Effective batching for recurrent neural network grammars. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4340–4352.

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

Norris, D. (2009). Putting it all together: A unified account of word recognition and reaction-time distributions. *Psychological Review*, 116(1):207–219.

Oh, B.-D. and Linzen, T. (2025). To model human linguistic prediction, make LLMs less superhuman. *arXiv preprint*, arXiv:2510.05141v1.

Oh, B.-D. and Schuler, W. (2023a). Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921.

Oh, B.-D. and Schuler, W. (2023b). 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.

Oh, B.-D., Yue, S., and Schuler, W. (2024). Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2644–2663.

Paape, D. and Vasishth, S. (2022). Conscious rereading is confirmatory: Evidence from bidirectional self-paced reading. *Glossa Psycholinguistics*, 1(1):20.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Rayner, K. and Sereno, S. C. (1994). Regressive eye movements and sentence parsing: On the use of regression-contingent analyses. *Memory & Cognition*, 22(3):281–285.

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

Reichle, E. D. and Sheridan, H. (2015). E-Z Reader: An overview of the model and two recent applications. In Pollatsek, A. and Treiman, R., editors, *The Oxford Handbook of Reading*, pages 277–290. Oxford University Press.

Reichle, E. D., Warren, T., and McConnell, K. (2009). Using E-Z Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, 16(1):1–21.

Richter, E. M., Engbert, R., and Kliegl, R. (2006). Current advances in SWIFT. *Cognitive Systems Research*, 7(1):23–33.

Ryskin, R. and Nieuwland, M. S. (2023). Prediction during language comprehension: What is next? *Trends in Cognitive Sciences*, 27(11):1032–1052.

van Schijndel, M. and Linzen, T. (2018). Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 40, pages 2603–2608.

van Schijndel, M. and Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45:e12988.

van Schijndel, M., Mueller, A., and Linzen, T. (2019). Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5831–5837.

Schotter, E. R. and Dillon, B. (2025). A beginner's guide to eye tracking for psycholinguistic studies of reading. *Behavior Research Methods*, 57(2):68.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Shain, C. (2024). Word frequency and predictability dissociate in naturalistic reading. *Open Mind*, 8:177–201.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., and 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.

Smith, N. J. and Levy, R. (2008). Optimal processing times in reading: A formal model and empirical investigation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30, pages 595–600.

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

Stanojević, M., Brennan, J. R., Dunagan, D., Steedman, M., and Hale, J. T. (2023). Modeling structure-building in the brain with CCG parsing and large language models. *Cognitive Science*, 47(7):e13312.

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, 116(1):71–86.

Staub, A. (2011). Word recognition and syntactic attachment in reading: Evidence for a staged architecture. *Journal of Experimental Psychology: General*, 140(3):407–433.

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, 9(8):311–327.

Staub, A. (2024). Predictability in language comprehension: Prospects and problems for surprisal. *Annual Review of Linguistics*, 11:17–34.

Staub, A., Dillon, B., and Clifton Jr., C. (2017). The matrix verb as a source of comprehension difficulty in object relative sentences. *Cognitive Science*, 41(S6):1353–1376.

Stern, M., Fried, D., and Klein, D. (2017). Effective inference for generative neural parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1695–1700.

Stowe, L. (1986). Parsing wh-constructions: Evidence for on-line gap location. *Language and Cognitive Processes*, 1:227–245.

Sturt, P., Pickering, M. J., and Crocker, M. W. (1999). Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, 40(1):136–150.

Tabor, W., Galantucci, B., and Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, 50(4):355–370.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Traxler, M. J., Morris, R. K., and Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, 47(1):69–90.

Traxler, M. J., Pickering, M. J., and Clifton Jr., C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, 39(4):558–592.

Vasishth, S., Suckow, K., Lewis, R. L., and Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, 25(4):533–567.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 6000–6010.

Vigly, J., Qian, P., Sonderegger, M., and O'Donnell, T. (2025). Comprehension effort as the cost of inference. *PsyArXiv preprint*, 2498w_v1.

Vonk, W. and Cozijn, R. (2003). Chapter 15 - On the treatment of saccades and regressions in eye movement measures of reading time. In Hyönä, J., Radach, R., and Deubel, H., editors, *The Mind's Eye*, pages 291–311. North-Holland.

Wagers, M. and Dillon, B. (2025). Sentence processing. In Frank, M. C. and Majid, A., editors, *Open Encyclopedia of Cognitive Science*. MIT Press.

Warren, T. and McConnell, K. (2007). Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. *Psychonomic Bulletin & Review*, 14(4):770–775.

Warren, T., Reichle, E. D., and Patson, N. D. (2011). Lexical and post-lexical complexity effects on eye movements in reading. *Journal of Eye Movement Research*, 4(1):1–10.

Wilcox, E., Vani, P., and Levy, R. (2021). A targeted assessment of incremental processing in neural language models and humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 939–952.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., and Levy, R. P. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 42, pages 1707–1713.

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

Wilcox, E. G., Pimentel, T., Meister, C., and Cotterell, R. (2024). An information-theoretic analysis of targeted regressions during reading. *Cognition*, 249:105765.

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., and Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Appendices

A Overview of the evaluated LLMs

We generated predictions of reading measures from a total of 407 LLMs from six model families that span different architectures, training objectives, number of parameters, and the amount and composition of training data. Where we report just a single measure per model family in the main paper, we show the effects predicted by the LLM from each of the six model families that achieved the best fit to the *filler* reading data across the four reading measures (that is, models were not selected based on their fit to the critical items). The goodness-of-fit to the filler reading data was measured as the increase in likelihood due to including LLM surprisal over a regression model that contains only the baseline predictors.

The model families we evaluated are:

Pythia (Biderman et al., 2023): A family of transformer-based (Vaswani et al., 2017) autoregressive language models trained on the next-word prediction objective over ∼300B tokens from the Pile dataset (Gao et al., 2020). The Pythia family spans a wide range of model parameters (eight sizes with ∼70M, ∼160M, ∼410M, ∼1B, ∼1.4B, ∼2.8B, ∼6.9B, ∼12B parameters), with checkpoints of each model available at various stages of training, when they have seen different amounts of training data. We evaluated models of all eight sizes at 19 stages during training (when models have seen 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1000, 2000, 4000, 8000, 16000, 32000, 64000, 128000, 143000 training batches of ∼2M tokens), resulting in a total of 152 models.

GPT-2 (Radford et al., 2019): A family of transformer-based autoregressive language models trained on the next-word prediction objective over 40GB of internet text. We evaluated the four officially released GPT-2 models, which vary in their number of parameters (∼124M, ∼355M, ∼774M, ∼1.6B parameters).

Mamba (Gu and Dao, 2023): A family of autoregressive language models based on the Mamba architecture (Gu and Dao, 2023), which rely on recurrence-like operations that are constrained to be parallelizable. These models were trained on the next-word prediction objective over ∼300B tokens from the Pile dataset (Gao et al., 2020). We evaluated models of all five sizes released by the authors, with ∼130M, ∼370M, ∼790M, ∼1.4B, and ∼2.8B parameters.

LSTM (vS) (van Schijndel et al., 2019): A family of language models based on the long short-term memory architecture (LSTM; Hochreiter and Schmidhuber, 1997), trained on the

next-word prediction objective over the WikiText-103 dataset (Merity et al., 2016). We evaluated all 125 released models; these included all combinations of five different sizes (with 100, 200, 400, 800, 1600 hidden units in each of the two LSTM layers), five different amounts of training data (2M, 10M, 20M, 40M, 80M words), and five distinct subsets of the WikiText-103 dataset.

RNNG: A family of recurrent neural network grammar (RNNG; Dyer et al., 2016; Noji and Oseki, 2021) models. These are recurrent neural networks that explicitly build syntactic structure by jointly predicting parsing operations and the upcoming word (unlike standard recurrent neural networks, which only predict the upcoming word). These models require a parsed version of the training corpus. We trained 10 models on a 42M-word machine-parsed subset of the BLLIP corpus (Charniak et al., 2000), using all combinations of five different random seeds and two parsing methods (top-down parsing and left-corner parsing). We then calculated surprisal estimates from these 10 models using word-synchronous beam search (Stern et al., 2017) with 12 different beam sizes (1, 2, 3, 4, 5, 10, 25, 50, 100, 250, 500, 1000 states), resulting in a total of 120 models.

LSTM (G) (Gulordava et al., 2018): A two-layer LSTM language model with 650 hidden units on each layer, trained on the next-word prediction objective over 90M tokens from the English Wikipedia.

B Why we use regression-contingent eye-movement measures

The eye-movements measures we use are FORWARD READING TIME, REGRESSIVE GAZE DURATION, and REGRESSIVE GO-PAST TIME, which are regression-contingent (Altmann, 1994; Altmann et al., 1992; Rayner and Sereno, 1994; Vonk and Cozijn, 2003) versions of the common measures GAZE DURATION and GO-PAST TIME (Schotter and Dillon, 2025). The "classic" (non-regression-contingent) measure GAZE DURATION is the sum of all first-pass fixations on w_i before it is exited in either direction, and is equivalent to FORWARD READING TIME when w_i is exited to the right, and REGRESSIVE GAZE DURATION when existed to the left. Likewise, the non-regression-contingent measure GO-PAST TIME is the sum of all fixations on any word from the first fixation on w_i until any word $w_{>i}$ is fixated, and is equivalent to FORWARD READING TIME when w_i is exited to the right, and REGRESSIVE GO-PAST TIME when exited to the left.

We chose to analyze the regression-contingent versions of these more common measures

for several reasons. Empirically, GAZE DURATION is known to be significantly longer when words are exited via a forward eye movement than when they are exited via a regression (Altmann, 1994; Rayner and Sereno, 1994; Vonk and Cozijn, 2003). One intuition for this pattern is that error-driven saccades will occur as soon as a processing error is detected, possibly before w_i is fully processed, while forward saccades from w_i occur after w_i has been fully processed. In addition, models of eye-movement control during reading such as like E-Z Reader 10 (Reichle et al., 2009) argue that forward saccades are often triggered when only the first of two lexical processing stages is complete, and that processing errors and regressive saccades due to integration failure can manifest on the following word (w_{i+1}). Processing difficulty is known to manifest as increases in REGRESSIONS OUT and as increases in GAZE DURATION (Clifton Jr. et al., 2007; Reichle et al., 2009; Staub, 2011), but it is also the case empirically that increases in REGRESSIONS OUT are associated with *decreases* in GAZE DURATION (Altmann, 1994; Altmann et al., 1992; Rayner and Sereno, 1994; Vonk and Cozijn, 2003). Thus, differences in the rate of regressive saccades across conditions can mask an underlying positive effect in GAZE DURATION (Altmann et al., 1992), or create a negative effect when none is in fact present, if GAZE DURATION is aggregated across both regressive and non-regressive trials. To avoid this confound, we separate this measure into two—FORWARD READING TIME and REGRESSIVE GAZE DURATION—based on whether or not a regression has occurred.

The non-regression-contingent GO-PAST TIME measure likewise confounds the rate of regressions with the duration of rereading. If we were to observe an increase in GO-PAST TIME, it would not be clear whether this effect is simply driven by an increase in the rate of regressions, or whether it also reflects an increase in the duration of rereading given that readers have regressed. Because our goal is to assess the ability of surprisal to differentially explain distinct reading measures—separating the rate of regression from rereading time given that a regression has occurred—we analyze GAZE DURATION and GO-PAST TIME in a regression-contingent manner, which eliminates the confound with regression rate.

B.1 Results from non-gaze-contingent measures

We next present results from an analysis of the non-regression-contingent aggregated measures (GAZE DURATION and GO-PAST TIME). Results from all constructions and all measures, including the non-regression-contingent measures, are summarized in Figure 8.

B.1.1 Gaze duration

In the garden path constructions, empirical effects in GAZE DURATION appear to reflect a mixture of the patterns in REGRESSIVE GAZE DURATION and FORWARD READING TIME. In all garden path constructions, the effect is localized to the critical word (MV/RR: 30 ms [17–44 ms], NP/Z: 24 ms [9–40 ms], NP/S: 17 ms [3–31 ms]). The predicted effects of several LLM families overlapped with the empirical effects in the NP/Z and NP/S constructions, the closest of which was `GPT-2`, with a predicted effect of 14 ms [8–19 ms] in the NP/Z construction and 7 ms [2–12 ms] in the NP/S construction. All LLMs underpredicted the magnitude of the empirical ambiguity effect in the MV/RR construction, with the largest LLM-predicted effect being 9 ms [6–13 ms] from `GPT-2`. There was no detectable difference in the magnitude of the empirical ambiguity effect across constructions (MV/RR vs. NP/Z: [-25–13 ms], MV/RR vs. NP/S: [-32–5 ms]).

In the relative clause subset, there was no detectable empirical effect at the determiner region [-2–22 ms], but there was a negative effect (i.e. a subject relative clause advantage) at the noun region of -39 ms -[51–27 ms]. The LLM-predicted effect of several models at this region was in the opposite direction of the empirical effect. There was also an empirical effect at the verb region of 17 ms [5–30 ms], which overlapped with the predictions of several LLMs, though the LLM-predicted effect appears to be driven by baseline predictors rather than surprisal, as LLM predictions did not differ substantially from those of the baseline model. The negative empirical effect at the noun region is particularly puzzling given that an empirical effect in the opposite direction at this region is observed in both REGRESSIONS OUT and GO PAST TIME. An identical pattern was observed by Staub (2010), which used identical materials to those in the present study. Staub suggests that this pattern may be driven by the increased rate of regressions at the noun region, as GAZE DURATION tends to be shorter when fixations on a word are followed by a regression compared to a forward saccade. Indeed, when we control for regressions as in our regression-contingent analyses, we show that the magnitude of the effect at the noun in both FORWARD READING TIME and REGRESSIVE GAZE DURATION is substantially reduced. The remaining effect in the regression contingent analyses may still be driven by an increase in regressions at the PREVIOUS word. The regression-contingent analyses we report in the main body of the paper ensure that any effect driven by rates of regressions manifests only in REGRESSIONS OUT, and does not confound the other measures.

B.1.2 Go-past time

At the critical region of the garden path subset, GO-PAST durations were 262 ms [196–325 ms] higher in the ambiguous conditions (757 ms [684–830 ms]) than the unambiguous conditions (467 ms [436–499 ms]) when averaged across the three constructions. As with REGRESSIVE GO-PAST TIME, all LLMs drastically underpredict the magnitude of this effect. The largest LLM-predicted effect was 24 ms [18–30 ms] from the `Pythia` model. The ambiguity effect was significantly larger in MV/RR construction (276 ms [209–340 ms]) than in the NP/S construction (177 ms [88–266 ms]), but there was no evidence for a difference between the ambiguity effect in the MV/RR and NP/Z (332 ms [228–436 ms]) constructions. Among the LLMs, only the `RNNG` model predicts this difference across constructions. At the first spillover word of the garden path subset, there was a large ambiguity effect of 415 ms [308–514 ms] across all constructions. This effect was significantly larger in the MV/RR construction (544 ms [436–644 ms]) and the NP/Z construction (523 ms [380–644 ms]) than the NP/S construction (178 ms [50–303 ms]). The LLMs once again drastically underpredict the magnitude of the ambiguity effects at the spillover word. The largest predicted ambiguity effect at the spillover word was 33 ms [24–41 ms] in `GPT-2`, twelve times smaller than the empirical effect. In the garden path constructions, the empirical pattern of results in GO-PAST TIME was largely similar to that observed in REGRESSIVE GO-PAST TIME, suggesting that the empirical effects in GO-PAST TIME are not just driven by increased *frequency* of regressions in the ambiguous condition, but also increased *duration* of rereading given that a regression has occurred. Our analysis in the "Targets of regressive eye movements" section suggests that this increased duration reflects selective reanalysis of the comprehender's parse of the sentence.

In the relative clause subset, there was a positive empirical effect (i.e. a subject relative clause advantage) at the determiner region of 124 ms [88–159 ms]. The largest LLM-predicted effect was only 20 ms [4–36 ms] from the `Pythia` model. There was also an empirical effect in the same direction at the noun region of 191 ms [153–228 ms]. The effect at this region was also underpredicted by LLM surprisal. The largest LLM-predicted effect at this region was 41 ms [33–49 ms] from `GPT-2`. The effect at the verb region trended in the negative direction, but its 95% credible intervals overlapped with zero. The LLM-predicted effects at the verb region were all positive, but this appears to be driven by the baseline predictors, as the baseline model predictions did not differ substantially from the predictions of the LLM surprisal models.

C Analysis of filler models

C.1 Filler model surprisal coefficients

The current word surprisal coefficients in the best filler model from each LLM family are presented in Figure 7. We report both *raw* surprisal coefficients (i.e. the increase in milliseconds of reading, or increase in the log odds of a regression, associated with a one bit increase in surprisal) and *scaled* surprisal coefficients (i.e. the number of standard deviations increase in the reading measure associated with one standard deviation increase in surprisal). Scaled coefficients generally quantify the strength of association between surprisal and the reading measure. We present scaled coefficients to allow for direct comparison across LLMs and measures. We find that scaled surprisal coefficients are consistently larger in FORWARD READING TIME than in REGRESSIVE GAZE or REGRESSIVE GO-PAST, suggesting that surprisal is more strongly associated with FORWARD READING TIME than the other measures.

C.2 Filler model predictive power

We also quantify the predictive power of LLM surprisal in each of the four measures. A standard measure of the goodness-of-fit for a regression model is the log-likelihood the response data given the model. To evaluate how much a single predictor improves a model's fit to the data, we can take the difference in the log-likelihood (ΔLL) of a regression model with only baseline predictors, and a model that also includes the predictor of interest (Wilcox et al., 2020). In this case, the response data is reading data from a particular eyetracking measure, and the predictor of interest is surprisal from a particular LLM. Concretely, the predictive power of surprisal in a given LLM and reading measure is the difference in log likelihood (ΔLL) of the filler reading data in a full regression model with both surprisal and baseline predictors, and a regression model with only baseline predictors.

To contextualize the predictive power of surprisal, we also compute a theoretical upper-bound ΔLL for each reading measure, which takes into account the intrinsic noise of the measure, and the fact that a single word-level difficulty estimate like surprisal cannot account for participant-level variance and trial-level variance. To compute this upper-bound, we follow methods for computing variance ceilings from prior work (Huang et al., 2024; Schrimpf et al., 2021). First, we randomly split our 368 participants into two equally sized subsets, A and B. For each measure, we compute the per-word mean reading time or regression rate across all

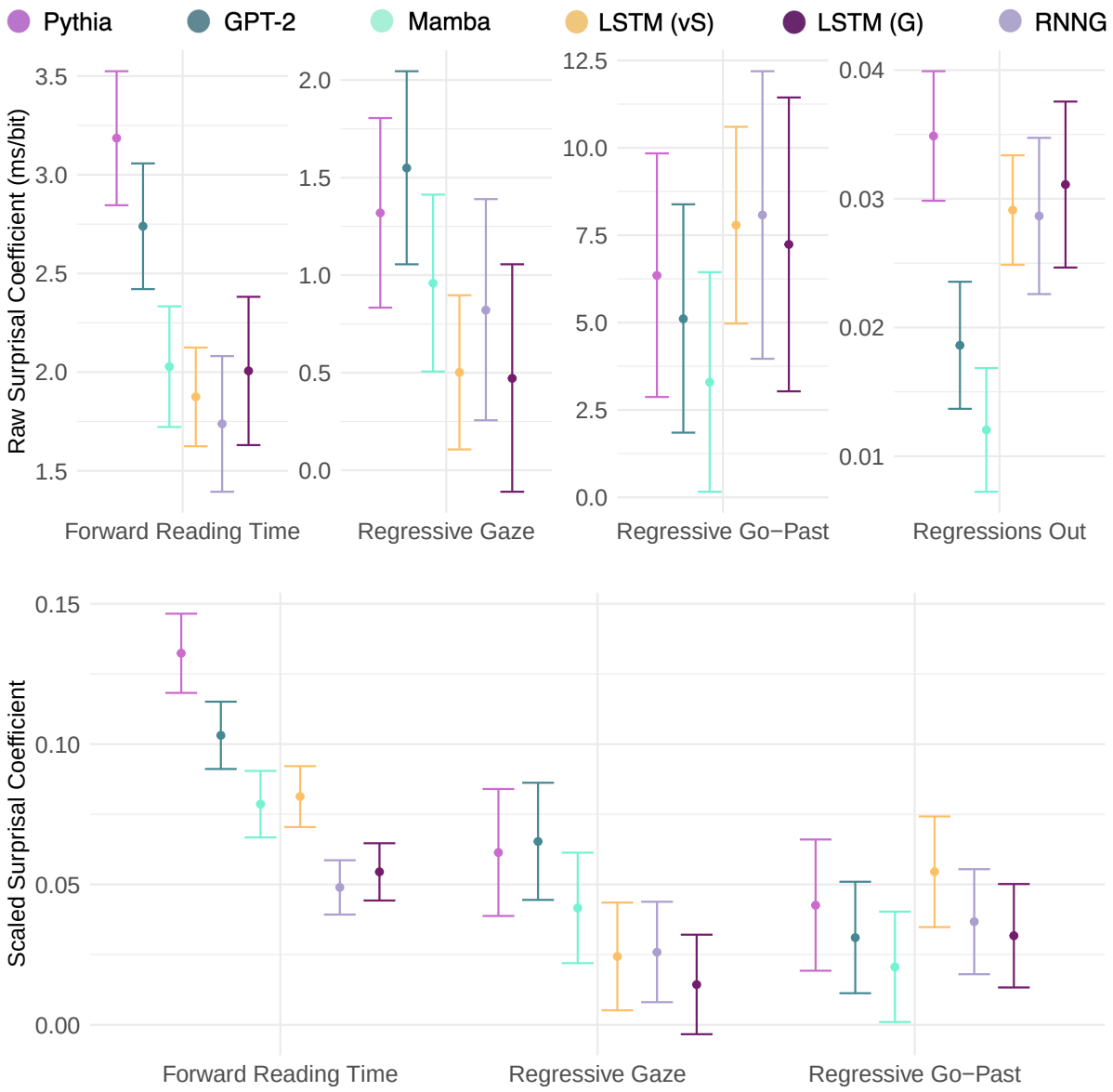


Figure 7: Current word raw surprisal coefficients (top) and scaled surprisal coefficients (bottom) from filler models of six representative LLMs, grouped by reading measure. Each point in this figure represents a coefficient from a single filler model, which is trained to predict one of the four reading measures using surprisal values from a single LLM. The raw coefficients quantify the expected increase in milliseconds of reading time (or log-odds of a regression in the case of REGRESSIONS OUT) of a word as that word's surprisal increases by one bit. The scaled coefficients quantify the expected increase in standard deviations of reading times as surprisal increases by one standard deviation. Error bars represent bootstrapped 95% confidence intervals. Note that only scaled surprisal coefficients can be directly compared across LLMs and measures.

participants in subset A, and then use this as a predictor instead of surprisal in a new filler model fit to reading data from all participants in subset B; see Section E for the upper-bound filler model formula. We then compute the ΔLL of the subset B data between the full filler model and a baseline filler model fit to the same data. Next, we repeat this process by comput-

ing word-level mean reading times from subset B and fitting the filler model to the subset A data. Finally we sum the ΔLL from both subsets to get the total ΔLL. This process is repeated for 15 iterations with different random participant splits. We report the per-observation ΔLL, averaged across these 15 iterations.

The results are summarized in graphically in Figure 3 and discussed in Section 3 of the main body of this work.

D Results from other constructions

D.1 Subject-verb agreement violation

We find no evidence in this subset for any effects of agreement violations in GAZE DURATION, or in its regression contingent counterparts of FORWARD READING TIME and REGRESSIVE GAZE DURATION. We do, however, find an effect in REGRESSIONS OUT. The rate at which readers regressed was 50% [24–79%] higher in the agreement violation condition than the no-violation condition. Interestingly, the predicted effect from `GPT-2` of 29% [26–31%] overlaps with the human results. This was the second largest effect in REGRESSIONS OUT predicted by `GPT-2` across all constructions (behind only the predicted ambiguity effect in the NP/Z construction), yet the magnitude of the empirical effect was much smaller than many of the effects seen in the the other constructions. It is possible that the trigger for regressions in agreement violation configurations might be distinct from that of garden path configurations, and that the former might be better explained by word surprisal than the latter.

Results for GO-PAST DURATION and REGRESSIVE GO-PAST DURATION are more consistent with the general pattern observed in the other constructions, where LLM surprisal cannot explain the full magnitude of the effect. In GO-PAST DURATION, there was an empirical effect of 168 ms [94–240 ms] at the critical word, and a smaller effect of 69 ms [8–130 ms] at the spillover word. The closest LLM-predicted effect at the critical word was 46 ms [42–49 ms] from `GPT-2`. No LLM predicted the existence of an effect at the spillover word. In REGRESSIVE GO-PAST DURATION, there was an empirical effect at the critical word of 272 ms [102–433 ms], which was larger than that observed in the non-regression-contingent measure. Interestingly two LLMs predicted any effect at the critical word, the larger of which was 26 ms [13–41 ms] from the `LSTM (vS)`. The LLM predictions in REGRESSIVE GO-PAST DURATION were not only well below the empirical effect magnitude, but were also smaller than the predicted effect in

the non-regression contingent GO PAST measure, opposite to the empirical pattern.

D.2 Attachment ambiguity

Evidence for ambiguity effects in the attachment ambiguity subset was limited. Empirically, we observe ambiguity effects only in REGRESSIONS OUT and GO-PAST DURATION, and only at the spillover word. These effects were relatively small in magnitude compared effects in the other constructions. Regression rates were 23% [3–47%] higher in the unambiguous condition than the ambiguous condition, and GO-PAST times were 80 ms [19–139 ms] longer. This was a similar pattern to the self-paced reading results reported in Huang et al. (2024), where no ambiguity effect was found in the low attachment condition, and the ambiguity effect in the high attachment condition was smaller in magnitude than the effects observed in any other constructions. In REGRESSIONS OUT, 95% credible intervals for the LLM predicted effects overlapped with those of human effects in all LLMs. The largest predicted effect was 10% [8–13%] for `GPT-2`. In GO-PAST TIME, only the effect predicted by `GPT-2` (19 ms [14–24 ms]) overlapped with the empirical effect.

E Regression model formulae

Model formulae are provided in `R` formula notation. The variable to the left of the "$\sim$" symbol is the response variable, and all variables to its right are predictors. The "$a \cdot b$" symbol denotes that a and b are combined both additively and multiplicitively. Random effects are denoted with "|" for uncorrelated random intercepts and slopes, and "||" for correlated random slopes. The random effect structure appears to the left of the "|" or "||" symbol, and the grouping variable appears to the right. Random intercepts are denoted with "1", and random slopes are denoted with their variable name. "$Surp(w_i)$ denotes the surprisal of the ith word of a sentence. Omitted from the formula below is the fact that all predictors are scaled to a mean of one and a standard deviation of zero. For the upper-bound filler models, $Measure_B$ denotes a reading measure from only participants in subset B, and $\overline{Measure_A}(w_i)$ is the mean of this measure at the ith word of the sentence across all participants in subset A.

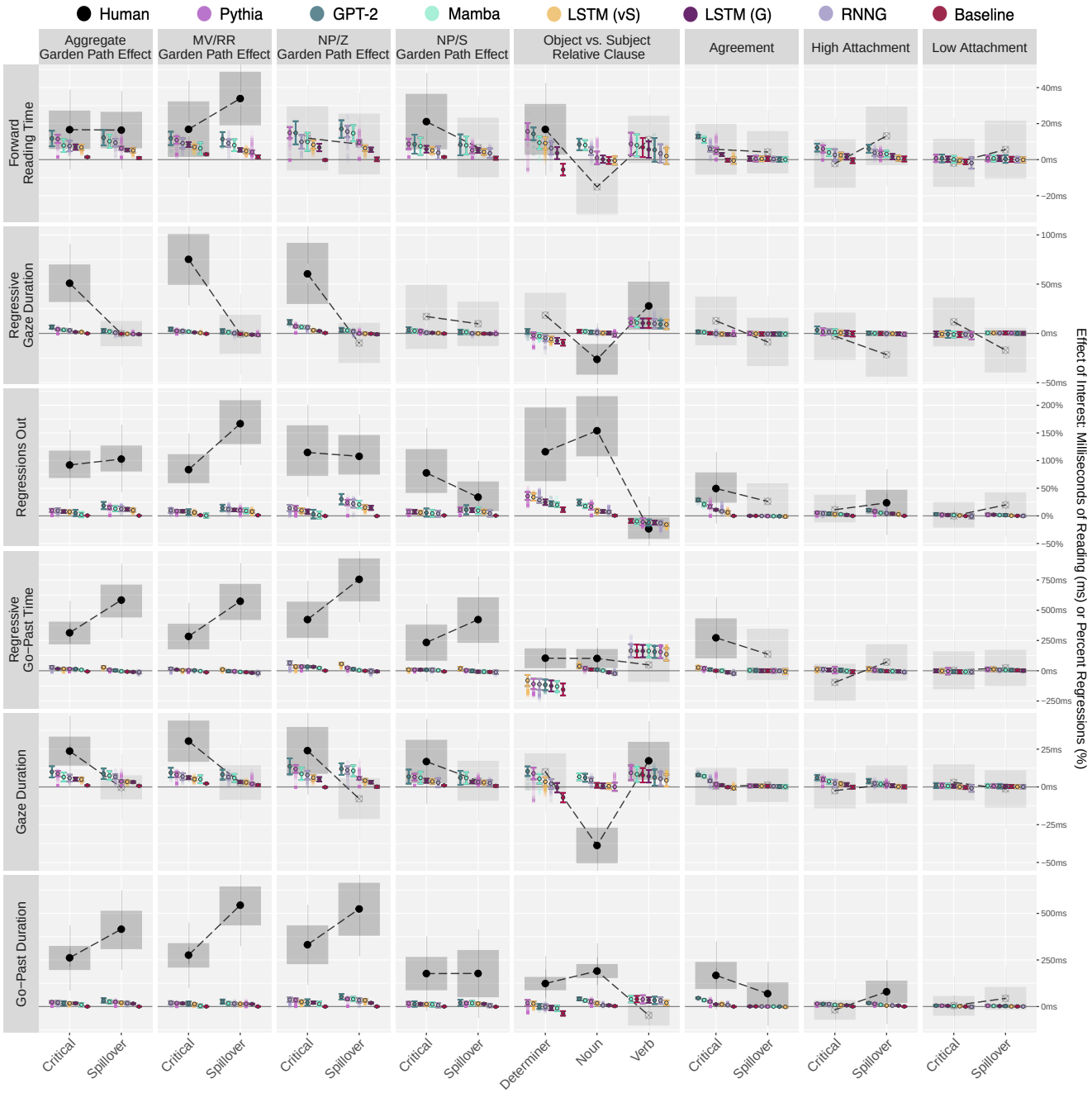


Figure 8: Empirical (human) and predicted (LLM) effects of interest at the regions of interest in all constructions (columns) and all measures (rows). Human effect estimates are represented with black points for the mean, with shaded regions representing 95% credible intervals. Empirical null effects are denoted with lighter shading and an X on the mean estimate point. Predicted effect sizes from the best LLM within each architecture (measured by goodness-of-fit to the filler sentences), and are sorted from largest to smallest, with error bars denoting 95% credible intervals. Mean effect estimates from all 407 LLM-derived surprisal estimates are displayed as square semi-transparent colored points without error bars.

Full filler model formula:

$$\begin{aligned} Measure \sim\ & Surp(w_i) + Surp(w_{i-1}) + \\ & \big(Logfreq(w_i) \cdot Length(w_i)\big) + \big(Logfreq(w_{i-1}) \cdot Length(w_{i-1})\big) + \\ & Position(w_i) + \\ & \big(1 + Surp(w_i) + Surp(w_{i-1}) + Logfreq(w_i) + Logfreq(w_{i-1}) \parallel participant\big) + \\ & (1 \mid item) \end{aligned}$$

Baseline filler model formula:

$$\begin{aligned} Measure \sim\ & \big(Logfreq(w_i) \cdot Length(w_i)\big) + \big(Logfreq(w_{i-1}) \cdot Length(w_{i-1})\big) + \\ & Position(w_i) + \\ & \big(1 + Logfreq(w_i) + Logfreq(w_{i-1}) \parallel participant\big) + \\ & (1 \mid item) \end{aligned}$$

Upper-bound full filler model formula:

$$\begin{aligned} Measure_B \sim\ & \overline{Measure_A}(w_i) + \overline{Measure_A}(w_{i-1}) + \\ & \big(Logfreq(w_i) \cdot Length(w_i)\big) + \big(Logfreq(w_{i-1}) \cdot Length(w_{i-1})\big) + \\ & Position(w_i) + \\ & \big(1 + \overline{Measure_A}(w_i) + \overline{Measure_A}(w_{i-1}) + \\ & \ Logfreq(w_i) + Logfreq(w_{i-1}) \parallel participant\big) + \\ & (1 \mid item) \end{aligned}$$

Upper-bound baseline filler model formula:

$$\begin{aligned} Measure_B \sim\ & \big(Logfreq(w_i) \cdot Length(w_i)\big) + \big(Logfreq(w_{i-1}) \cdot Length(w_{i-1})\big) + \\ & Position(w_i) + \\ & \big(1 + Logfreq(w_i) + Logfreq(w_{i-1}) \parallel participant\big) + \\ & (1 \mid item) \end{aligned}$$

Agreement subset model formula:

$$\begin{aligned} Measure \sim\ & agree + \\ & (1 + agree \parallel subject) + \\ & (1 + agree \parallel item) \end{aligned}$$

Attachment ambiguity subset model formula:

$$\begin{aligned} Measure \sim\ & ambiguity + height + \\ & (1 + ambiguity + height \parallel subject) + \\ & (1 + ambiguity + height \parallel item) \end{aligned}$$

Garden path subset model formula:

$$\begin{aligned} Measure \sim\ & ambiguity \cdot (NPS + NPZ) + \\ & \big(1 + ambiguity \cdot (NPS + NPZ) \parallel subject\big) + \\ & \big(1 + ambiguity \cdot (NPS + NPZ) \parallel item\big) \end{aligned}$$

Relative clause subset word position residualization model formula:

$$\begin{aligned} Measure \sim\ & Position(w_i) + \\ & \big(1 + Position(w_i) \parallel subject\big) + \\ & \big(1 \mid item\big) \end{aligned}$$

Relative clause subset model formula:

$$\begin{aligned} Measure\_corrected \sim\ & RC\_type + \\ & (0 + RC\_type \parallel subject) + \\ & (1 + RC\_type \parallel item) \end{aligned}$$

F Distributions of surprisal values

Histograms of surprisals assigned to filler and garden path sentences in the model from each LLM family that produced the highest ΔLL (i.e. highest increase in the likelihood of the empirical data compared to a baseline model that did not include LLM surprisal) are provided in Figure 9. In general, there is substantial overlap in the distribution of surprisal values in the filler sentences and the garden path sentences across all LLM families. In all LLM families except `LSTM (G)` there is at least one word in the filler sentences with a higher surprisal than any of the critical words in the garden path sentences. This means that the failure of LLM surprisal to capture the magnitude of syntactic processing effects in the regressive measures cannot be explained by a difference in the distribution of surprisal values that the filler models are fit to, and those at the critical region of syntactically challenging sentences.

G Influence of LLM size and amount of training data on garden path effect magnitudes

The number of parameters (size) of the LLMs we evaluated varied in four model families: `GPT-2`, `Mamba`, `LSTM (vS)`, and `Pythia`. The amount of training data only varied systematically in `LSTM (vS)` and `Pythia`. We evaluate the relationship between model size/training data size and the magnitude of predicted garden path effects through simple visual inspection of plots

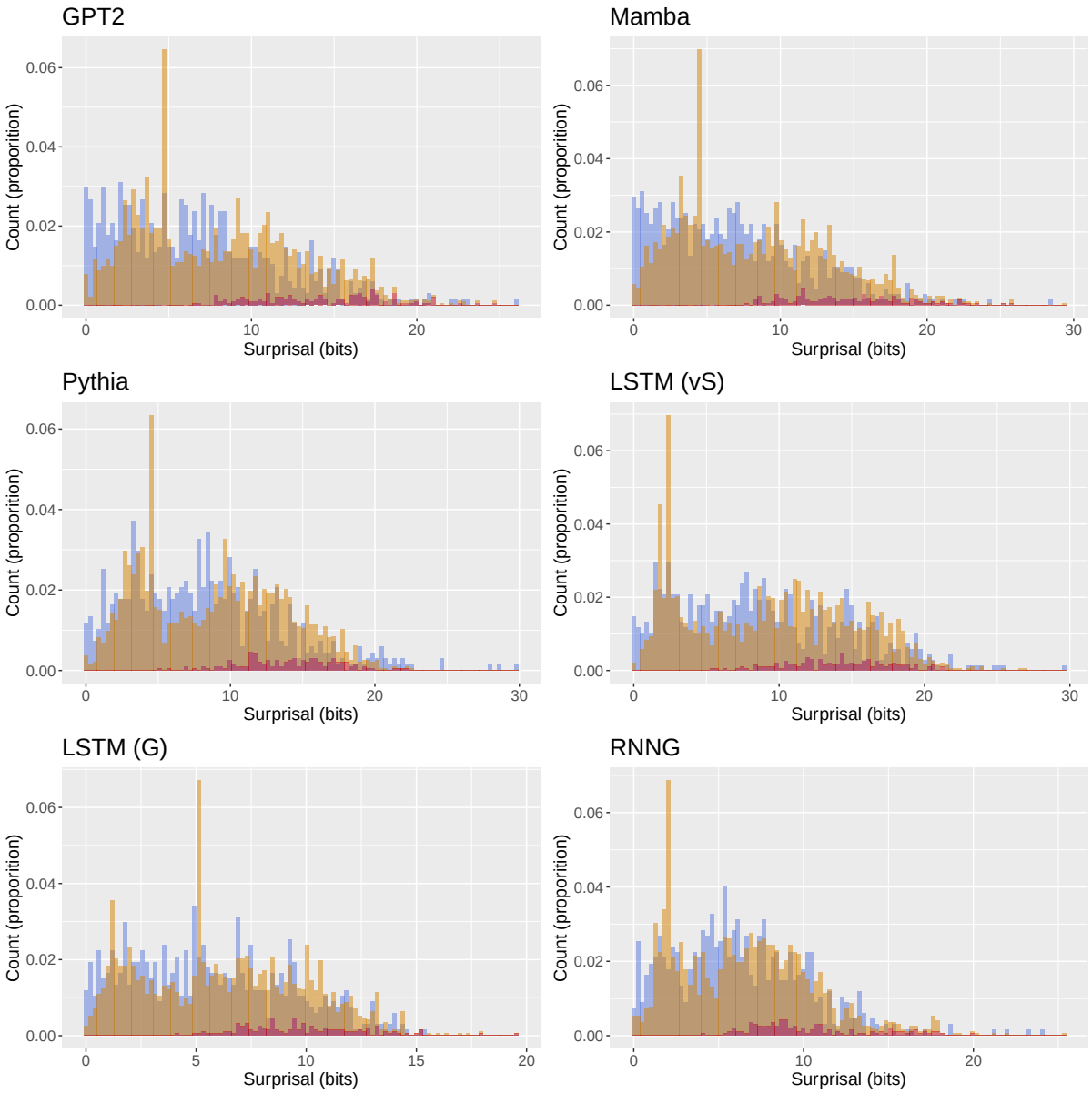


Figure 9: Distribution of surprisal values assigned to filler words (blue) and garden path sentence words, where words in the critical region are colored red, and words outside the critical region are colored orange.

within model families. Garden path effects reported for each model were estimated using frequentist linear mixed effects models fit with the `lme4` package in `R` with default settings. The model formulas were identical in specification to the Bayesian models presented in the main body of the paper.

GPT-2 family. In all measures, there appears to be a negative correlation between model size and the magnitude of predicted GP effects, where larger garden path effects are predicted by the models with fewer parameters. Effects in all measures are summarized in Figure 10 (top).

Mamba family. In FORWARD READING TIME and REGRESSIVE GAZE, there appears to be a small negative correlation between model size and the magnitude of predicted GP effects, where larger garden path effects are predicted by the models with fewer parameters, but does not appear to hold in REGRESSIONS OUT or REGRESSIVE GO-PAST TIME. Effects in all measures are summarized in Figure 10 (bottom).

LSTM (vS) family. In the LSTM (vS) models, it appears to be the case that for a fixed amount of training data, larger model sizes are weakly associated with larger predicted garden path effects in REGRESSIONS OUT, FORWARD READING TIME, REGRESSIVE GAZE and perhaps to a lesser extent in REGRESSIVE GO-PAST. The same general pattern also appears to hold for the amount of training data. Models of a given size that were trained on a larger corpus appear to predict larger garden path effects, though this relationship was less clear on visual inspection in REGRESSIVE GO-PAST compared to the other measures. Effects in REGRESSIONS OUT are summarized in Figure 11, FORWARD READING TIME in Figure 12, REGRESSIVE GAZE in Figure 13 and REGRESSIVE GO-PAST in Figure 14.

Pythia family. Unlike the `LSTM (vS)` models, `Pythia` LLMs with more parameters generally predicted smaller garden path effects for a given amount of training data, with the four smallest LLMs generally predicting the largest garden path effects. Note however that even the smallest Pythia models are larger in their total parameter count than the largest `LSTM (vS)` models. Interestingly, it was not the case that larger LLMs predicted larger garden path effects early in training and smaller LLMs predicted larger garden path effects later in training. Instead, the smaller LLMs generally predicted larger garden path effects both early and late in training. This trend appears to hold across all measures. As for the effect of training data,

it appears that garden path effect magnitudes peak around 3,000–16,000 steps, and are consistently smaller than this peak in the models trained on the most data (143,000 steps). The effect of training data amount appears to be weaker in REGRESSIONS OUT than the other measures. Effects in REGRESSIONS OUT are summarized in Figure 15, FORWARD READING TIME in Figure 16, REGRESSIVE GAZE in Figure 17 and REGRESSIVE GO-PAST in Figure 18.

Summary. In general, larger garden path effects were predicted by LLMs with fewer parameters, but only beyond a certain point, and in models trained on a relatively modest amount of trianing data. In the relatively small `LSTM (vS)` models, larger model sizes and more training data predicted larger garden path effects across all measures, while in the `Mamba`, `GPT-2` and `Pythia` models, we generally saw the opposite pattern. Overall, our results are consistent with several works demonstrating that larger LLMs trained on more data achieve lower perplexity (i.e. they are better at their training objective of predicting the next word), but provide a poorer fit to human reading times in naturalistic reading corpora (Oh and Schuler, 2023a,b; Oh et al., 2024). These results present a serious challenge to the view that the gap between LLM-surprisal-based predictions and human reading behavior can simply be bridged through improvements to LLMs on their training objective alone (Oh and Linzen, 2025; Wilcox et al., 2025).

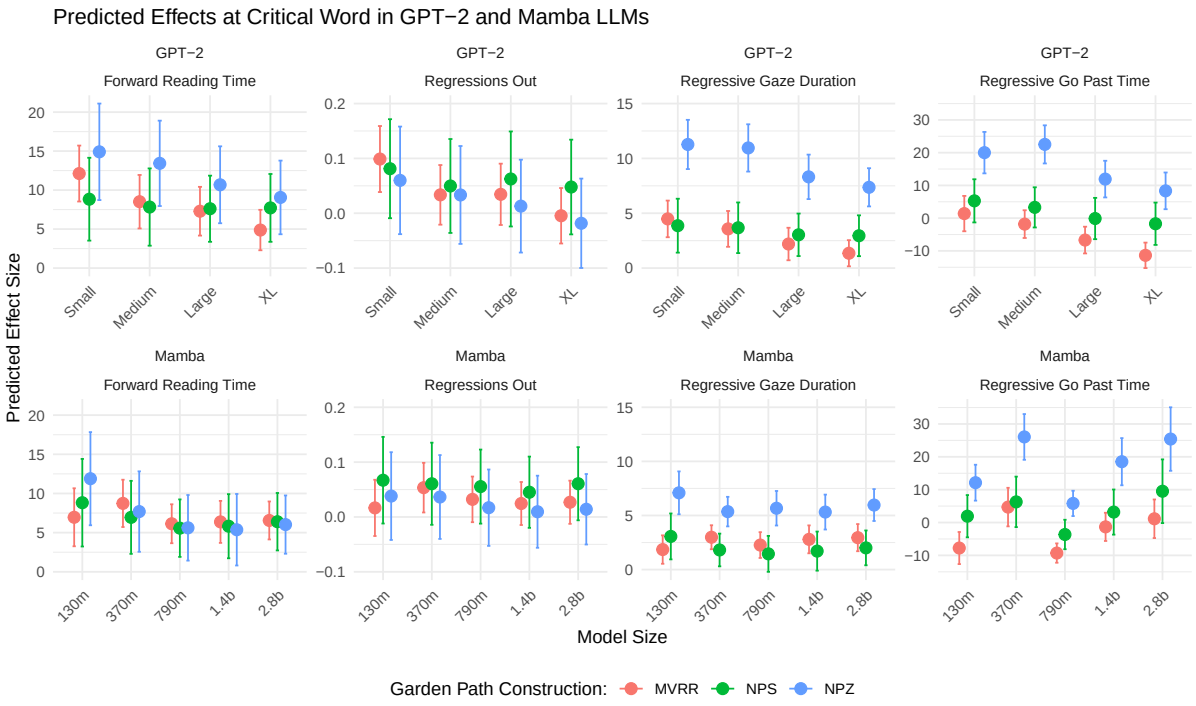


Figure 10: Predicted garden path effects in all four measures (columns) from `GPT-2` (top) and `Mamba` (bottom) LLM families. Effects in REGRESSIONS OUT are measured as the difference in log odds of a regression between the ambiguous and unambiguous conditions. All other effects are measured as the difference in milliseconds of predicted reading time between conditions. Within each plot, models are arranged from smallest to largest in parameter count.

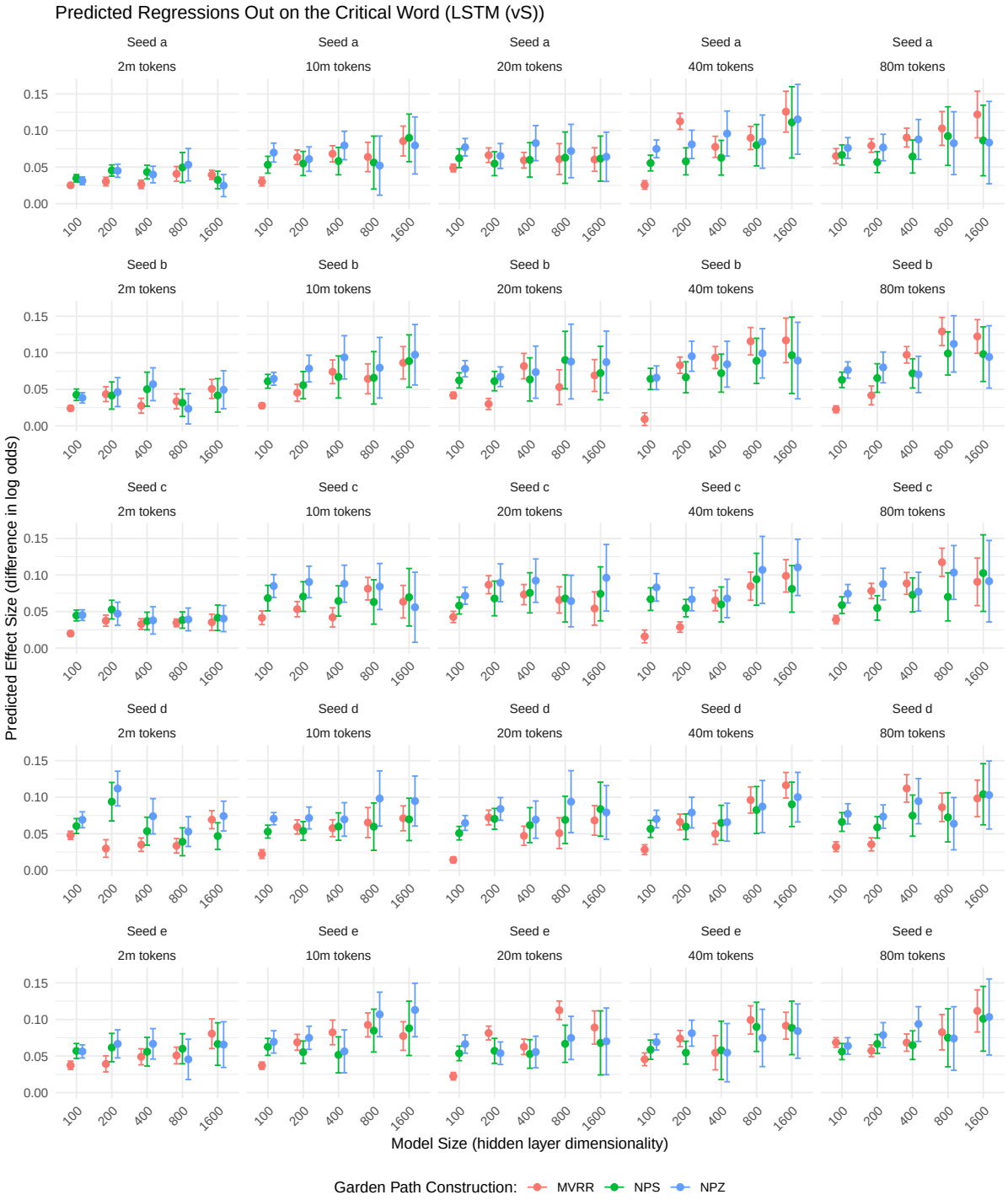


Figure 11: LSTM (vS) Predicted garden path effects in REGRESSIONS OUT, measured as the difference in log odds of a regression between the ambiguous and unambiguous conditions. Models are grouped by random seed (the initial parameter weights and training data partition) and amount of training data. Within each plot, the x-axis denotes the model size (number of dimensions in each of the LLM's two hidden layers), and the y-axis denotes the magnitude of the predicted garden path effect, with bootstrapped 95% confidence intervals.

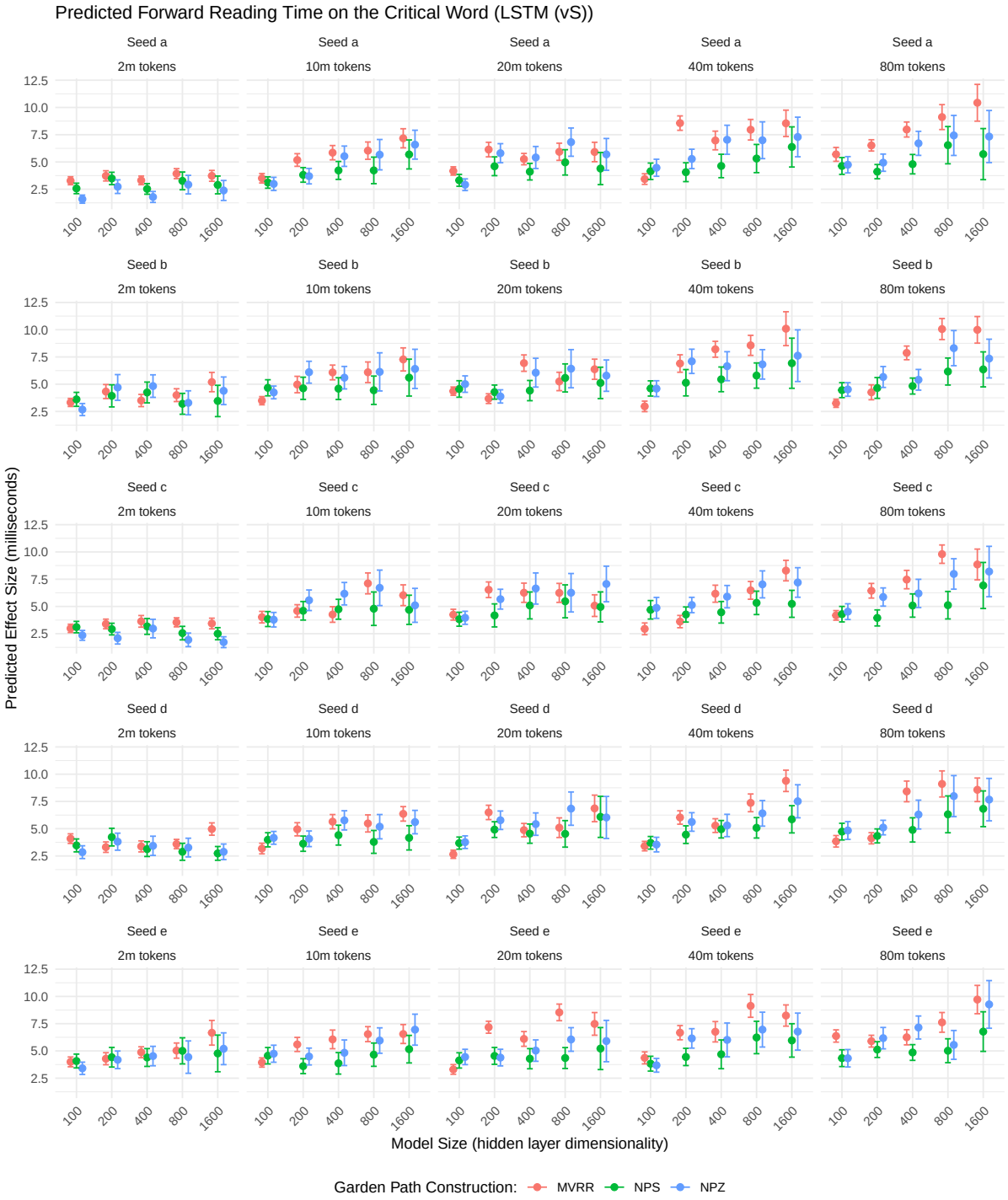


Figure 12: LSTM (vS) Predicted garden path effects in REGRESSIONS OUT, measured as the difference in log odds of a regression between the ambiguous and unambiguous conditions. Models are grouped by random seed (the initial parameter weights and training data partition) and amount of training data. Within each plot, the x-axis denotes the model size (number of dimensions in each of the LLM's two hidden layers), and the y-axis denotes the magnitude of the predicted garden path effect, with bootstrapped 95% confidence intervals.

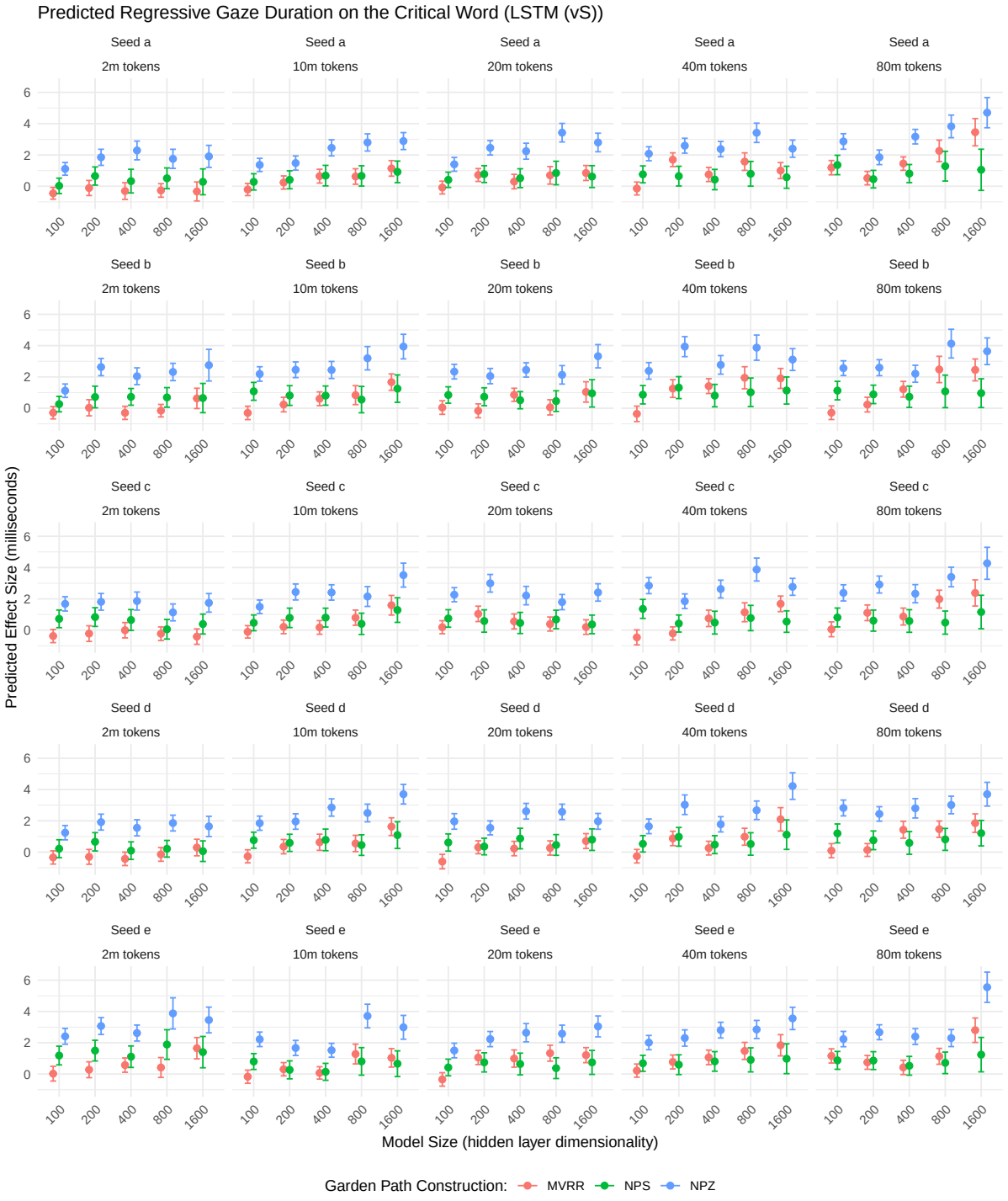


Figure 13: LSTM (vS) Predicted garden path effects in REGRESSIONS OUT, measured as the difference in log odds of a regression between the ambiguous and unambiguous conditions. Models are grouped by random seed (the initial parameter weights and training data partition) and amount of training data. Within each plot, the x-axis denotes the model size (number of dimensions in each of the LLM's two hidden layers), and the y-axis denotes the magnitude of the predicted garden path effect, with bootstrapped 95% confidence intervals.

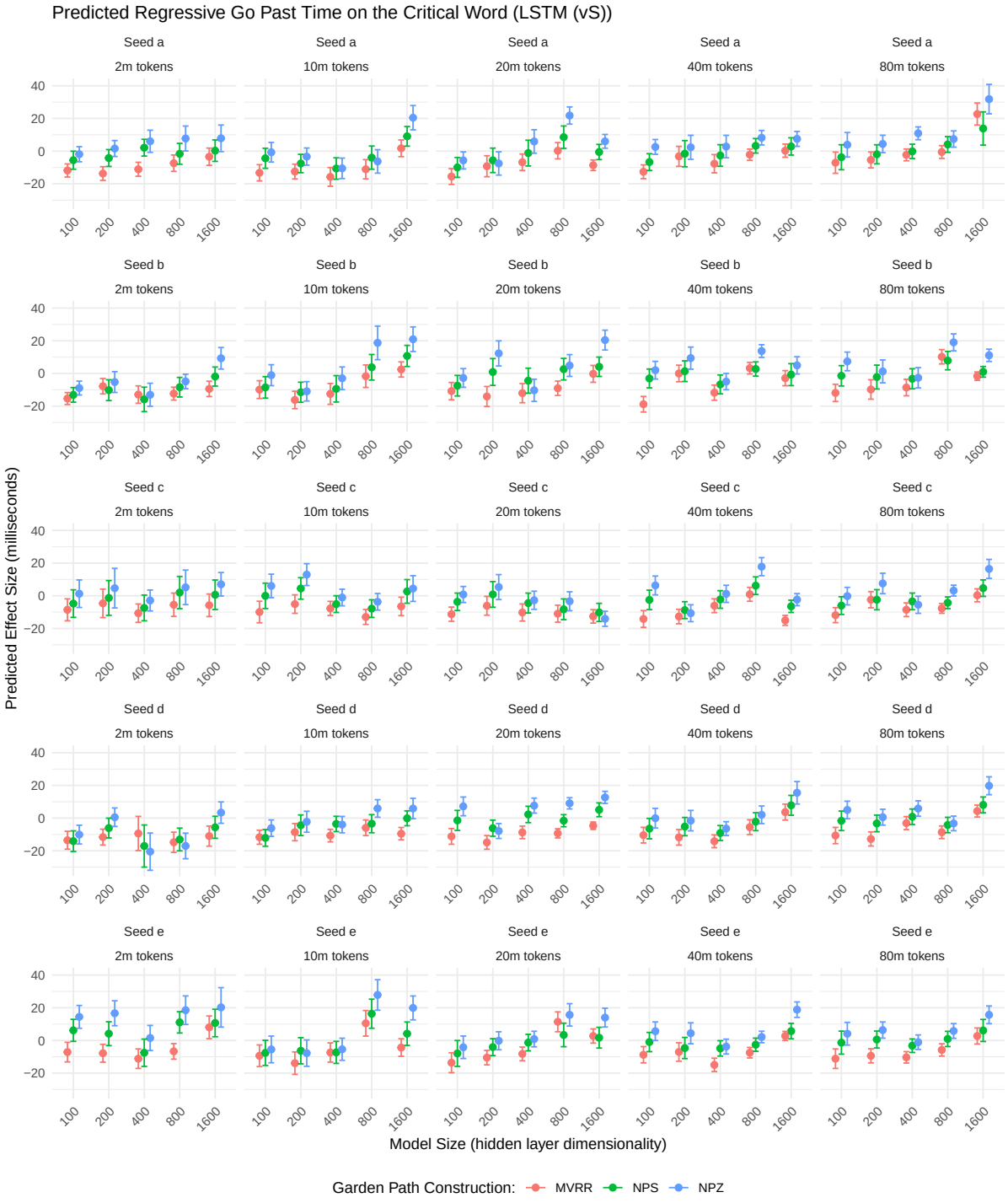


Figure 14: LSTM (vS) Predicted garden path effects in REGRESSIONS OUT, measured as the difference in log odds of a regression between the ambiguous and unambiguous conditions. Models are grouped by random seed (the initial parameter weights and training data partition) and amount of training data. Within each plot, the x-axis denotes the model size (number of dimensions in each of the LLM's two hidden layers), and the y-axis denotes the magnitude of the predicted garden path effect, with bootstrapped 95% confidence intervals.

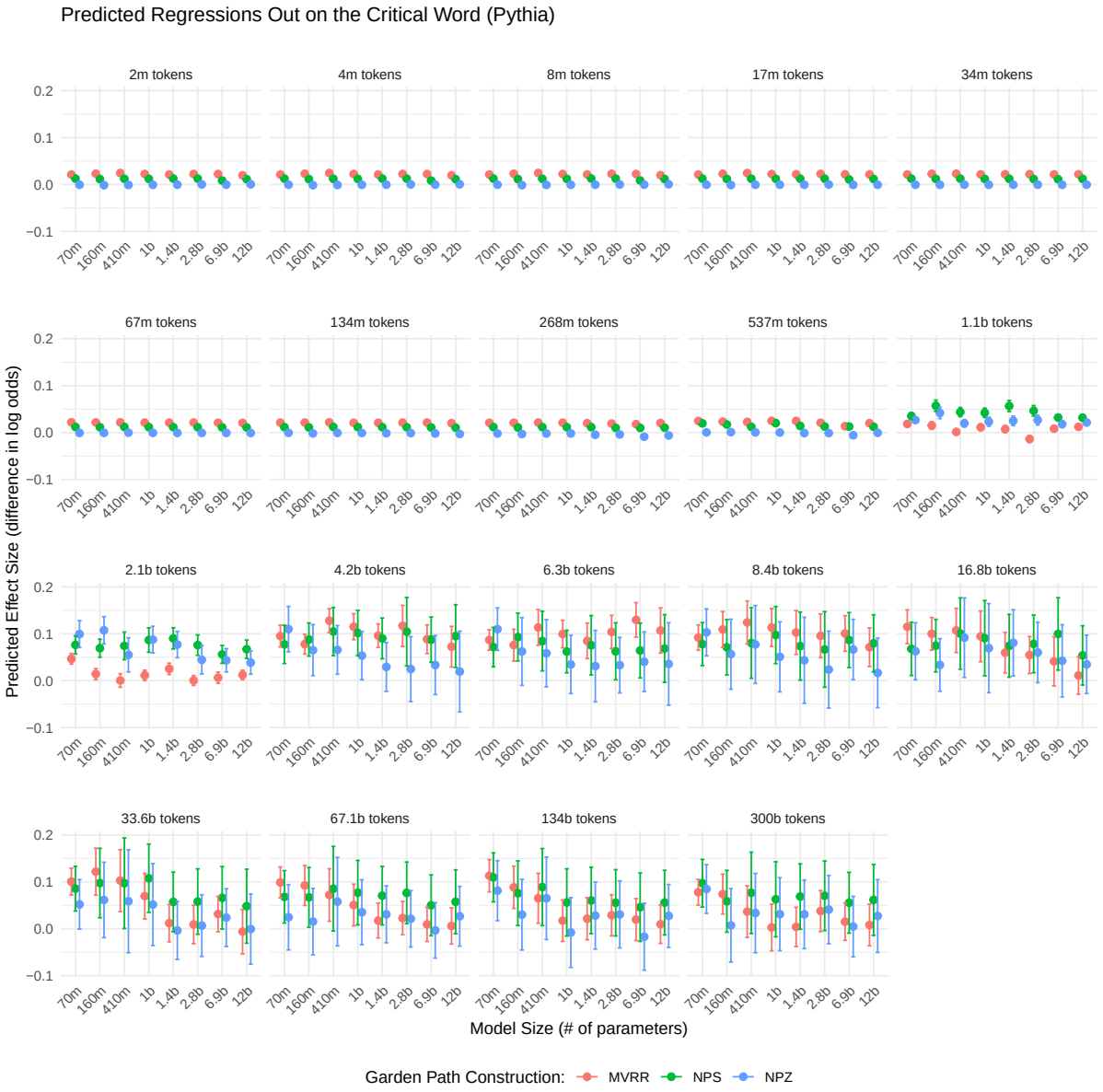


Figure 15: Pythia Predicted garden path effects in REGRESSIONS OUT, measured as the difference in log odds of a regression between the ambiguous and unambiguous conditions. Plots are grouped by the LLM's training data quantity. Within each plot, the x-axis denotes LLM size (total number of parameters), the y-axis denotes the magnitude of the predicted garden path effect, with bootstrapped 95% confidence intervals.

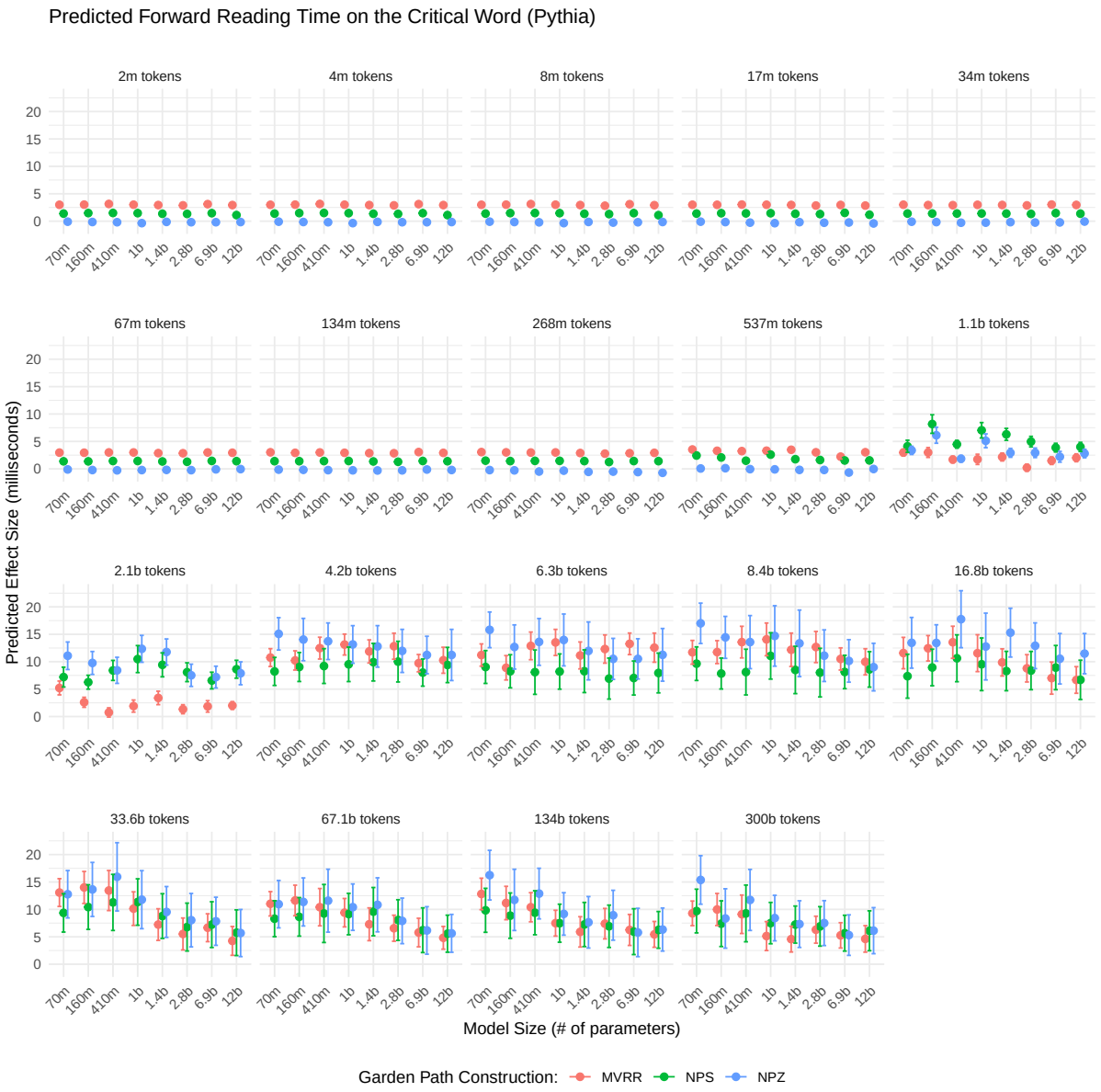


Figure 16: Pythia Predicted garden path effects in REGRESSIONS OUT, measured as the difference in log odds of a regression between the ambiguous and unambiguous conditions. Plots are grouped by the LLM's training data quantity. Within each plot, the x-axis denotes LLM size (total number of parameters), the y-axis denotes the magnitude of the predicted garden path effect, with bootstrapped 95% confidence intervals.

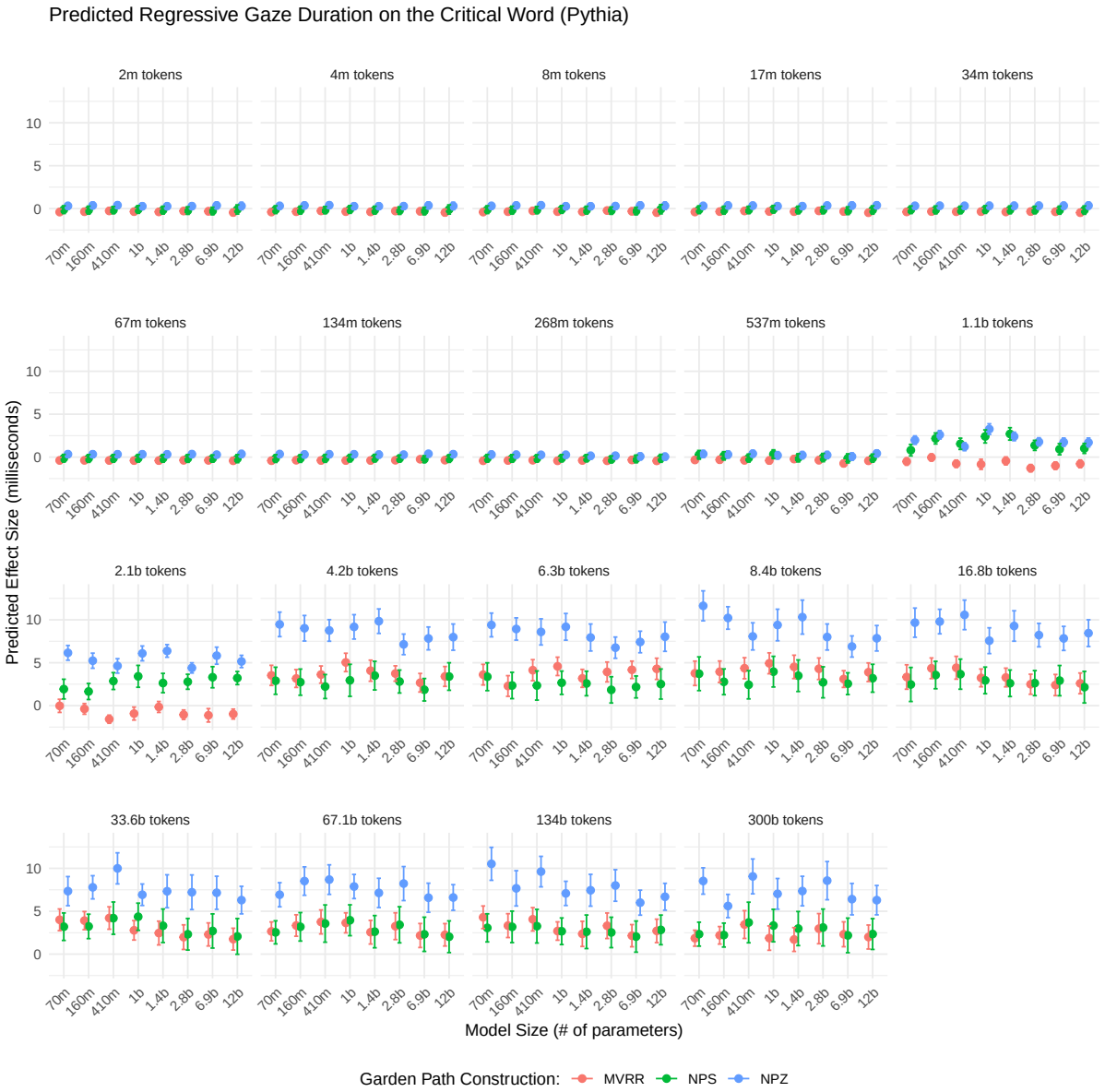


Figure 17: Pythia Predicted garden path effects in REGRESSIONS OUT, measured as the difference in log odds of a regression between the ambiguous and unambiguous conditions. Plots are grouped by the LLM's training data quantity. Within each plot, the x-axis denotes LLM size (total number of parameters), the y-axis denotes the magnitude of the predicted garden path effect, with bootstrapped 95% confidence intervals.

Figure 18: Pythia Predicted garden path effects in REGRESSIONS OUT, measured as the difference in log odds of a regression between the ambiguous and unambiguous conditions. Plots are grouped by the LLM's training data quantity. Within each plot, the x-axis denotes LLM size (total number of parameters), the y-axis denotes the magnitude of the predicted garden path effect, with bootstrapped 95% confidence intervals.

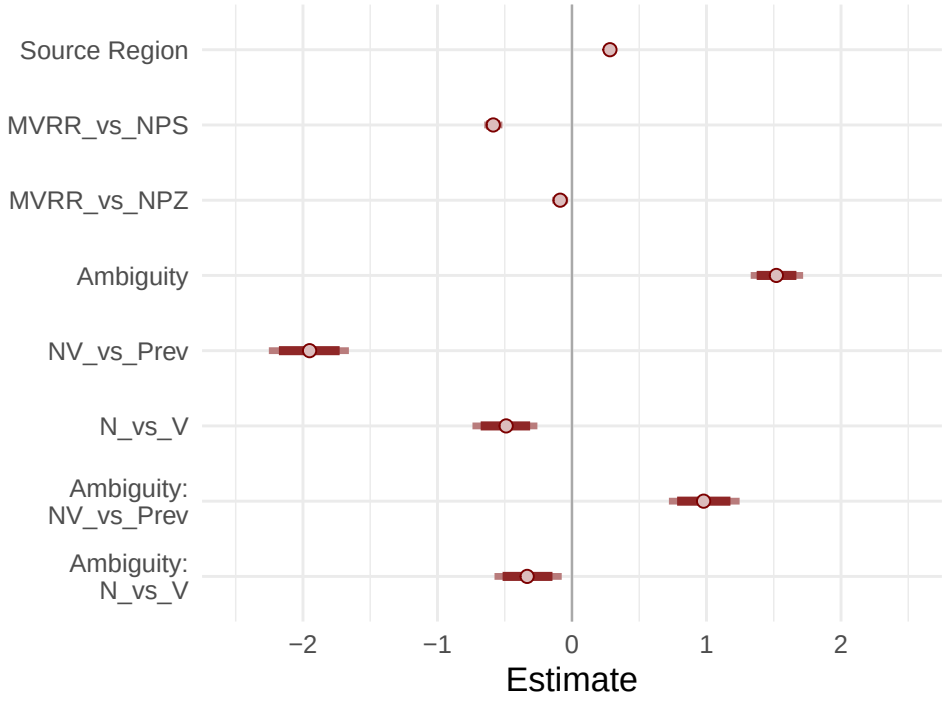


Figure 19: Posterior estimates of parameters from the Bayesian Poisson regression model of the number of fixations on earlier words. The inner bars and outer bars denote 95% and 99% credible intervals respectively.

H Parameters and predictions from Poisson regression model of the number of fixations

The posterior estimates of parameters from the Bayesian Poisson regression model of the number of fixations on earlier words can be found in Figure 19. The resulting posterior means of model predictions (i.e. log λ) for each target word and ambiguity condition can be found in Figure 20.

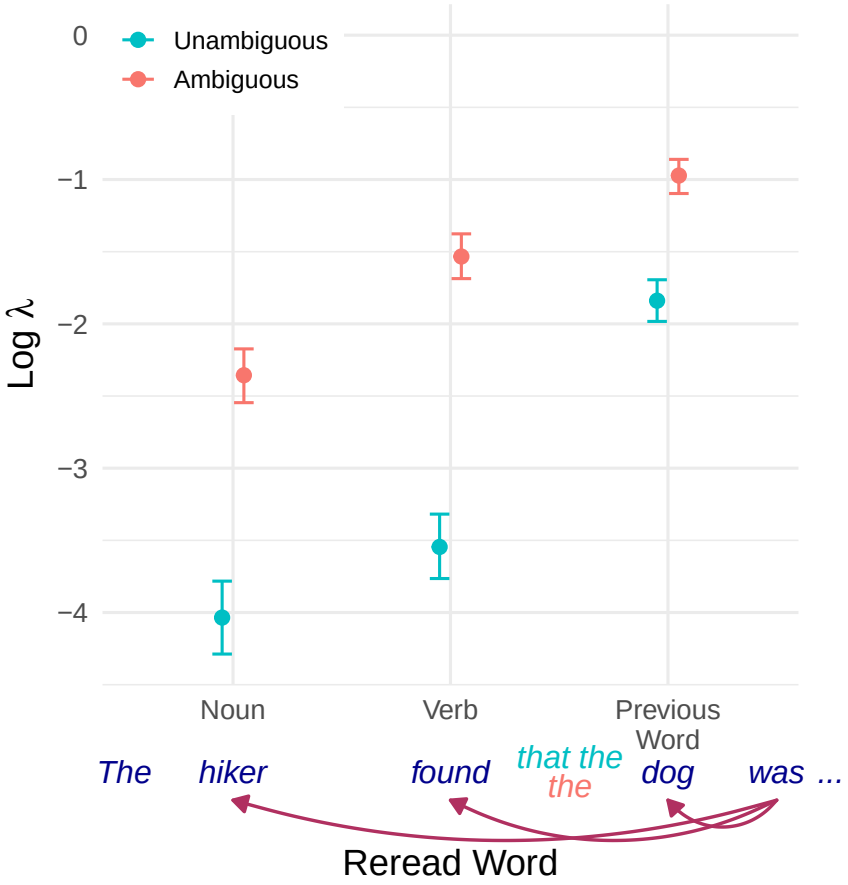


Figure 20: Posterior means of predictions from the Bayesian Poisson regression model of the number of fixations on earlier words (log λ) by target region and ambiguity, averaged across constructions and source regions. There are significant interaction effects between ambiguity and target region, such that the effect of ambiguity is larger on the verb (*found*) than on the noun (*hiker*) or on the previous word (*dog* or *was*). The error bars reflect highest-density 95% credible intervals of the posterior distribution over model parameters.

I Full list of experimental sentences

This section provides a complete list of the experimental items in each subset.

I.1 Garden path subset

For the first item, we fully spell out all six conditions. From the second item onward, we use a more concise notation, where the unambiguous conditions are differentiated by the material in parentheses.

1. (a) **Main Verb/Reduced Relative (MV/RR)**
    i. **Ambiguous**: The suspect sent the file deserved further investigation after the murder trial.
    ii. **Unambiguous**: The suspect who was sent the file deserved further investigation after the murder trial.

   (b) **Transitive/Intransitive (NP/Z)**
    i. **Ambiguous**: Because the suspect changed the file deserved further investigation after the murder trial.
    ii. **Unambiguous**: Because the suspect changed, the file deserved further investigation after the murder trial.

   (c) **Direct Object/Sentential Complement**
    i. **Ambiguous**: The suspect showed the file deserved further investigation after the murder trial.
    ii. **Unambiguous**: The suspect showed that the file deserved further investigation after the murder trial.

2. (a) The corrupt politician (who was) handed the bill received unwelcome attention from southern voters.

   (b) After the corrupt politician signed(,) the bill received unwelcome attention from southern voters.

   (c) The corrupt politician mentioned (that) the bill received unwelcome attention from southern voters.

3. (a) The woman (who was) brought the mail disappeared mysteriously after reading the bad news in it.

   (b) After the woman moved(,) the mail disappeared mysteriously from the delivery system.

   (c) The woman maintained (that) the mail disappeared mysteriously from her front porch.

4. (a) The boy (who was) fed the chicken stayed surprisingly happy despite having a mild allergic reaction.

   (b) Although the boy attacked(,) the chicken stayed surprisingly happy as if nothing happened.

   (c) The boy found (that) the chicken stayed surprisingly happy in the new barn.

5. (a) The new doctor (who was) offered the operation appeared increasingly likely to succeed in her career.

   (b) After the new doctor left(,) the operation appeared increasingly likely to succeed.

   (c) The new doctor demonstrated (that) the operation appeared increasingly likely to succeed.

6. (a) The professor (who was) awarded the grant gained more attention from marine biologists.

   (b) After the professor read(,) the grant gained more attention due to her excellent description.

   (c) The professor noticed (that) the grant gained more attention from marine biologists.

7. (a) The technician (who was) refused the service stopped working almost immediately after the argument.

   (b) After the technician called(,) the service stopped working almost immediately to his surprise.

   (c) The technician reported (that) the service stopped working almost immediately after the storm started.

8. (a) The mechanic (who was) brought the truck needed several more hours to fully repair it.

(b) Because the mechanic stopped(,) the truck needed several more hours before it could be fully repaired.

(c) The mechanic observed (that) the truck needed several more hours to be repaired.

9. (a) The guitarist (who was) assigned the song failed dramatically because he never practiced enough.

(b) After the guitarist began(,) the song failed dramatically because he skipped the sound check.

(c) The guitarist knew (that) the song failed dramatically because of the tensions within the band.

10. (a) The player (who was) paid the bonus remained essentially the same despite his sudden fame and wealth.

(b) Although the player lost(,) the bonus remained essentially the same as in the original contract.

(c) The player revealed (that) the bonus remained essentially the same as in the original contract.

11. (a) The recent hire (who was) offered the job prepared many students for careers in media.

(b) Once the recent hire started(,) the job prepared many students for careers in media.

(c) The recent hire claimed (that) the job prepared many students for careers in media.

12. (a) The assistant manager (who was) assigned the training seemed unnecessarily demanding to new staff.

(b) While the assistant manager worked(,) the training seemed unnecessarily demanding to him.

(c) The assistant manager discovered (that) the training seemed unnecessarily demanding for new staff.

13. (a) The mayor (who was) sent the document provided sufficient evidence that it was simply blackmail.

(b) Although the mayor changed(,) the document provided sufficient evidence for what he had promised.

(c) The mayor showed (that) the document provided sufficient evidence to prove her innocence.

14. (a) The basketball player (who was) handed the contract created another controversy in the NBA.

    (b) After the basketball player signed(,) the contract created another political controversy in the NBA.

    (c) The basketball player mentioned (that) the contract created another controversy in the NBA.

15. (a) The engineer (who was) brought the equipment required constant supervision from senior technicians.

    (b) After the engineer moved(,) the equipment required constant supervision from senior technicians.

    (c) The engineer maintained (that) the equipment required constant supervision from senior technicians.

16. (a) The little girl (who was) fed the lamb remained relatively calm despite having asked for beef.

    (b) When the little girl attacked(,) the lamb remained relatively calm despite the sudden assault.

    (c) The little girl found (that) the lamb remained relatively calm despite the absence of its mother.

17. (a) The yoga instructor (who was) offered the position demanded immense physical effort from everyone.

    (b) Before the yoga instructor left(,) the position demanded immense physical effort from everyone.

    (c) The yoga instructor demonstrated (that) the position demanded immense physical effort from everyone.

18. (a) The governor (who was) awarded the contract received sweeping support across the entire state.

(b) While the governor read(,) the contract received sweeping support from the audience at the rally.

(c) The governor noticed (that) the contract received sweeping support across the entire state.

19. (a) The patient (who was) refused the treatment continued causing uncomfortable scenes in the ER.

(b) Before the patient called(,) the treatment continued causing uncomfortable side effects like nausea.

(c) The patient reported (that) the treatment continued causing uncomfortable side effects like nausea.

20. (a) The operator (who was) brought the machine started working efficiently with the added automation.

(b) Once the operator stopped(,) the machine started working efficiently without any supervision.

(c) The operator observed (that) the machine started working efficiently all of a sudden.

21. (a) The dancer (who was) assigned the ballet achieved incredible success for a new performer.

(b) Once the dancer began(,) the ballet achieved incredible success for a show with a new performer.

(c) The dancer knew (that) the ballet achieved incredible success for a small local production.

22. (a) The contestant (who was) paid the money became unavailable and suddenly terminated his contract.

(b) After the contestant lost(,) the money became unavailable despite his previous three wins in a row.

(c) The contestant revealed (that) the money became unavailable to him when the show's budget shrank.

23. (a) The new chef (who was) offered the restaurant separated mediocre cooks from gifted ones.

(b) Once the new chef started(,) the restaurant separated mediocre cooks from gifted ones.

(c) The new chef claimed (that) the restaurant separated mediocre cooks from gifted ones.

24. (a) The apprentice baker (who was) assigned the oven produced smaller cakes because he lacked experience.

(b) When the apprentice baker worked(,) the oven produced smaller cakes because he lacked experience.

(c) The apprentice baker discovered (that) the oven produced smaller cakes because it heated too fast.

I.2 Relative clause subset

1. (a) **Object Relative Clause (ORC)**: The bus driver who the kids followed wondered about the location of a hotel.

(b) **Subject Relative Clause (SRC)**: The bus driver who followed the kids wondered about the location of a hotel.

2. (a) The chef who the cameraman distracted poured the flour onto the counter.

(b) The chef who distracted the cameraman poured the flour onto the counter.

3. (a) The children who the father woke bothered him about the trip to the beach.

(b) The children who woke the father bothered him about the trip to the beach.

4. (a) The class that the teacher disliked skimmed the reading for the week.

(b) The class that disliked the teacher skimmed the reading for the week.

5. (a) The dancer that the audience loved ignored some basic principles.

(b) The dancer that loved the audience ignored some basic principles.

6. (a) The employees that the fireman noticed hurried across the open field.

(b) The employees that noticed the fireman hurried across the open field.

7. (a) The farmer that the customers approached lifted the chickens from their coop.

(b) The farmer that approached the customers lifted the chickens from their coop.

8. (a) The farmer who the rancher hired piled the seeds in long rows.

   (b) The farmer who hired the rancher piled the seeds in long rows.

9. (a) The firemen that the residents called attacked the house with high-powered hoses.

   (b) The firemen that called the residents attacked the house with high-powered hoses.

10. (a) The girl who the parents watched changed a critical part of the story.

    (b) The girl who watched the parents changed a critical part of the story.

11. (a) The investigator who the agency phoned considered Ms. Reynolds from accounting.

    (b) The investigator who phoned the agency considered Ms. Reynolds from accounting.

12. (a) The judge who the witnesses addressed noticed the defense attorneys.

    (b) The judge who addressed the witnesses noticed the defense attorneys.

13. (a) The manager who the boss visited remembered some inconvenient facts.

    (b) The manager who visited the boss remembered some inconvenient facts.

14. (a) The mathematician who the chairman visited created a solution to the well-known problem.

    (b) The mathematician who visited the chairman created a solution to the well-known problem.

15. (a) The monkeys that the zookeepers watched charged the bars of their cage.

    (b) The monkeys that watched the zookeepers charged the bars of their cage.

16. (a) The movie star who the organizers visited proposed an annual prize.

    (b) The movie star who visited the organizers proposed an annual prize.

17. (a) The neighbor who the couple observed purchased the old Victorian house.

    (b) The neighbor who observed the couple purchased the old Victorian house.

18. (a) The pilot who the ground crew delayed remained on the runway for a long time.

    (b) The pilot who delayed the ground crew remained on the runway for a long time.

19. (a) The soldiers that the natives helped climbed the big rock that blocked the path.

    (b) The soldiers that helped the natives climbed the big rock that blocked the path.

20. (a) The speaker who the economists entertained predicted a good year for the industry.

    (b) The speaker who entertained the economists predicted a good year for the industry.

21. (a) The table top that the box rested on screwed directly to the legs.

    (b) The table top that rested on the box screwed directly to the legs.

22. (a) The trainer who the jockey called rubbed the horse's skin.

    (b) The trainer who called the jockey rubbed the horse's skin.

23. (a) The veteran who the coach admired defeated his greatest rival.

    (b) The veteran who admired the coach defeated his greatest rival.

24. (a) The visitor who the student introduced walked across the quad.

    (b) The visitor who introduced the student walked across the quad.

I.3 Attachment ambiguity subset

1. (a) **Low Attachment**: In the lobby, Clyde bumped into the chauffeurs of the CEO who is reckless and very unpopular with the company.

    (b) **High Attachment**: In the lobby, Clyde bumped into the chauffeur of the CEOs who is reckless and very unpopular with the company.

    (c) **Ambiguous**: In the lobby, Clyde bumped into the chauffeur of the CEO who is reckless and very unpopular with the company.

2. (a) Edwin has been reading about the sisters of the actor who was visiting the resort in Death Valley.

    (b) Edwin has been reading about the sister of the actors who was visiting the resort in Death Valley.

    (c) Edwin has been reading about the sister of the actor who was visiting the resort in Death Valley.

3. (a) From the gallery, Franny observed the nurses of the surgeon who was in charge of the operation currently underway.

(b) From the gallery, Franny observed the nurse of the surgeons who was in charge of the operation currently underway.

(c) From the gallery, Franny observed the nurse of the surgeon who was in charge of the operation currently underway.

4. (a) Gerald introduced himself to the nieces of the billionaire who sails vintage yachts around the Vineyard.

(b) Gerald introduced himself to the niece of the billionaires who sails vintage yachts around the Vineyard.

(c) Gerald introduced himself to the niece of the billionaire who sails vintage yachts around the Vineyard.

5. (a) At the potluck, Marcus chatted with the aunts of the nun who bakes sugar cookies with cute designs.

(b) At the potluck, Marcus chatted with the aunt of the nuns who bakes sugar cookies with cute designs.

(c) At the potluck, Marcus chatted with the aunt of the nun who bakes sugar cookies with cute designs.

6. (a) During the budget negotiation, Janet charmed the assistants of the executive who decides almost everything in secret.

(b) During the budget negotiation, Janet charmed the assistant of the executives who decides almost everything in secret.

(c) During the budget negotiation, Janet charmed the assistant of the executive who decides almost everything in secret.

7. (a) On the fishing trip, we laughed at the uncles of the sailor who was confused about the motor on the boat.

(b) On the fishing trip, we laughed at the uncle of the sailors who was confused about the motor on the boat.

(c) On the fishing trip, we laughed at the uncle of the sailor who was confused about the motor on the boat.

8. (a) At trial, we scrutinized the prisoners of the FBI agent who was lying about the incident at the casino.

   (b) At trial, we scrutinized the prisoner of the FBI agents who was lying about the incident at the casino.

   (c) At trial, we scrutinized the prisoner of the FBI agent who was lying about the incident at the casino.

9. (a) During the demonstration, someone photographed the soldiers of the lieutenant who was camouflaged and hiding in the trees.

   (b) During the demonstration, someone photographed the soldier of the lieutenants who was camouflaged and hiding in the trees.

   (c) During the demonstration, someone photographed the soldier of the lieutenant who was camouflaged and hiding in the trees.

10. (a) Karl recognized the hostages of the pirate who was on TV this morning on the local news.

    (b) Karl recognized the hostage of the pirates who was on TV this morning on the local news.

    (c) Karl recognized the hostage of the pirate who was on TV this morning on the local news.

11. (a) During the play, we all heckled the murderers of the prince who was disguised as a peasant from nearby Trosselheim.

    (b) During the play, we all heckled the murderer of the princes who was disguised as a peasant from nearby Trosselheim.

    (c) During the play, we all heckled the murderer of the prince who was disguised as a peasant from nearby Trosselheim.

12. (a) At the charity show, Noreen nodded to the sidekicks of the actor who was juggling sharp knives and glass bottles.

    (b) At the charity show, Noreen nodded to the sidekick of the actors who was juggling sharp knives and glass bottles.

(c) At the charity show, Noreen nodded to the sidekick of the actor who was juggling sharp knives and glass bottles.

13. (a) No one quite knew how to respond to the buddy of the janitors who burp without excusing themselves.

    (b) No one quite knew how to respond to the buddies of the janitor who burp without excusing themselves.

    (c) No one quite knew how to respond to the buddies of the janitors who burp without excusing themselves.

14. (a) The cunning Wally outmaneuvered the henchman of the villains who often fail to carry out the plot.

    (b) The cunning Wally outmaneuvered the henchmen of the villain who often fail to carry out the plot.

    (c) The cunning Wally outmaneuvered the henchmen of the villains who often fail to carry out the plot.

15. (a) Down at the pub, Ollie gossiped about the daughter of the nurses who were at church last Sunday in grimy shorts.

    (b) Down at the pub, Ollie gossiped about the daughters of the nurse who were at church last Sunday in grimy shorts.

    (c) Down at the pub, Ollie gossiped about the daughters of the nurses who were at church last Sunday in grimy shorts.

16. (a) From the lounge everyone could see the pilot of the millionaires who were distrusted by everyone at the company.

    (b) From the lounge everyone could see the pilots of the millionaire who were distrusted by everyone at the company.

    (c) From the lounge everyone could see the pilots of the millionaires who were distrusted by everyone at the company.

17. (a) On the news they showed the accomplice of the thieves who were indicted for stealing the Mona Lisa.

(b) On the news they showed the accomplices of the thief who were indicted for stealing the Mona Lisa.

(c) On the news they showed the accomplices of the thieves who were indicted for stealing the Mona Lisa.

18. (a) Everyone at the party groaned at the bodyguard of the divas who smoke clove cigarettes constantly.

(b) Everyone at the party groaned at the bodyguards of the diva who smoke clove cigarettes constantly.

(c) Everyone at the party groaned at the bodyguards of the divas who smoke clove cigarettes constantly.

19. (a) At the summit, Ursula warmly greeted the advisor of the tycoons who snowboard in Aspen in January.

(b) At the summit, Ursula warmly greeted the advisors of the tycoon who snowboard in Aspen in January.

(c) At the summit, Ursula warmly greeted the advisors of the tycoons who snowboard in Aspen in January.

20. (a) Rosalina testified against the detective of the senators who were caught spying on his colleagues.

(b) Rosalina testified against the detectives of the senator who were caught spying on his colleagues.

(c) Rosalina testified against the detectives of the senators who were caught spying on his colleagues.

21. (a) Before the exhibition, Silas telephoned the friend of the bodybuilders who write fan fiction about Batman.

(b) Before the exhibition, Silas telephoned the friends of the bodybuilder who write fan fiction about Batman.

(c) Before the exhibition, Silas telephoned the friends of the bodybuilders who write fan fiction about Batman.

22. (a) At her orientation, Tamara recently met the nephew of the professors who paint beautiful portraits of local celebrities.

    (b) At her orientation, Tamara recently met the nephews of the professor who paint beautiful portraits of local celebrities.

    (c) At her orientation, Tamara recently met the nephews of the professors who paint beautiful portraits of local celebrities.

23. (a) Everyone at the coffee shop sympathized with the courier of the florists who were complaining about the weather.

    (b) Everyone at the coffee shop sympathized with the couriers of the florist who were complaining about the weather.

    (c) Everyone at the coffee shop sympathized with the couriers of the florists who were complaining about the weather.

24. (a) Despite the good press, we didn't really like the commander of the soldiers who whistle very loudly and for no reason at all.

    (b) Despite the good press, we didn't really like the commanders of the soldier who whistle very loudly and for no reason at all.

    (c) Despite the good press, we didn't really like the commanders of the soldiers who whistle very loudly and for no reason at all.

I.4 Agreement violation subset

1. (a) **Violation**: If the supervisor changes, the schedules deserves further inspection by the rest of the staff.

   (b) **No Violation**: If the supervisor changes, the schedule deserves further inspection by the rest of the staff.

2. (a) When the magician moves, the cards disappears mysteriously from his assistant's hand.

   (b) When the magician moves, the card disappears mysteriously from his assistant's hand.

3. (a) Whenever the lawyer leaves, his clients appears increasingly uncomfortable in the courtroom.

(b) Whenever the lawyer leaves, his client appears increasingly uncomfortable in the courtroom.

4. (a) After the esteemed reviewer reads, the books gains more attention due to his glowing praise.

   (b) After the esteemed reviewer reads, the book gains more attention due to his glowing praise.

5. (a) Whenever the nurse calls, the doctors stops working immediately to check on the patient.

   (b) Whenever the nurse calls, the doctor stops working immediately to check on the patient.

6. (a) When the lecturer stops, her audiences needs several minutes to reflect on the content.

   (b) When the lecturer stops, her audience needs several minutes to reflect on the content.

7. (a) When the actress begins, the scenes fails dramatically despite the months she spent rehearsing.

   (b) When the actress begins, the scene fails dramatically despite the months she spent rehearsing.

8. (a) After the worst team loses, the tournaments remains essentially the same for the rest of the year.

   (b) After the worst team loses, the tournament remains essentially the same for the rest of the year.

9. (a) When the supervisor works, the shifts seems unnecessarily stressful on a Friday night.

   (b) When the supervisor works, the shift seems unnecessarily stressful on a Friday night.

10. (a) After the diplomat signs, the agreements creates another border conflict as a side effect.

(b) After the diplomat signs, the agreement creates another border conflict as a side effect.

11. (a) Whenever the reporter moves, the cameras requires constant adjustment from the director.

    (b) Whenever the reporter moves, the camera requires constant adjustment from the director.

12. (a) Unless the dog attacks, the cats remains relatively tranquil throughout the day.

    (b) Unless the dog attacks, the cat remains relatively tranquil throughout the day.

13. (a) Until the lead architect leaves, the projects demands immense patience from the engineers.

    (b) Until the lead architect leaves, the project demands immense patience from the engineers.

14. (a) Even if the mother calls, her boys continues causing problems with the other kids on the playground.

    (b) Even if the mother calls, her boy continues causing problems with the other kids on the playground.

15. (a) After the tutor stops, the students starts working independently on the questions.

    (b) After the tutor stops, the student starts working independently on the questions.

16. (a) Once the head surgeon begins, the operations achieves incredible results given the risks involved.

    (b) Once the head surgeon begins, the operation achieves incredible results given the risks involved.

17. (a) After the producer starts, the auditions separates mediocre actors from talented ones.

    (b) After the producer starts, the audition separates mediocre actors from talented ones.

18. (a) However hard the scientist works, his experiments produces smaller amounts of alcohol than expected.

    (b) However hard the scientist works, his experiment produces smaller amounts of alcohol than expected.

I.5 Filler items

1. There are now rumblings that Apple might soon invade the smart watch space, though the company is maintaining its customary silence.
2. A bill was drafted and introduced into Parliament several times but met with great opposition, mostly from farmers.
3. The human body can tolerate only a small range of temperature, especially when the person is engaged in vigorous activity.
4. Seeing Peter slowly advancing upon him through the air with dagger poised, he sprang upon the bulwarks to cast himself into the sea.
5. Some months later, Michael Larson saw another opportunity to stack the odds in his favor with a dash of ingenuity.
6. Bob Murphy, the Senior PGA Tour money leader with seven hundred thousand, says heat shouldn't be a factor.
7. Greg Anderson, considered a key witness by the prosecution, vowed he wouldn't testify when served a subpoena last week.
8. Owls are more flexible than humans because a bird's head is only connected by one socket pivot.
9. Even in the same animal, not all bites are the same.
10. Buck did not like it, but he bore up well to the work, taking pride in it.
11. These days, neuroscience is beginning to catch up to musicians who practice mentally.
12. Hybrid vehicles have a halo that makes owners feel righteous and their neighbors feel guilty for not doing as much to save the planet.
13. Binge drinking may not necessarily kill or even damage brain cells, as commonly thought, a new animal study suggests.
14. When attacked, a skunk's natural inclination is to turn around, lick its tail and spray a noxious scent.

15. All that the brain has to work with are imperfect incoming electrical impulses announcing that things are happening.

16. There often seems to be more diving in soccer than in the Summer Olympics.

17. Susan B. Anthony spent nearly sixty years of her life devoted to the cause of social justice and equality for all.

18. Unfortunately, for every six water bottles we use, only one makes it to the recycling bin.

19. As in the United States, Colombian legislation requires travelers entering the country to declare cash in excess of ten thousand dollars.

20. Stress is a risk factor for both depression and anxiety, he says.

21. When it comes to having a lasting and fulfilling relationship, common wisdom says that feeling close to your romantic partner is paramount.

22. Voltaire himself probably won around half a million livres, a large fortune, which he then made even larger.

23. When preparing to check out of their hotel room, some frequent travelers pile up their used bath towels on the bathroom floor.

24. Research showing that a tiny European river bug called the water boatman may be the loudest animal on earth.

25. When the new world was first discovered it was found to be, like the old, well stocked with plants and animals.

26. Police in Georgia have shut down a lemonade stand run by three girls trying to save up for a trip to a water park.

27. An early task will be to make sure the newfound microbes were not introduced while drilling through the ice into the lake.

28. Lady Gaga's YouTube account was suspended Thursday.

29. John Thornton asked little of man or nature.

30. Proper ventilation will make a backdraft less likely.

31. For centuries, time was measured by the position of the sun with the use of sundials.

32. The girl's feet were then re-wrapped even tighter than before, causing her footprint to shrink further.

33. The astronauts used a hefty robotic arm to move the bus-size canister, stuffed with nearly three tons of packing foam.

34. Very similar, but even more striking, is the evidence from athletic training.

35. I agree that California's "three strikes and you're out" law will be a financial disaster for taxpayers who care about education.

36. It was a forbidding challenge, and it says much for Winstanley's persuasive abilities, not to mention his self-confidence.

37. With schools still closed, cars still buried and streets still blocked by the widespread weekend snowstorm, officials are asking people to help out.

38. Steam sterilization is limited in the types of medical waste it can treat, but is appropriate for laboratory substances contaminated with infectious organisms.

39. From coal to cars, Chinese floods tangle supply chains worldwide.

40. This new film marks 10 years since the death of the superstar.