Journal of Memory and Language 143 (2025) 104645

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

# Dissociable frequency effects attenuate as large language model surprisal predictors improve

Byung-Doh Oh [a,*], William Schuler [b]

[a] Center for Data Science, New York University, 60 5th Avenue, New York, NY 10011, USA
[b] Department of Linguistics, The Ohio State University, 1712 Neil Avenue, Columbus, OH 43210, USA

ARTICLE INFO

Dataset link: https://osf.io/8v5qb/, https://osf.io/3527a/, https://github.com/coryshain/cdr, https://github.com/modelblocks/modelblocks-release, https://github.com/byungdoh/llm_surprisal

*Keywords:*
Sentence processing
Surprisal
Frequency
Computational modeling
Large language models

ABSTRACT

Recent psycholinguistic modeling work using surprisal from Transformer-based language models has reported separable effects of frequency and predictability on real-time processing difficulty. However, it has also been shown that as Transformer-based language models become larger and are trained on more data, they are able to predict low-frequency words more accurately, which has a deleterious effect on fit to reading times. This article examines the impact of this property of language models on the dissociability of frequency effects and predictability effects in naturalistic reading. Regression results show robust positive effects of language model size and training data amount on the ability of word frequency to explain variance in held-out reading times as the contribution due to surprisal declines, which suggests a strong compensatory relationship between frequency and language model surprisal. Additionally, an analysis of the learning trajectories of low-frequency tokens reveals that the influence of model size is strongest on the prediction of tokens that are not part of a bigram sequence observed earlier in the context that models can readily copy, which suggests that limitations in model size create pressures toward learning more general associations. Taken together, these results suggest that the observed frequency effects may be due to imperfect estimates of predictability, and may disappear entirely as better-fitting language models are discovered. This further highlights the importance of exploring additional language models as models of human sentence processing.

## Introduction

A long-standing core research agenda of psycholinguistics is to provide an account of the cognitive mechanism underlying human sentence processing. In other words, what are the processes that enable the rapid and efficient comprehension of linguistic input? How is this achieved in real time with limited cognitive resources? Decades of psycholinguistics research have highlighted the role of prediction in sentence processing (for overviews, see Kuperberg & Jaeger, 2016; Staub, 2015), which has received support from a large body of experimental and computational studies showing how predictive processing influences real-time reading behavior. However, there are still largely open questions about the characteristics and scope of prediction in human sentence processing, such as whether the increased processing difficulty observed at low-frequency words is also a by-product of predictive processing or driven by a separate lexical access mechanism.

Previous experiments that factorially manipulate frequency and predictability have reported independent effects of the two factors on reading times (Kretzschmar, Schlesewsky, & Staub, 2015; Staub & Benatar, 2013), which have provided support for the latter view. However, stimuli in psycholinguistic experiments are usually short and presented in isolation without much surrounding context, which raises concerns about their ecological validity (Hasson & Honey, 2012). This can be addressed through naturalistic experiments, where subjects are instructed to read naturalistic stimuli (e.g. short stories, newspaper articles). The resulting data is then analyzed by first operationalizing constructs of interest like frequency and predictability, and then evaluating how much separate variance they explain. Often, predictability is operationalized by surprisal (i.e. negative log probabilities) from Transformer-based large language models (LLMs), which are artificial neural networks that are trained to predict upcoming words in a corpus. While surprisal from such LLMs has been shown to be predictive of measures of processing difficulty such as self-paced reading times and eye fixation durations (Shain, Meister, Pimentel, Cotterell, & Levy, 2024), the precise hypotheses about predictive processing that they represent remain unclear due to the opaque nature of their computations.

However, it has been shown that factors like the number of model parameters and the amount of training data have a reliable effect on

* Corresponding author.
*E-mail address:* oh.b@nyu.edu (B.-D. Oh).

https://doi.org/10.1016/j.jml.2025.104645
Received 1 November 2023; Received in revised form 19 March 2025; Accepted 13 April 2025
Available online 14 May 2025

the fit of LLM surprisal to measures of processing difficulty. More specifically, LLMs that have more parameters and are trained on more data generally yield surprisal estimates that are *poorer* predictors of processing difficulty that manifests in naturalistic reading times (Oh & Schuler, 2023a, 2023b). This appears to be driven by the LLMs' capability to predict rare words accurately, which is readily learned with more parameters and large amounts of training data (Oh, Yue, & Schuler, 2024). In addition to providing an explanation for the discrepancy between LLMs and human-like predictive processing, this finding has crucial methodological implications for studying whether frequency effects are separable from predictability effects in naturalistic reading (e.g. Goodkind & Bicknell, 2021; Shain, 2019, 2024). That is, given this strong relationship between word frequency and LLM surprisal, using surprisal from larger models trained on more data is likely to result in an underestimate of predictability effects and an overestimate of frequency effects, as the excessive number of parameters and training data effectively serves to wash out difficulty associated with infrequent words that could otherwise be explained by predictability.

The present article demonstrates this point by conducting regression analyses on multiple reading time datasets that span different languages, modalities, and genres, using LLMs that vary in model size and training data amount. The results reveal a robust positive effect of both model size and training data amount of the LLM on the ability of word frequency to predict human reading times, which indicates that frequency compensates for surprisal to a greater degree as bigger LLMs trained on more data are used to calculate surprisal. Subsequent follow-up analyses examine how the increase in model size helps the prediction of low-frequency tokens. To this end, low-frequency tokens are first categorized according to factors informed by architectural properties of Transformers and contemporary language modeling practices, whose learning trajectories are subsequently analyzed. The results show that the influence of model size is strongest on tokens that are not part of a bigram sequence observed earlier in the context, which cannot be predicted by simply copying. This suggests that the limitations in model size may cause a bottleneck for learning specific associations during training, which results in less accurate predictions of the correct token, improved fit to human reading times, and a correspondingly lower contribution due to frequency.

To contextualize the article, the remainder of this section defines Transformer-based LLMs, draws a theoretical connection to human sentence processing, provides a review of empirical work on the dissociability of frequency effects and predictability effects, and introduces the framework of continuous-time deconvolutional regressive neural network for modeling reading times.

### *Transformer-based large language models*

LLMs are a class of language models that are trained on the in-context word prediction objective. These models are typically based on the Transformer neural network architecture (Vaswani et al., 2017), which does not maintain a vector representation of the context that is updated at each timestep (cf. recurrent neural networks; Elman, 1991; Hochreiter & Schmidhuber, 1997) but newly calculates a contextualized representation at each timestep through its self-attention mechanism. More specifically, autoregressive language models (e.g. Brown et al., 2020; Radford et al., 2019) are trained to predict the 'next' word given the sequence of previous words, and are therefore closer to the traditional definition of language models. These models are typically trained on large amounts of Internet text, although the exact details about their training data are usually not disclosed.

More recent approaches also employ reinforcement learning and use predicted human preferences for the generated response as a reward to fine-tune language models to general-purpose dialogue agents (e.g. reinforcement learning from human feedback; OpenAI, 2023; Ouyang et al., 2022). However, such methods entail a domain shift in their probability distribution and thereby weakens the interpretation of LLMs as models of next-word prediction trained on large-scale corpora. Therefore, throughout this article, LLMs refer specifically to autoregressive language models that have not been adapted to specific tasks.

### *Theoretical link between LLMs and human sentence processing*

There are two conceptually similar senses in which LLMs are relevant for studying human sentence processing. The first is as a computational model based on surprisal theory (Hale, 2001; Levy, 2008), which posits that the processing difficulty of a word in context is proportional to its surprisal (Shannon, 1948), or negative log probability. Surprisal theory views prediction as ongoing probabilistic inference over possible structure- or message-level analyses given the context, which are continuously updated upon observing the bottom-up input (e.g. reading the next word). Assuming that the human comprehender maintains multiple probabilistic analyses in parallel, surprisal of the observed word is equivalent to the Kullback–Leibler divergence between the probability distribution over analyses before observing the word and after observing it (Levy, 2008). Therefore, surprisal has the interpretation of the amount of 'cognitive effort' taken to readjust the analyses after observing a word. As such, early surprisal-based processing models explicitly modeled this process of probabilistic inference, mostly in the form of maintaining and updating partial syntactic structures generated by probabilistic incremental parsers. Examples of incremental parsers that have been applied as models of sentence processing include Earley parsers (Hale, 2001), top-down parsers (Roark, Bachrach, Cardenas, & Pallier, 2009), Recurrent Neural Network Grammars (Dyer, Kuncoro, Ballesteros, & Smith, 2016; Hale, Dyer, Kuncoro, & Brennan, 2018), and left-corner parsers (Jin & Schuler, 2020; van Schijndel, Exley, & Schuler, 2013). Naturally, these models were employed to study the role of syntactic expectation in human sentence processing.

Non-structural 'sequential' language models, which do not explicitly maintain multiple structure- or message-level representations of the partial sentence, have also been evaluated as expectation-based models of human sentence processing, as they directly define and estimate a conditional probability distribution necessary for surprisal calculation. As neural networks were increasingly being trained as language models, surprisal estimates from both *n*-gram language models and those based on neural network architectures[1] such as Simple Recurrent Networks (Elman, 1991), Long Short-Term Memory networks (Hochreiter & Schmidhuber, 1997), Gated Recurrent Unit networks (Cho et al., 2014), and Transformers (Vaswani et al., 2017) have been evaluated against behavioral measures of processing difficulty (Aurnhammer & Frank, 2019; Fossum & Levy, 2012; Goodkind & Bicknell, 2018; Hao, Mendelsohn, Sterneck, Martinez, & Frank, 2020; Merkx & Frank, 2021; Smith & Levy, 2013; Wilcox, Gauthier, Hu, Qian, & Levy, 2020). In addition to high-level neural network architectures, this line of research also studies the impact of factors that influence language model probabilities on the fit of surprisal, such as the amount of input context or various decoding strategies (Kuribayashi, Oseki, Brassard, & Inui, 2022; Liu, Škrjanec, & Demberg, 2024).

The other closely related sense in which LLMs are relevant for human sentence processing is quantifying predictability in psycholinguistic experiments, or how predictable the word is given its context. Probabilities derived from data collected through the cloze task (Taylor, 1953) have traditionally been used to demonstrate the effect of predictability on real-time processing in early psycholinguistic research (Ehrlich & Rayner, 1981; Kutas & Hillyard, 1980, 1984). However, it is prohibitively expensive to collect data using the cloze task on various stimuli of interest, as large samples are required for reliable estimates of predictability. Additionally, even with large samples, it

[1] Levy (2008) notes the potential similarity between surprisal from incremental parsers and prediction-based connectionist architectures.

is often difficult to make fine-grained distinctions at the low end of the predictability scale, as many words are unobserved as responses to the cloze task. In order to mitigate these issues with the cloze task, more recent research has begun to rely on predictability estimates that are approximated using corpus statistics. These estimates include conditional probabilities from bigram (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; McDonald & Shillcock, 2003), trigram (Smith & Levy, 2013), 5-gram (Shain, 2019), and Simple Recurrent Network language models (Hofmann, Remus, Biemann, Radach, & Kuchinke, 2022). From this perspective, Transformer-based LLMs are appealing as more accurate approximations of corpus statistics (cf. computational models of predictive processing that relate linguistic input to intermediate representations) in that they are trained to estimate probabilities based on very large amounts of text and can condition on a large number of preceding words, unlike earlier $n$-gram models. As such, conditional probabilities from LLMs have recently been used to examine the shape of the linking function between predictability and reading times (Hoover, Sonderegger, Piantadosi, & O'Donnell, 2023; Shain et al., 2024) and whether frequency effects dissociate from predictability effects in naturalistic reading (Shain, 2024).

### *Dissociability of frequency and predictability effects*

It is a well-established finding in experimental psycholinguistics that less frequent words take longer to read (Juhasz & Rayner, 2006; Just & Carpenter, 1980; Rayner & Duffy, 1986). However, different theoretical views about sentence processing have posited different explanations for this effect. A *procedural* view of sentence processing, which emphasizes the role of retrieval, integration, and construction of meaning (Gibson, 2000; Lewis & Vasishth, 2005), argues that this frequency effect is due to differential encoding strength in the mental lexicon, where more frequent words have stronger representations that are easier to retrieve (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Just & Carpenter, 1980; Reichle, Pollatsek, Fisher, & Rayner, 1998). This view also construes these processes as distinct from prediction, and therefore predicts dissociable frequency effects from predictability effects.

In contrast, an *inferential* view (e.g. surprisal theory; Hale, 2001; Levy, 2008) emphasizes the probabilistic inference over possible structure- or message-level analyses given the partial sentence, and posits that the contextual predictability of a word determines its processing difficulty. According to this view, frequency effects should be subsumed by predictability effects, because more frequent words are also more predictable than less frequent words in unconstrained contexts (i.e. they have higher prior probabilities).

Modeling studies that aim to answer this question using naturalistic reading times have yielded mixed results. For example, Goodkind and Bicknell (2021) analyzed fixation durations in the Dundee dataset (Kennedy, Hill, & Pynte, 2003) and found dissociable frequency effects from predictability effects that were operationalized by $n$-gram language models. This is in contrast to the earlier work by Shain (2019), who found that word frequency did not improve fit to unseen data on the Dundee, Natural Stories (Futrell et al., 2021), and UCL (Frank, Monsalve, Thompson, & Vigliocco, 2013) datasets over a baseline including 5-gram surprisal. In light of these conflicting findings, Shain (2024) revisited this question at scale with more reading time datasets and a more flexible modeling approach that relaxes assumptions that are inappropriate for modeling naturalistic reading. Using surprisal estimates from the GPT-2 language model (Radford et al., 2019) to operationalize predictability, Shain (2024) found a dissociable frequency effect on most datasets, which is consistent with the predictions of the procedural view.

While Shain (2024) acknowledges that model size has an impact on the quality of surprisal estimates as predictors of reading times (based on results in e.g. Oh, Clark, & Schuler, 2022; Shain et al., 2024), the potential influence of training data amount was not considered. More importantly, word frequency was found to modulate the influence of model size and training data amount on the ability of LLM surprisal to predict human reading times (Oh et al., 2024). It is therefore likely that different estimates of frequency effects could be derived depending on the LLM used to operationalize predictability in a modeling study using naturalistic reading times. We illustrate this point through an experiment that closely follows the procedures of Shain (2024), as well as an experiment using multilingual reading time data.

### *Continuous-time deconvolutional regressive neural network*

Shain (2024) studies the dissociation of frequency and predictability effects at scale using the modeling framework of continuous-time deconvolutional regressive neural network (CDR-NN; Shain, 2021; Shain & Schuler, 2024), which we also employ in our first experiment. While standard approaches like linear mixed-effects models (LMEM; Bates, Mächler, Bolker, & Walker, 2015) and generalized additive models (GAM; Wood, 2006) are typically used to study human reading times (e.g. Hoover et al., 2023; Oh & Schuler, 2023b; Wilcox et al., 2020), these modeling approaches assume that the current response reading time $y_i$ depends solely on the corresponding predictors $\mathbf{x}_i$ and is independent of any preceding predictors. This limits LMEMs and GAMs in capturing the lingering influence of the *current* word on *future* reading times, which is well-known as 'spillover' effects in psycholinguistics (Rayner, Carlson, & Frazier, 1983; Vasishth, 2006). While this issue is commonly addressed by including 'spillover variants' of predictors from preceding words as predictors of the current response reading time, this may lead to identifiability issues in LMEM/GAMs and additionally makes the assumption that previously processed words are dispersed evenly throughout time.

Continuous-time deconvolutional regression (CDR; Shain & Schuler, 2021) was developed to address these limitations by estimating a parametric continuous-time impulse response function (e.g. the three-parameter shifted Gamma function) in a data-driven manner. CDR-NN models are extensions of CDR models based on deep neural networks that estimate a nonlinear function that relates a set of predictors (e.g. unigram surprisal) to its effect on the parameters of the predictive distribution over the response (i.e. reading times) with some continuous time delay. In doing so, CDR-NN models additionally relax assumptions that the influence of the predictor on the response is linear and homoscedastic (constant error), which are also implausible for modeling the time course of naturalistic reading (Shain & Schuler, 2024).

## Experiment 1: Effects of LLM size and training data amount on frequency effects

The first experiment evaluates the ability of word frequency to predict naturalistic reading times in the presence of surprisal estimates from LLMs that systematically vary in model size and training data amount. To this end, a series of CDR-NN (Shain, 2021; Shain & Schuler, 2024) models was fit and evaluated.

### *Methods*

#### *Response data*

This experiment used reading times from five self-paced reading (SPR) and eye-tracking (ET) corpora, which are the Brown and Natural Stories SPR corpora (Futrell et al., 2021; Smith & Levy, 2013) and the GECO, Dundee, and Provo ET corpora (Cop, Dirix, Drieghe, & Duyck, 2017; Kennedy et al., 2003; Luke & Christianson, 2018). Data preprocessing and modeling procedures closely follow those of Shain (2024), which are described below.

For the SPR datasets, the by-word reading times (i.e. the time taken between keystrokes to advance to the next word) provided the response variables. Data points were filtered to exclude those of sentence-initial and -final words and those shorter than 100 ms or longer than 3000 ms. For the ET datasets, as eye movements recorded through eye-tracking are non-linear, the following by-word measures were analyzed.

- Scan path duration: Time taken between entering a word region (from the left or right) and entering a different word region (to the left or right). Unlike total fixation duration, this treats rereading of the same word as a different event from its first reading.
- First-pass duration: Time elapsed between entering a word region from the left in the first pass and entering a different word region (to the left or right).
- Go-past duration: Time elapsed between entering a word region from the left and entering a word region to the right (including all regressive fixations).

The response reading times were filtered to remove those for unfixated words, fixations interrupted by blinks, words following saccades longer than 20 words, and words at starts and ends of sentences and documents. For the Dundee Corpus (Kennedy et al., 2003) that further provides annotations of positions within lines and screens, response reading times of words at starts and ends of lines and screens were also removed.

Prior to regression modeling, all datasets were split into fit, exploratory, and held-out partitions roughly consisting of 50%, 25%, and 25% of data points respectively. This partitioning was conducted based on the sum of the subject index and the sentence index,[2] meaning that all data points from a subject reading a particular sentence was kept intact in one partition. The fit partition was used to fit the regression models, and the exploratory partition was used to diagnose convergence and inform early stopping. Model comparison and statistical significance tests were conducted on the held-out partition.

*Brown SPR.* The Brown SPR Corpus (Smith & Levy, 2013) contains reading times from 35 subjects that read 13 English passages from the Brown Corpus (Kučera & Francis, 1967) that consist of a total of 7188 words. The partitioning and filtering procedures resulted in 59,617, 29,693, and 29,810 data points in the fit, exploratory, and held-out partitions respectively.

*Natural Stories SPR.* The Natural Stories Corpus (Futrell et al., 2021) contains reading times from 181 subjects that read 10 naturalistic English stories consisting of 10,256 words. Data points were further filtered to remove those from subjects who answered fewer than four comprehension questions correctly. The partitioning and filtering procedures resulted in 386,249, 192,455, and 190,475 data points in the fit, exploratory, and held-out partitions respectively.

*GECO ET.* The Ghent Eye-Tracking Corpus (GECO; Cop et al., 2017) contains eye-tracking data from 14 monolingual subjects that read the English version of the novel *The Mysterious Affair at Styles* (Agatha Christie, 1920), which consists of 13 chapters and 56,441 words. The partitioning and filtering procedures resulted in 149,392, 74,846, and 74,532 first-pass and go-past durations[3] in the fit, exploratory, and held-out partitions respectively.

*Dundee ET.* The Dundee Corpus (Kennedy et al., 2003) contains eye-tracking data from 10 subjects that read 67 English newspaper editorials consisting of a total of 51,501 words. The partitioning and filtering procedures resulted in 126,970, 63,469, and 63,580 scan path durations, as well as 96,661, 48,196, and 48,276 first-pass and go-past durations in the fit, exploratory, and held-out partitions respectively.

[2] If this value modulo four is zero or one, the data point was assigned to the fit partition; if this value modulo four is two, the data point was assigned to the exploratory partition; if this value modulo four is three, the data point was assigned to the held-out partition.

[3] The scan path durations could not be calculated as the GECO dataset does not provide raw fixation durations.

**Table 1**
Hyperparameters and model sizes of Pythia LLM variants whose surprisal estimates were used in this experiment. #L, #H, and $d_{\text{model}}$ refer to number of layers, number of attention heads per layer, and embedding size, respectively.

| Model | #L | #H | $d_{\text{model}}$ | #Parameters |
|---|---|---|---|---|
| Pythia 70M | 6 | 8 | 512 | ~70M |
| Pythia 160M | 12 | 12 | 768 | ~160M |
| Pythia 410M | 24 | 16 | 1024 | ~410M |
| Pythia 1B | 16 | 8 | 2048 | ~1B |
| Pythia 1.4B | 24 | 16 | 2048 | ~1.4B |
| Pythia 2.8B | 32 | 32 | 2560 | ~2.8B |
| Pythia 6.9B | 32 | 32 | 4096 | ~6.9B |
| Pythia 12B | 36 | 40 | 5120 | ~12B |

*Provo ET.* The Provo Corpus (Luke & Christianson, 2018) contains eye-tracking data from 84 subjects that read 55 short English passages consisting of a total of 2746 words that range from news articles, science magazines, and works of fiction. The partitioning and filtering procedures resulted in 72,801, 36,633, and 36,397 scan path durations, as well as 50,534, 25,349, and 25,406 first-pass and go-past durations in the fit, exploratory, and held-out partitions respectively.

*Predictors*

*Surprisal predictors.* As a measure of word frequency, unigram surprisal was estimated using the OpenWebText Corpus (Gokaslan & Cohen, 2019), which is a replication of GPT-2's (Radford et al., 2019) WebText training corpus and contains about 6.5 billion words. Unigram probabilities were estimated using the KenLM toolkit (Heafield, Pouzyrevsky, Clark, & Koehn, 2013) and its interpolation protocols for out-of-vocabulary words, which were subsequently converted to surprisal scale.

The unigram surprisal predictor was evaluated in the presence of surprisal estimates from Pythia LLMs (Biderman et al., 2023) that differ in model size and training data amount. Pythia LLMs are decoder-only autoregressive Transformer-based models whose variants differ primarily in their model size. The relevant hyperparameters and resulting model sizes of the Pythia variants are outlined in Table 1.

Importantly for this experiment, all eight Pythia variants were trained using identical batches of training examples that were presented in the same order. These training examples are from the Pile (Gao et al., 2020), which is a collection of English language datasets that consists of around 300 billion tokens. A total of 143,000 batches of 1024 examples with 2048 tokens (i.e. 2,097,152 tokens in each batch) was used to train the eight variants, which amounts to about one epoch over the entire Pile dataset. Model parameters that were saved during early training stages (i.e. after 1, 2, 4, ..., 256, 512 batches) as well as after every 1000 batches are publicly available.[4] Surprisal estimates from the eight variants after {256, 512, 1000, 2000, 4000, 8000, 143 000} training batches were used in this experiment for computational efficiency of regression modeling. These steps were chosen based on earlier results that show a peak in fit to human reading times at around 1000 training batches, and relatively little change after batch 8000 onwards for the Natural Stories and Dundee datasets (Oh & Schuler, 2023a).

Each article or story of the reading time corpora was tokenized by Pythia's byte-pair encoding (BPE; Sennrich, Haddow, & Birch, 2016) tokenizer and provided as input to each model variant. When a word $w_t$ was tokenized into multiple tokens, negative log probabilities of subword tokens corresponding to $w_t$ were summed to calculate $\mathsf{S}(w_t) = -\log \mathsf{P}(w_t \mid w_{1..t-1})$ according to the chain rule of conditional probabilities.[5] In cases where each story or article was longer than a single context window of 2048 tokens, surprisal estimates for the remaining tokens were calculated by using the second half of the previous context window as the first half of a new context window.

[4] To our knowledge, this makes Pythia models the most comprehensive in terms of model size and training data amount, which motivates their use in this experiment.

*Baseline predictors.* In addition to the surprisal predictors, the following set of baseline predictors was also included in all regression models.

- Rate: The "deconvolutional intercept" that depends solely on stimulus timing regardless of the properties of a word (Shain & Schuler, 2021).
- Word length: Length of the word in characters.
- End of sentence: Whether the word ends a sentence, which is designed to capture lingering effects of sentence boundaries.

For the models fit to ET datasets, the following modality-specific baseline predictors were also included.

- Saccade length: Length of the incoming saccade in words.
- Regression: Whether the fixation follows a regressive saccade.

For models fit to the Dundee ET dataset that provides annotations of screen and line boundaries, the following baseline predictors were included for similar reasons as the 'end of sentence' predictor above.

- End of line: Whether the word ends a line.
- End of screen: Whether the word ends a screen.

*Regression modeling*

This experiment fit a series of CDR-NN models to evaluate the contribution of word frequency over predictability. Following Shain (2024), the CDR-NN models estimated the parameters of an exGaussian distribution, which is the convolution of a Gaussian distribution (with location $\mu$ and dispersion $\sigma$ parameters) with an exponential distribution (with a skewness $\tau$ parameter). The exGaussian distribution has been used in previous work to study the distributional influence of word frequency and predictability on reading times (Staub, 2010, 2011). To assess the contribution of frequency over predictability, two sets of CDR-NN models were fit to reading times in the fit partition of each corpus. The 'full' models are CDR-NN models that include both unigram surprisal and LLM surprisal as described in Section "Predictors", and their corresponding 'ablated' models are those that ablate the unigram surprisal predictor. All models include maximal by-subject random intercepts, slopes, and neural network parameters, as well as random intercepts for each word position in each text. Convergence is diagnosed based on a time-loss correlation criterion on the exploratory partition, where the correlation between training epoch number and likelihood every 10 epochs is calculated over a window of 250 consecutive epochs (Shain & Schuler, 2021). More specifically, models are considered to be converged whenever likelihood on the exploratory partition is statistically non-increasing at $\alpha = 0.5$ for at least 13 of the preceding 25 evaluations.

After the CDR-NN models have converged, the change in conditional log-likelihood ($\Delta$LL) on the held-out partition of each corpus was calculated between a full model and its counterpart ablated model. To minimize variation due to stochastic optimization, the change in log-likelihood was calculated based on ensembles of 10 training runs; the median held-out log-likelihood of 10 runs of an ablated model was subtracted from that of 10 runs of a full model to calculate $\Delta$LL.

[5] Concurrent work (Oh & Schuler, 2024b; Pimentel & Meister, 2024) has shown that this method of aggregating subword probabilities results in incorrect word probabilities if the tokens have leading whitespaces like those in Pythia's vocabulary. Follow-up analysis with corrected surprisal from LLM variants showed that the numerical difference in regression model likelihood is not large, and therefore we conclude that this would not drastically change the results of this experiment.

*Results*

First, we present how well surprisal from each LLM variant predicts the reading times from the 10 datasets in aggregate over just the baseline predictors, as a measure of how strong they are as predictors of comprehension difficulty. Fig. 1 generally replicates previous results using linear mixed-effects modeling that show a 'peak' at around 1000 training batches (the 70M variant) and an adverse effect of model size by the end of training (Oh & Schuler, 2023a).

On top of these surprisal predictors, the ability of word frequency to predict naturalistic reading times decreases as the LLMs used to generate surprisal are trained on fewer data (Fig. 2). The slope of the best-fitting line between log number of training batches and $\Delta$LL due to unigram surprisal is greater than zero on all 10 datasets, eight of which obtain statistical significance by a permutation test that permutes the independent variable. This overall trend is highly significant by a binomial test (8 out of 10 successes; $p < 0.001$). The lack of significance on the Brown and Natural Stories SPR datasets is likely due to the high variance in held-out log-likelihood between training runs (see e.g. Shain et al., 2024), which itself may be due to the smaller number of predictors compared to ET datasets for a highly flexible model like CDR-NN.

The same data visualized as a function of model size (Fig. 3) reveals a similar trend, where $\Delta$LL due to unigram surprisal decreases as smaller LLMs are used to calculate surprisal. Again, the slope of the best-fitting line between log number of parameters and $\Delta$LL is greater than zero on all 10 datasets, seven of which obtain statistical significance by a permutation test. This overall trend is again highly significant by a binomial test (7 out of 10 successes; $p < 0.001$). These results indicate that frequency compensates for surprisal less and less as smaller LLMs trained on fewer data are used to calculate surprisal. Together with the results in Fig. 1, this shows that generally speaking, word frequency plays a smaller role in accounting for processing difficulty in the presence of surprisal estimates that approximate human reading times more closely.

*Discussion*

Through the first experiment, we provide evidence that the ability of frequency to predict naturalistic reading times over and above predictability depends on the model size and training data amount of the LLM used to calculate surprisal. More specifically, frequency becomes less predictive of reading times when the underlying LLM is smaller, trained on fewer data, and provides a better surprisal predictor. This confirms the strong relationship between LLM surprisal and frequency that is modulated by model size and training data amount (Oh et al., 2024), and indicates that limitations in these two variables result in surprisal estimates that are more consistent with the inferential view of frequency effects (i.e. frequency effects as a by-product of predictive processing). The finding that surprisal from smaller models trained on fewer data is generally more predictive of reading times (Fig. 1, but also Oh & Schuler, 2023b; Shain et al., 2024) further suggests that Shain's (2024) estimates of frequency effects on top of surprisal from the GPT-2 model (Radford et al., 2019) may be systematically overestimated.

It should be noted that the results of this experiment are based on surprisal from Pythia models, which have a specific Transformer-based architecture and a specific training dataset. As a consequence, while these results may not generalize to other models, this serves as a proof of concept that different conclusions about the dissociability of frequency effects from predictability effects could be drawn depending on the underlying LLM, and that there are reasons to be cautious of Shain's (2024) conclusions. An active line of research in both language modeling and computational psycholinguistics involves studying how the architectural properties of a neural network influence the learned probabilities (Bietti, Cabannes, Bouchacourt, Jegou, & Bottou,

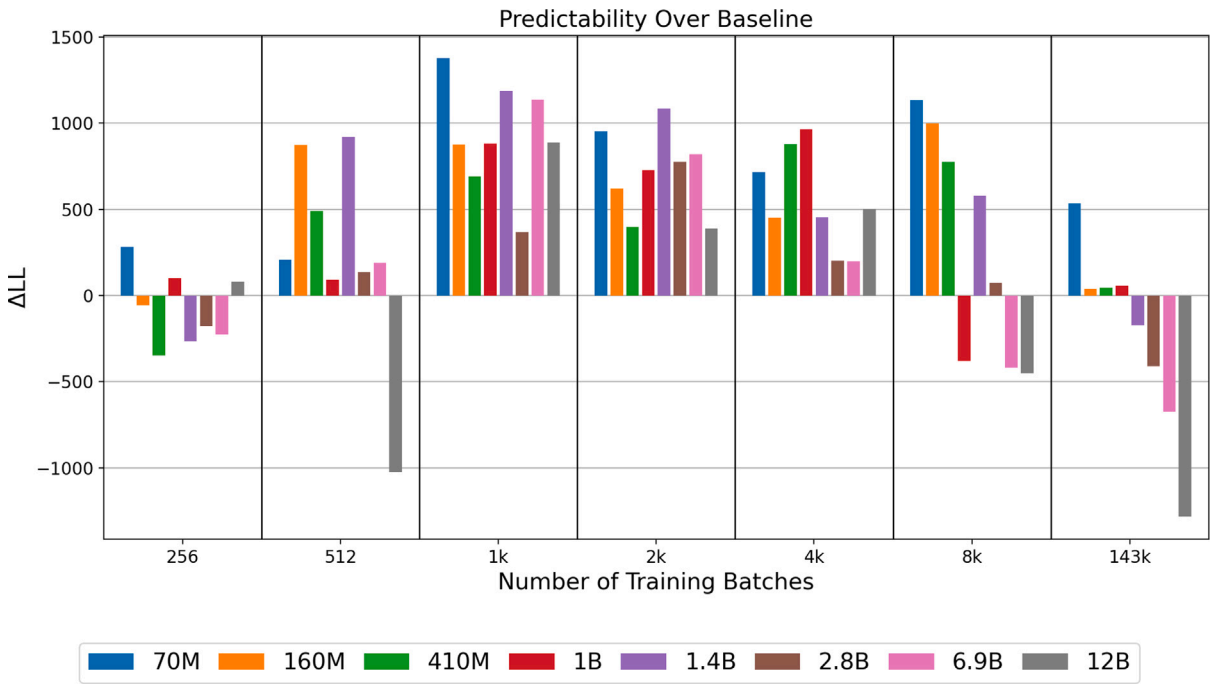


**Fig. 1.** Increase in CDR-NN model log-likelihood due to including LLM surprisal in the presence of baseline predictors as a function of number of training batches and model size, aggregated across all reading time datasets.

2023; Jelassi, Brandfonbrener, Kakade, & Malach, 2024) and their fit to measures of processing difficulty (Clark, Oh, & Schuler, 2025; de Varda & Marelli, 2024), which indicates that exploring additional language models that capture human-like predictive processing remains as an open agenda. Therefore, we refrain from adjudicating the underlying relationship between frequency and predictability effects based on these results, but rather emphasize declining effects of frequency as LLM-based predictors improve and the possibility of dissociable frequency effects disappearing entirely as better-fitting language models are discovered.

## Experiment 2: Effects of LLM size replicated on multilingual datasets

The previous experiment revealed a strong influence that the model size of LLMs used for surprisal calculation has on the ability of word frequency to predict naturalistic reading times. However, as the experiment relies solely on LLMs trained on English text and data from native speakers of English, it remains to be seen whether the results generalize to data from other languages. To this end, the second experiment evaluates the influence of LLM size on word frequency's fit to reading times in a multilingual eye-tracking dataset using linear mixed-effects regression models.

### *Methods*

#### *Response data*

This experiment used reading times from the Multilingual Eye-Movement Corpus (MECO; Siegelman et al., 2022), which contains eye-tracking data collected in 13 languages.[6] The data were collected from 29–54 subjects (depending on language) that read a total of 12 Wikipedia-style entries covering a variety of topics, which were chosen to minimize the potential influence of academic knowledge or cultural bias across collection sites. As in Experiment 1, these fixation durations were processed to calculate the scan path, first-pass, and go-past durations for each word region. Subsequently, the data were filtered to remove those for unfixated words, words following saccades longer than four words, and words at starts and ends of sentences, lines, and documents. Additionally, data from subjects that answered fewer than two out of four article-level comprehension questions correctly were excluded from analysis. The data were then split into fit, exploratory, and held-out partitions roughly consisting of 50%, 25%, and 25% of data points respectively, following the same procedures as Experiment 1. The fit partition was used to fit the regression models, and all results are reported on the exploratory partition. The final number of observations in each partition of MECO is outlined in Table 2.

#### *Predictors*

*Surprisal predictors.* As a measure of word frequency, unigram surprisal was calculated using the `wordfreq` toolkit (Robyn Speer, 2022), which is based on word counts from corpora that span multiple genres. This unigram surprisal predictor was evaluated in the presence of surprisal estimates from publicly available LLMs trained on each language covered by MECO. For comparability with Experiment 1, decoder-only autoregressive Transformer-based models that had multiple variants of different sizes were chosen for analysis.[7] Whenever available, LLMs trained only on data in the language of interest (i.e. 'monolingual' models) were chosen. For languages in which such models were not available, the 'multilingual' mGPT models (Shliazhko et al., 2024) were used to calculate surprisal estimates.[8] The relevant hyperparameters and model sizes of each LLM variant are outlined in Table 3.

[6] We exclude the Estonian data from our analysis, as there were no publicly available estimates of word frequency.

[7] Unlike the English Pythia models, none of the other languages in MECO had publicly available LLMs that vary systematically in the amount of their training data.

[8] As all of these models have leading whitespaces in their subword tokens, correction to their word probabilities (Oh & Schuler, 2024b; Pimentel & Meister, 2024) was applied prior to surprisal calculation.

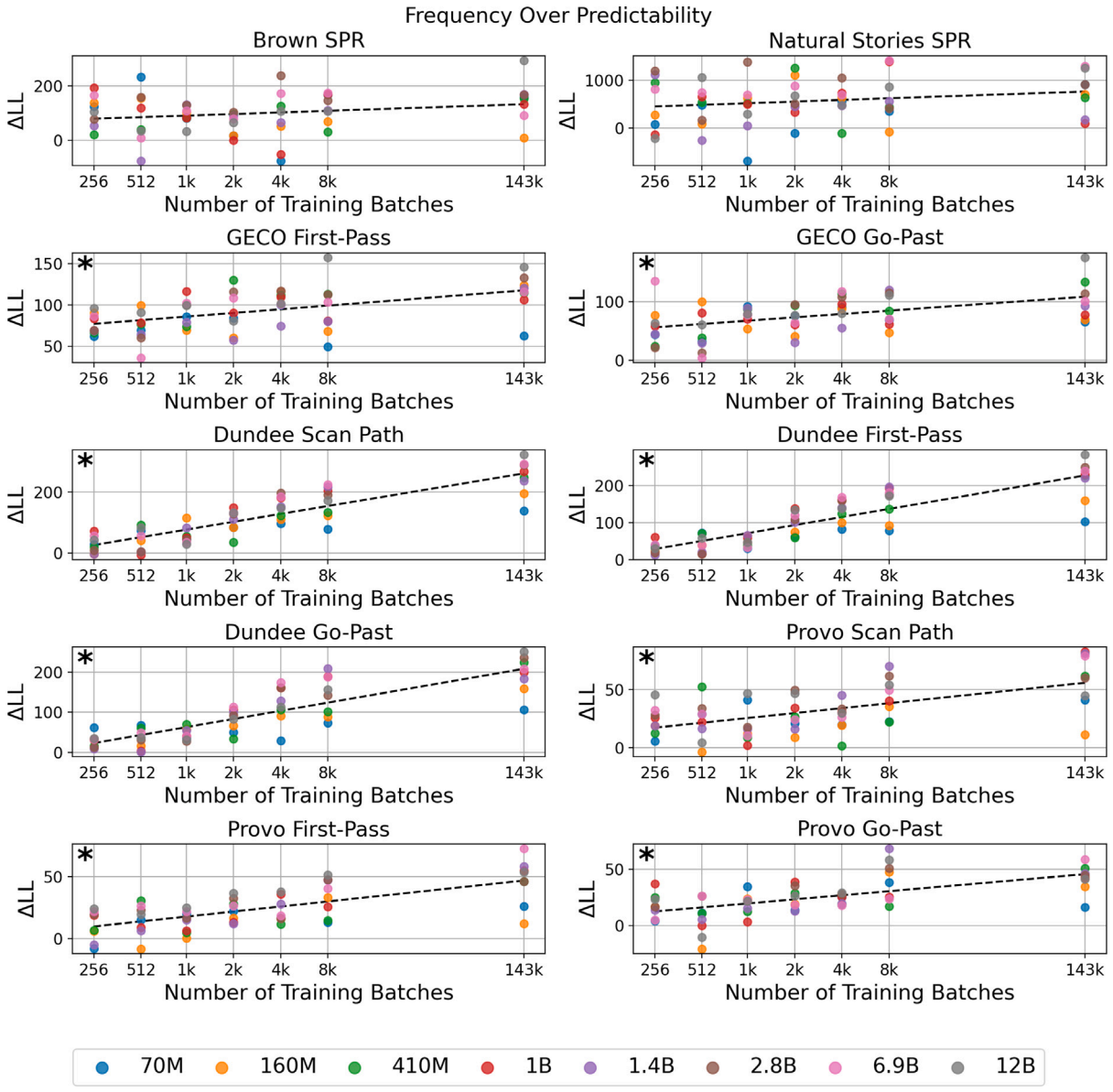


**Fig. 2.** Increase in CDR-NN model log-likelihood due to including unigram surprisal in the presence of surprisal estimates from Pythia LLMs as a function of number of training batches on each reading time dataset. Note that the $x$-axis is on log scale. Statistical significance of the slope of the best-fitting line is tested by a permutation test that permutes the independent variable. *: $p < 0.05$.

**Table 2**
Number of data points in the fit and exploratory partitions of the MECO dataset by each language. SP: scan path, FP: first-pass, GP: go-past, Total: total across three by-word measures.

| Measure (Partition) | NL | EN | FI | DE | EL | HE | IT | KO | NB | RU | ES | TR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SP (Fit) | 25,301 | 30,520 | 29,520 | 38,977 | 24,585 | 31,700 | 41,603 | 14,868 | 20,011 | 24,287 | 36,756 | 17,550 |
| SP (Exploratory) | 13,183 | 14,435 | 14,705 | 19,381 | 12,008 | 15,196 | 21,218 | 7,428 | 10,007 | 12,463 | 18,986 | 8,998 |
| FP/GP (Fit) | 3,176 | 9,256 | 11,636 | 12,069 | 8,166 | 8,225 | 12,411 | 4,530 | 6,502 | 10,218 | 7,283 | 4,491 |
| FP/GP (Exploratory) | 1,681 | 4,574 | 5,866 | 6,188 | 4,278 | 3,933 | 6,450 | 2,274 | 3,286 | 5,245 | 3,696 | 2,395 |
| Total (Fit) | 31,653 | 49,032 | 52,792 | 63,115 | 40,917 | 48,150 | 66,425 | 23,928 | 33,015 | 44,723 | 51,322 | 26,532 |
| Total (Exploratory) | 16,545 | 23,583 | 26,437 | 31,757 | 20,564 | 23,062 | 34,118 | 11,976 | 16,579 | 22,953 | 26,378 | 13,788 |

*Baseline predictors.* All regression models fit to eye-tracking data in MECO included word length in characters, as well as a binary variable indicating whether the previous word was fixated.

*Regression modeling*

Due to the smaller number of data points available in MECO, a series of linear mixed-effects (LME; Bates et al., 2015) models was fit to evaluate the contribution of word frequency over predictability. Their random effects structures also had to be kept simple due to convergence issues, which include by-subject random slopes for LLM surprisal and a by-subject random intercept. Similarly to Experiment 1, two sets of LME models were fit to reading times in the fit partition of each language subset. The 'full' models are LME models that include both unigram surprisal and LLM surprisal as described in Section "Predictors", and their corresponding 'ablated' models are those that ablate the unigram surprisal predictor. After model fitting, the $\Delta$LL on the exploratory

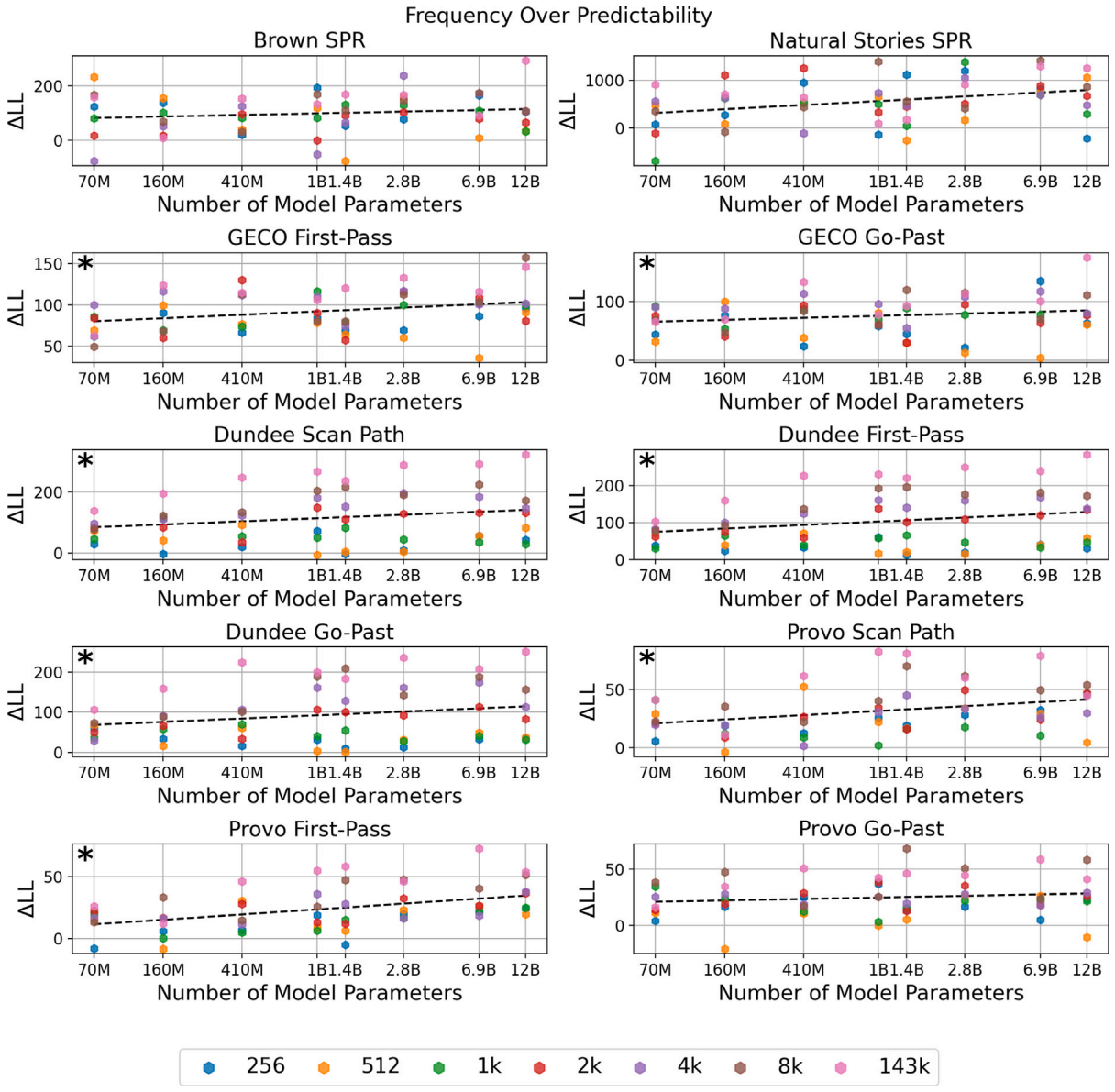


**Fig. 3.** Increase in CDR-NN model log-likelihood due to including unigram surprisal in the presence of surprisal estimates from Pythia LLMs as a function of number of model parameters on each reading time dataset. Note that the $x$-axis is on log scale. Statistical significance of the slope of the best-fitting line is tested by a permutation test that permutes the independent variable. *: $p < 0.05$.

partition of each language subset was calculated between a full model and its counterpart ablated model. To compensate for the smaller number of data points in MECO, ΔLL aggregated across the three by-word measures (i.e. scan path, first-pass, go-past) is reported.

*Results*

Again, we first present how well surprisal from each LLM variant predicts the reading times in each language subset over just the baseline predictors, as a measure of how strong they are as predictors of reading times. Fig. 4 replicates the adverse effect of model size on surprisal's fit to reading times strictly in 9 out of 12 languages (all except Dutch, English, and Korean), which is consistent with both Experiment 1 and previous work (Oh & Schuler, 2023b; Shain et al., 2024).

The results in Fig. 5 show that the trend observed in Experiment 1 is replicated in 11 out of 12 languages (all except Korean), where ΔLL due to unigram surprisal strictly increases as bigger LLMs are used to calculate surprisal. While both the number of reading time data points and the number of LLM variants are much smaller than in Experiment 1, the chance level probability of observing this trend across 12 languages is extremely low ($p < 0.001$).[9] This provides further support that the dissociability of frequency and predictability effects depends on the LLM used to calculate surprisal estimates.

*Discussion*

The second experiment provides additional evidence that the ability of frequency to predict reading times over and above predictability depends on the model size of the LLM used to calculate surprisal. The finding that the influence of model size replicates on 11 out of 12 languages examined further attests to the close relationship between LLM surprisal and frequency, and suggests that this is a general property of

[9] The chance level of observing this trend across 12 languages can be calculated as $\underbrace{\frac{1}{2!}}_{\text{Dutch}} \times \underbrace{\frac{1}{4!}}_{\text{English}} \times \underbrace{\frac{1}{4!}}_{\text{Finnish}} \times \ldots$.

**Table 3**
Hyperparameters and model sizes of LLM variants whose surprisal estimates were used in this experiment. #L, #H, and $d_{\text{model}}$ refer to number of layers, number of attention heads per layer, and embedding size, respectively. These models can be accessed on HuggingFace by appending each model identifier to the URL https://huggingface.co/. The English Pythia variants are those that were fully trained on 143,000 training batches.

| Language (Code) | Model identifier | #L | #H | $d_{\text{model}}$ | #Parameters |
|---|---|---|---|---|---|
| Dutch (NL)<br>Greek (EL)<br>Italian (IT) | ai-forever/mGPT | 24 | 16 | 2048 | ~1.4B |
| | ai-forever/mGPT-13B | 40 | 40 | 5120 | ~13.1B |
| English (EN) | EleutherAI/pythia-160m | 12 | 12 | 768 | ~160M |
| | EleutherAI/pythia-1b | 16 | 8 | 2048 | ~1B |
| | EleutherAI/pythia-2.8b | 32 | 32 | 2560 | ~2.8B |
| | EleutherAI/pythia-12b | 36 | 40 | 5120 | ~12B |
| Finnish (FI) | TurkuNLP/gpt3-finnish-small | 12 | 12 | 768 | ~186M |
| | TurkuNLP/gpt3-finnish-medium | 24 | 16 | 1024 | ~437M |
| | TurkuNLP/gpt3-finnish-large | 24 | 16 | 1536 | ~881M |
| | TurkuNLP/gpt3-finnish-xl | 24 | 24 | 2064 | ~1.5B |
| German (DE) | benjamin/gerpt2 | 12 | 12 | 768 | ~163M |
| | benjamin/gerpt2-large | 36 | 20 | 1280 | ~774M |
| Hebrew (HE) | Norod78/hebrew-gpt_neo-tiny | 6 | 12 | 768 | ~82M |
| | Norod78/hebrew-gpt_neo-xl | 24 | 16 | 2048 | ~1.3B |
| Korean (KO) | EleutherAI/polyglot-ko-1.3b | 24 | 16 | 2048 | ~1.3B |
| | EleutherAI/polyglot-ko-3.8b | 32 | 24 | 3072 | ~3.8B |
| | EleutherAI/polyglot-ko-5.8b | 28 | 16 | 4096 | ~5.9B |
| | EleutherAI/polyglot-ko-12.8b | 40 | 40 | 5120 | ~12.9B |
| Norwegian (NB) | NorGLM/NorGPT-369M | 24 | 16 | 1024 | ~370M |
| | NorGLM/NorGPT-3B | 32 | 32 | 2688 | ~3B |
| Russian (RU) | ai-forever/rugpt3small_based_on_gpt2 | 12 | 12 | 768 | ~125M |
| | ai-forever/rugpt3medium_based_on_gpt2 | 24 | 16 | 1024 | ~356M |
| | ai-forever/rugpt3large_based_on_gpt2 | 24 | 16 | 1536 | ~760M |
| Spanish (ES) | PlanTL-GOB-ES/gpt2-base-bne | 12 | 12 | 768 | ~124M |
| | PlanTL-GOB-ES/gpt2-large-bne | 36 | 20 | 1280 | ~773M |
| Turkish (TR) | ytu-ce-cosmos/turkish-gpt2 | 12 | 12 | 768 | ~124M |
| | ytu-ce-cosmos/turkish-gpt2-medium | 24 | 16 | 1024 | ~355M |
| | ytu-ce-cosmos/turkish-gpt2-large | 36 | 20 | 1280 | ~774M |

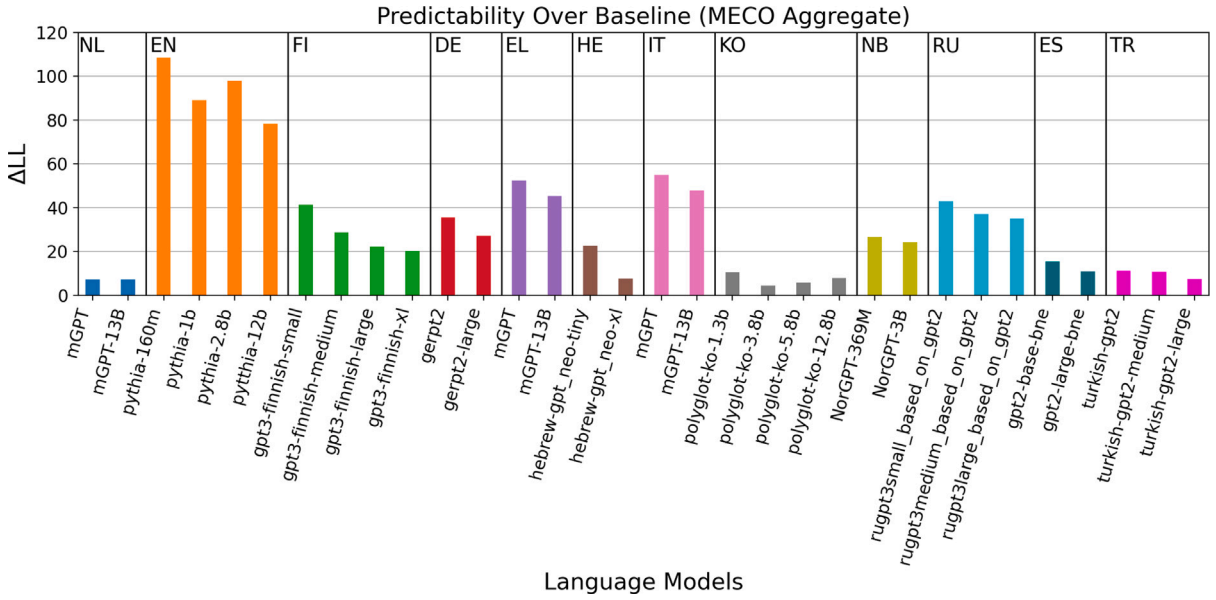


**Fig. 4.** Increase in LME model log-likelihood due to including surprisal from LLMs with different model sizes in the presence of baseline predictors.

language modeling that holds across languages. If so, the influence of training data amount may also replicate on reading time data across a wide range of languages.

One language subset that does not show this trend is Korean, despite the availability of monolingual LLMs that span a wider range of parameter count. A possible explanation for this is the incorrect tokenization of Korean words observed occasionally throughout the MECO dataset; the dataset treats some Korean words as consisting of multiple word regions, and therefore these words are incorrectly tokenized in the text stimuli.[10] This is likely to have resulted in incorrect estimates of both word frequency and surprisal that confound the experimental

[10] Some examples from just the first article include 야 누스 for 야누스 (*Janus*), 전 쟁시에는 for 전쟁시에는 (*at wartime*), and 들어오 는 for 들어오는 (*entering*).

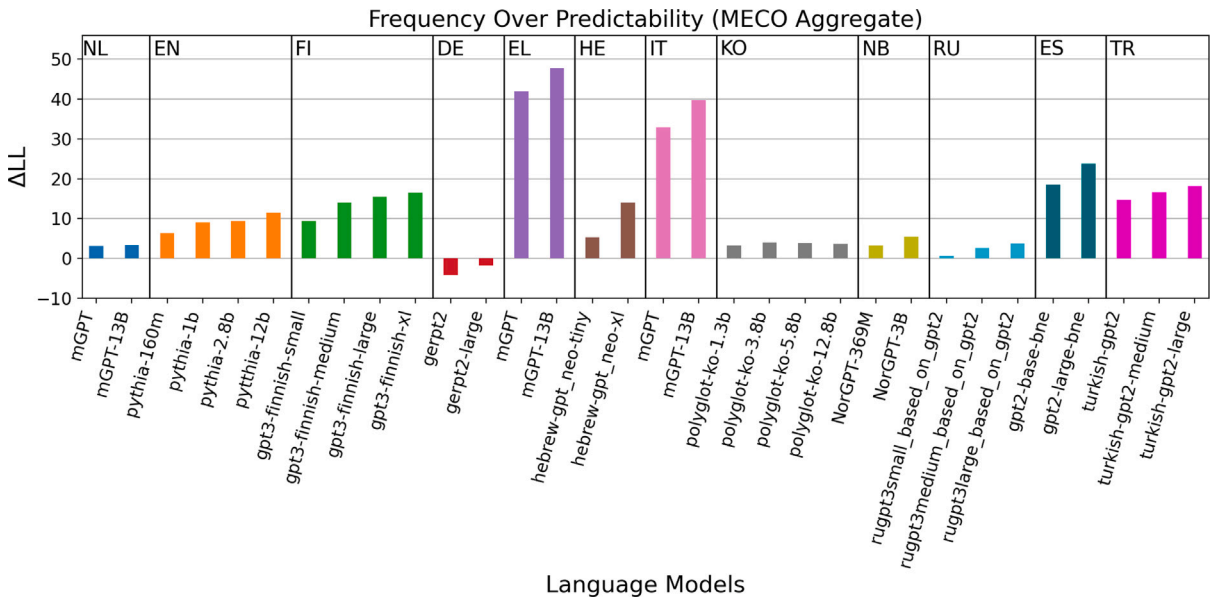


**Fig. 5.** Increase in LME model log-likelihood due to including unigram surprisal in the presence of surprisal estimates from LLMs with different model sizes.

results, with a potentially larger influence on surprisal due to the LLMs' dependence on previous incorrectly tokenized input.

## Follow-up analyses: Learning trajectories of low-frequency tokens

The two experiments revealed that as the LLMs for calculating surprisal become bigger and are trained on more data, the ability of word frequency to predict reading times reliably increases. This indicates that frequency compensates for surprisal either because bigger LLMs trained on more data systematically assign higher surprisal to high-frequency words or because they systematically assign lower surprisal to low-frequency words. Using the same corpora and LLMs as in Experiment 1, these follow-up analyses first establish that it is the latter of the two cases that primarily drives the effects of model size and training data amount. Subsequently, low-frequency tokens are further categorized into patterns that are informed by properties of the Transformer architecture or standard language modeling practices to study the impact of the two variables on token-level probabilities.

### *Analysis 1: The effects are driven by systematically lower surprisal assigned to low-frequency words*

#### *Methods*

The tokens in the five reading time corpora used in Experiment 1 were first divided into quintiles defined by token-level unigram surprisal estimated from the training data of the Pythia LLMs (Gao et al., 2020).[11] More specifically, counts from tokens in 16,000 training batches (~33 billion tokens of the Pile) were used to estimate token-level unigram probabilities over the vocabulary of the Pythia LLMs. Preliminary analyses showed that these ~33 billion tokens were enough to reliably estimate probabilities over 50,277 token types in the vocabulary. The average token-level surprisal on each quintile was then calculated using Pythia variants at various points throughout their training.

[11] Note that this is different from the word-level unigram surprisal that was used as a predictor of reading times in Experiment 1.

#### *Results*

Fig. 6 shows a clear interaction between token frequency, model size, and training data amount. The subset of most frequent tokens (top right panel) seem to be 'learned' early, with average surprisal dropping greatly with only 128 training batches. Moreover, after about 4000 training batches, additional training data does not seem to further improve the predictions of these tokens. On the other hand, the prediction of the least frequent tokens (bottom right panel) continue to be learned throughout the end of training, with the larger models showing a decrease in average surprisal between 64,000 and 143,000 training batches. In contrast, the smaller models do not seem to be able to learn to predict these tokens, even with large amounts of training data. The intermediate quintiles further support this general trend, with the difference in average surprisal as a function of model size and training data amount gradually increasing as token frequency decreases. These trends suggest that it is the systematically lower surprisal assigned to low-frequency words that drives the effects observed in the two experiments.

### *Analysis 2: Increase in model size helps LLMs form specific associations*

#### *Methods*

These results on low-frequency tokens pose additional questions about why limitations in model size appear to pose a bottleneck for learning to predict low-frequency tokens in spite of large amounts of training data. To answer these questions, low-frequency tokens were further categorized by properties that are informed by the inductive biases of the Transformer architecture or standard language modeling practices.

In these examples, $x$ denotes each subword token, and the language model uses the left context $x_{1..i-1}$ to predict $x_i$.

- *Repeated* tokens: Tokens $x_i$ that have been observed in the context, i.e. $x_k = x_i$ for some $k < i$. The Transformer architecture has access to veridical representations of the input sequence, unlike the class of recurrent neural networks that maintain and update a representation of fixed size. This property allows Transformers to reliably *repeat* input tokens that have been observed earlier in the sequence in comparison to architectures based on recurrence (Jelassi et al., 2024). The complement of this set is referred to as *non-repeated* tokens (i.e. tokens that are not part of the current input sequence).

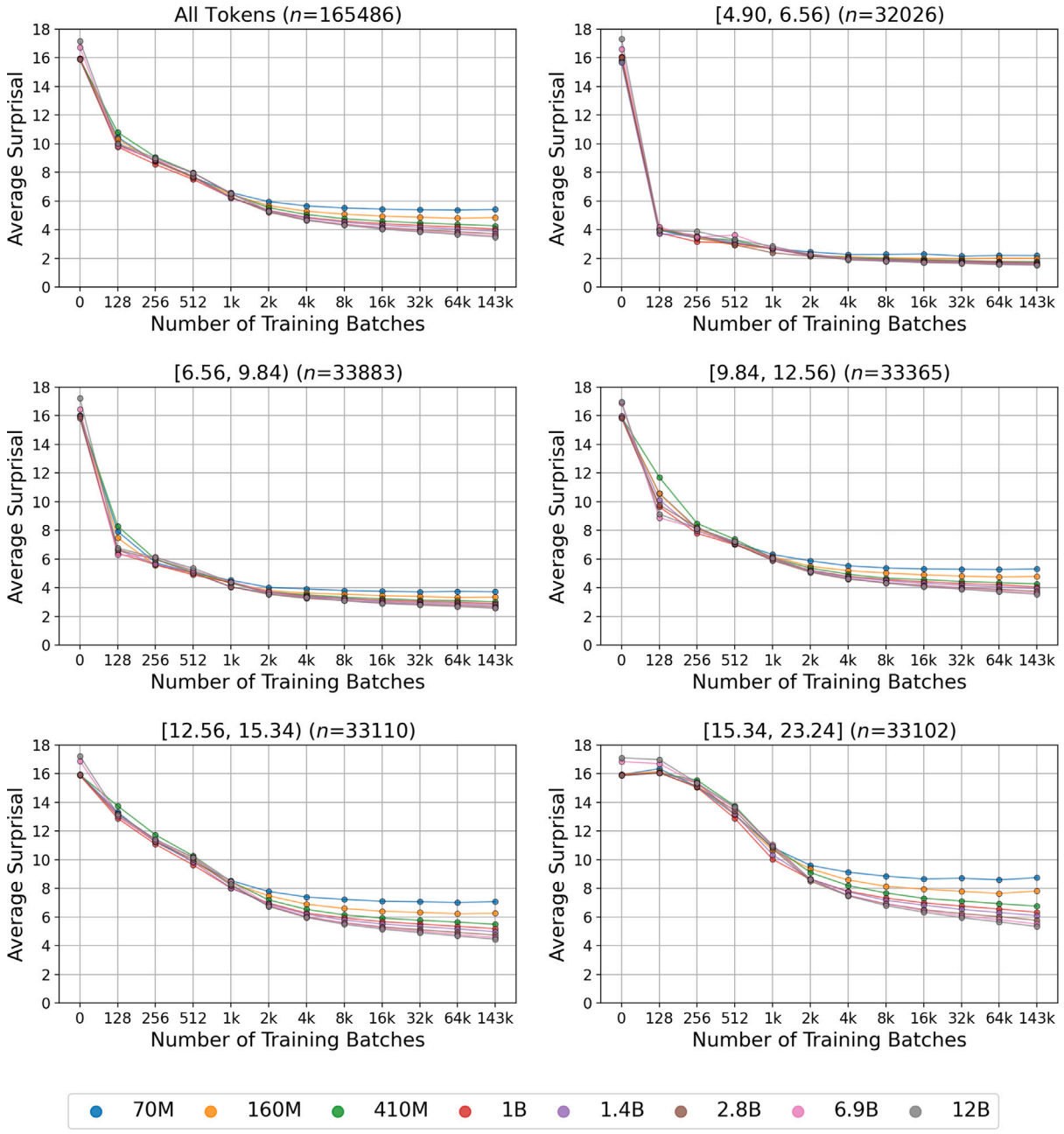


**Fig. 6.** Average token-level surprisal values from Pythia LLM variants aggregated across all reading time corpora (top left), and those by quintiles defined by token-level unigram surprisal. The titles of subplots denote the interval of each quintile.

- *Induction* tokens: Tokens $x_i$ that together with its preceding token $x_{i-1}$ have been observed in the context, i.e. $x_k = x_i$ and $x_{k-1} = x_{i-1}$ for some $k < i$. This is a subset of *repeated* tokens that are thought to be easier to predict due to the observation of $x_{i-1}$, which triggers the model to 'look back' at $x_{k-1}$ and 'copy over' $x_k$ to predict $x_i$ (Bietti et al., 2023; Olsson et al., 2022).
- *Word-internal* tokens: Tokens $x_i$ like *er* that are 'inside' of words like *miller*. Most contemporary LLMs use subword tokenization schemes to define their vocabulary, which allows arbitrary words to be handled by the model while keeping the vocabulary size tractable. This results in whitespace-delimited words that are tokenized into multiple tokens; *word-internal* tokens refer to those that are 'inside' whitespace-delimited words and include punctuation marks like commas and full stops. These tokens are likely to be predictable from the earlier subword token(s) of the same word.

The relationship between these categories is visualized in Fig. 7. Subsequently, the learning trajectories of subsets defined by these categories were analyzed to examine where the increase in model size and training data amount has the biggest influence on their accurate prediction.

*Results*

The results in Fig. 8 show that different subsets of tokens are learned at different points in training. Until Batch 512, all models seem to mainly learn to predict word-internal tokens that have not been seen in the context (Condition 1), where the largest decrease in average surprisal is observed. Subsequently, from Batch 512 to Batch 1000, a steep decrease in average surprisal occurs for word-internal tokens that are repeated from the input (Conditions 2 and 3). Across all models, there is very little change in average surprisal after this point, which indicates that these tokens are learned the earliest. From Batch

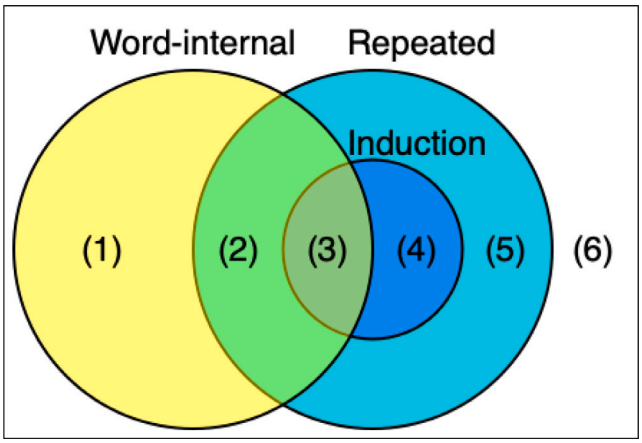


**Fig. 7.** Subsets defined by three properties that are thought to influence the learning of low-frequency tokens.

1000 to 2000, models appear to learn to predict repeated non-word-internal tokens (Conditions 4 and 5). Generally, tokens that follow the induction pattern (Conditions 3 and 4) are learned to be predicted more accurately, as can be seen by their lower average surprisal compared to their non-induction counterparts (Conditions 2 and 5) by the end of training. Finally, non-repeated non-word-internal tokens (Condition 6) seem to be learned most slowly and predicted least accurately, with average surprisal gradually decreasing throughout the course of training. By the end of training, tokens in Conditions 5 and 6 appear to show the biggest difference in average surprisal as a function of model size. These tokens are not part of a bigram sequence observed earlier in the context (i.e. non-repeated tokens, or repeated tokens in a different bigram context) in usually non-constraining contexts (cf. word-internal tokens).

*Discussion*

The follow-up analyses provide insight into the reasons underlying the trends observed in the two experiments. As a predictor of reading times, frequency appears to systematically compensate for lower surprisal assigned to low-frequency words, thereby achieving stronger fit in the presence of surprisal from bigger LLMs trained on more data. This is consistent with previous results showing that the adverse effects of model size and training data amount on surprisal's fit to reading times are driven by more severe mispredictions of reading times at low-frequency words (Oh et al., 2024). Additionally, this confirms the importance of training example frequency on LLMs' learned probabilities (Tirumala, Markosyan, Zettlemoyer, & Aghajanyan, 2022; Xia et al., 2023), and aligns with previous findings that neural language models first approximate unigram probabilities and then higher-order *n*-gram probabilities (Chang & Bergen, 2022; Chang, Tu, & Bergen, 2023) throughout training. Importantly for psycholinguistic modeling, model size and training data amount appear to shape predictability estimates from LLMs in a way that is relevant to extant research questions.

The second analysis delved more specifically into the predictive mechanisms that allow bigger LLMs to predict low-frequency tokens more accurately than their smaller counterparts. The learning trajectories of low-frequency tokens revealed that the influence of model size is strongest on tokens that are not repeated in the input, or are repeated in a different bigram context. To predict these tokens, models cannot simply copy observed bigrams, but rather must rely on information encoded in their parameters. If Transformer-based LLMs are viewed as associative memories between tokens (Bietti et al., 2023), this suggests that the limitations in model size create pressures toward learning more general associations, thereby resulting in less accurate predictions of rare tokens and better fits to reading times compared to larger models. Such tokens are likely to correspond to subsets of words that embody factual knowledge like named entity nouns, of which bigger LLMs have been shown to make 'superhuman' predictions (Oh & Schuler, 2023b).

## General discussion

While surprisal from Transformer-based LLMs is often used in psycholinguistic modeling work as predictors that operationalize predictability, factors like their number of parameters or amount of training data have a substantial influence on their fit to human reading times (Oh & Schuler, 2023a; Shain et al., 2024). Different sensitivity to word frequency has been proposed as an explanation for these trends (Oh et al., 2024), which has direct implications for using them to study the dissociability of frequency effects from predictability effects (Goodkind & Bicknell, 2021; Shain, 2019, 2024). This article illustrates this point through regression analyses using naturalistic reading time datasets spanning multiple genres and languages. The results show robust positive effects of LLM size and training data amount on the ability of word frequency to predict human reading times, which indicates that frequency predictors are compensating for surprisal predictors, and that this compensation decreases as LLM size and training data amount decrease and the fit of surprisal to reading times improves.

This may apply not only to the specific class of Transformer-based LLMs with different model sizes and training data amounts that were examined in this study, but more broadly to other language models with different prediction objectives (Arehalli, Dillon, & Linzen, 2022; Hale et al., 2018), neural network architectures (Merkx & Frank, 2021; Michaelov, Arnett, & Bergen, 2024; Wilcox et al., 2020), granularity of text prediction (Nair & Resnik, 2023; Oh, Clark, & Schuler, 2021; Oh & Schuler, 2024a), etc. With still largely open questions about modeling human-like predictive processing, drawing conclusions based on predictors from one particular computational model may be hasty. The finding that estimated frequency effects get larger in the presence of surprisal from LLMs that are bigger and trained on more data – which themselves are poorer predictors of reading times – raises the concern that the observed frequency effects may be due to imperfect estimates of predictability. As better-fitting language models are discovered by psycholinguists, dissociable frequency effects may disappear entirely.

In conjunction with exploring additional language models, efforts toward understanding their predictive mechanisms better are also required in order to make more precise claims about human sentence processing. An inherent shortcoming of studies using neural networks like LLMs to model human sentence processing is that it is difficult to explicate how the current state of the language processor influences the processing difficulty of the observed word. This is primarily because neural networks perform complex computations over continuous vector representations of each word, the dimensions of which are not directly interpretable in terms of e.g. linguistic abstractions. Therefore, studying the information encoded in the representations of these LLMs and how it influences their predictions (for overviews, see Belinkov, 2022; Linzen & Baroni, 2021; Rogers, Kovaleva, & Rumshisky, 2020) is important for characterizing their predictive mechanisms and appreciating them as models of human sentence processing.

In this spirit, the subsequent follow-up analyses examined how the increase in model size helps the prediction of low-frequency tokens. The Transformer architecture has veridical access to the input, which allows Transformer-based models to readily repeat sub-sequences from it. Additionally, the contemporary standard practice of using subword tokenizers results in highly constraining contexts such as 'insides' of long words, where the subsequent tokens are highly predictable given the first. An analysis of the learning trajectories of tokens categorized according to these properties showed that the influence of model size is strongest on tokens that are not part of a bigram sequence observed earlier in the context. As these tokens cannot be predicted by simply

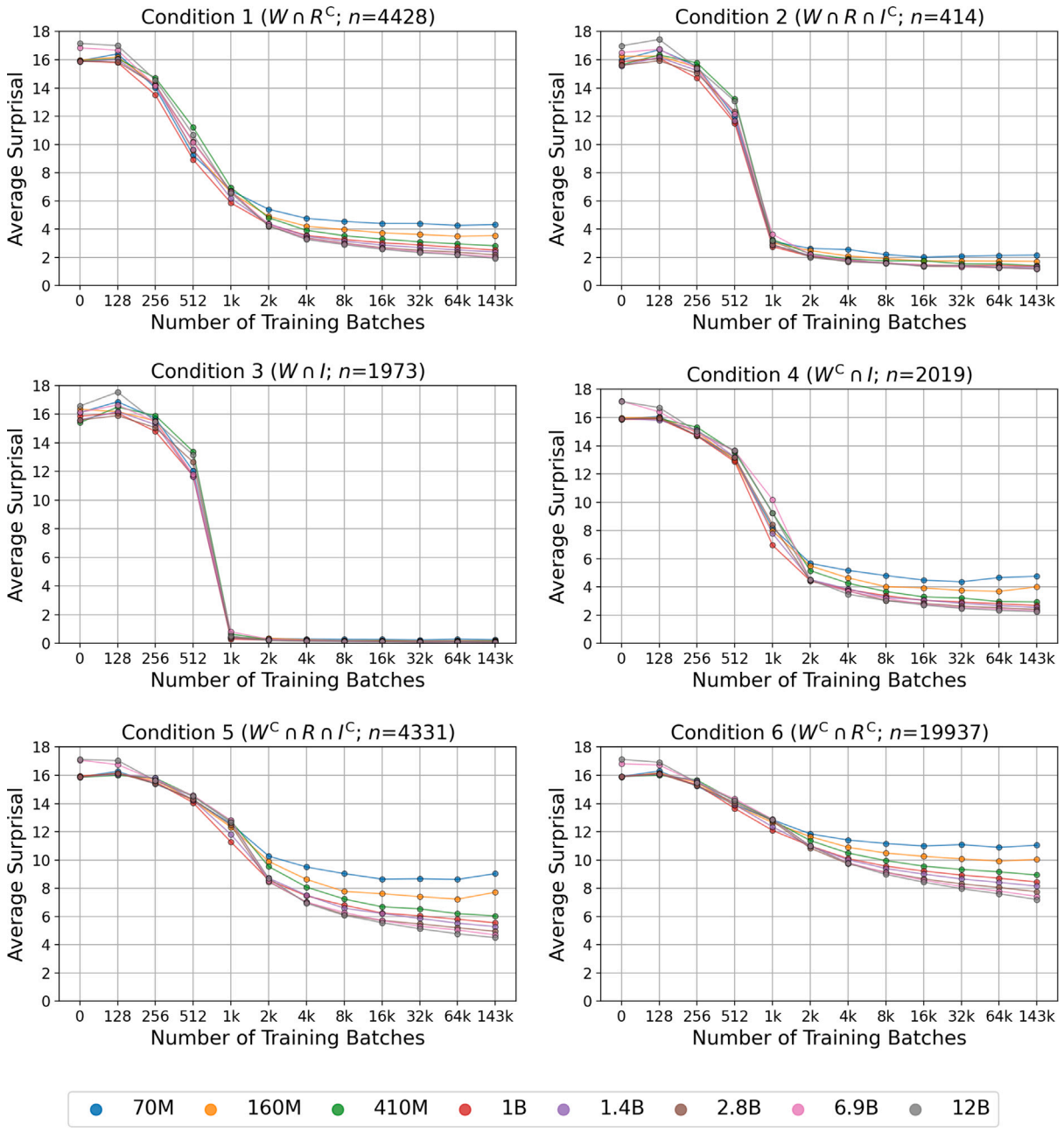


**Fig. 8.** Average token-level surprisal values from Pythia LLM variants on the quintile of least frequent tokens by each subset defined in Fig. 7. *R*: repeated, *W*: word-internal, *I*: induction. $^C$ denotes the complement set.

copying, this may be best explained from an associative-memory view of Transformers (Bietti et al., 2023); the limitations in the number of model parameters may enforce more general associations between tokens during training, which results in less accurate predictions of rare words and better fits to reading times.

These results suggest that more specific associations obtained from the training data make bigger LLMs less susceptible to incorrectly predicting low-frequency words compared to their smaller counterparts. Given that these models are trained on vast amounts of Internet text, this kind of association might correspond to factual knowledge that enables naturalistic text stimuli to be readily predicted. While such knowledge is an important determinant of an individual's reading behavior (Škrjanec, Broy, & Demberg, 2023; Smith, Snow, Serry, & Hammond, 2021) and that of their frequency effects (Monaghan, Chang, Welbourne, & Brysbaert, 2017), the knowledge embodied by LLMs is likely to be 'superhuman' compared to that of typical human readers (Oh & Schuler, 2023b). In this regard, applying model editing techniques (Meng, Bau, Andonian, & Belinkov, 2022; Wang et al., 2023) to tone down such knowledge in LLMs may be a promising direction for estimates of predictability that better align with measures of processing difficulty.

This is complementary to the perfect memory of the inference-time context (i.e. veridical access to the input) that is often considered to be problematic of LLMs under the framework of lossy-context surprisal (Futrell, Gibson, & Levy, 2020; Hahn, Futrell, Gibson, & Levy, 2022). The learning trajectories of low-frequency tokens show that models learn to accurately predict tokens that have been observed in the input regardless of size, indicating that they learn to implement a kind of copying mechanism that leverages their veridical access to the input (Elhage et al., 2021; Jelassi et al., 2024). This has been shown to cause a severe misalignment in modeling the processing behavior of repeated text (Gruteke Klein, Meiri, Shubi, & Berzak, 2024;

Vaidya, Turek, & Huth, 2023), suggesting that this is also problematic for human-like estimates of predictability. Explicitly implementing a form of decay to limit the models' access to earlier material in the input (Clark et al., 2025; de Varda & Marelli, 2024; Hahn et al., 2022) may alleviate this issue.

Finally, the fundamental source of LLM surprisal's sensitivity to frequency is likely to be the LLMs' training objective of *lexical* next-word prediction, which suggests another direction for improving LLMs as models of human sentence processing. That is, the training signal rewards only the correct prediction of the specific lexical form, which means that any linguistic abstraction has to be indirectly learned from distributional evidence. In contrast, humans make broader predictions based on linguistic abstractions such as semantic features, which have generally shown separable effects from the prediction of specific lexical forms during real-time processing (Federmeier & Kutas, 1999; Roland, Yun, Koenig, & Mauner, 2012). To close this gap, more models that make explicit predictions about linguistic abstractions (e.g. Oh et al., 2022) should be trained and evaluated against human subject data.

## Limitations

There are methodological choices adopted in this study that may limit the generalizability of the results. First, this study relies on reading time data of naturalistic text stimuli, which does not experimentally manipulate frequency and predictability. While the naturalistic reading paradigm may obtain higher ecological validity than controlled experiments, there may be confounding variables that are not controlled for by our regression modeling. Moreover, due to the partitioning protocols, both the CDR-NN and LME models are evaluated in terms of generalization likelihood across sentences for a given subject, but not across subjects. Future work could explore the generalizability of these results across subjects or datasets using different random effects structures.

In terms of the predictors, our experiments used token frequency as the measure of word frequency, when lemma frequency may reflect the ease of lexical access more accurately. However, estimates of lemma frequency are usually not available at scale, which limits their applicability to naturalistic corpora. Finally, the training data amounts of the Pythia LLMs used in Experiment 1 are potentially confounded by their learning rate schedule, which backpropagates the error from each training batch with different learning rates. As such, the different amounts of training data should be interpreted as lying on a continuum, with each training batch having a different degree of influence on the LLMs' learned probabilities.

## Conclusion

This article examines the dissociability of frequency and predictability effects through regression analyses on multiple reading time datasets using LLMs that vary in model size and training data amount. The results show that both factors have robust positive effects on the ability of word frequency to predict human reading times. This indicates that frequency compensates for surprisal, the degree to which decreases as the fit of surprisal predictors improves with smaller model sizes and fewer training data. The follow-up analyses that examined how the increase in model size helps the prediction of low-frequency tokens reveals that the influence of model size is strongest on tokens that are not part of a bigram sequence observed earlier in the context, which cannot be predicted by simply copying. This suggests that the limitations in model size may create pressures toward more general associations between tokens, which results in less accurate predictions.

Taken together, this study suggests that previously observed dissociable frequency effects may be due to imperfect estimates of predictability. As better-fitting, more human-like language models are developed by psycholinguists, these dissociable frequency effects may disappear entirely.

## CRediT authorship contribution statement

**Byung-Doh Oh:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **William Schuler:** Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

## Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Acknowledgments

This work was supported by the National Science Foundation grant #1816891. All views expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation. Computations for this work were mostly run using the Ohio Supercomputer Center (1987). This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

## Data availability

The reading time data used in these analyses is available at https://osf.io/8v5qb/ (Experiment 1) and https://osf.io/3527a/ (Experiment 2), code for CDR-NN modeling is available at https://github.com/coryshain/cdr, code for LME modeling is available at https://github.com/modelblocks/modelblocks-release, and code for LLM surprisal calculation is available at https://github.com/byungdoh/llm_surprisal.

## References

Agatha Christie (1920). *The mysterious affair at Styles*. John Lane, Retrieved from Project Gutenberg.

Arehalli, S., Dillon, B., & 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* (pp. 301–313).

Aurnhammer, C., & Frank, S. L. (2019). Comparing gated and simple recurrent neural network architectures as models of human sentence processing. In *Proceedings of the 41st annual meeting of the cognitive science society* (pp. 112–118).

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, *48*(1), 207–219.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., et al. (2023). Pythia: A suite for analyzing large language models across training and scaling. *202*, In *Proceedings of the 40th international conference on machine learning* (pp. 2397–2430).

Bietti, A., Cabannes, V., Bouchacourt, D., Jegou, H., & Bottou, L. (2023). Birth of a transformer: A memory viewpoint. *36*, In *Advances in neural information processing systems* (pp. 1560–1588).

Boston, M. F., Hale, J. T., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1), 1–12.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *33*, In *Advances in neural information processing systems* (pp. 1877–1901).

Chang, T. A., & Bergen, B. K. (2022). Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, *10*, 1–16.

Chang, T. A., Tu, Z., & Bergen, B. K. (2023). Characterizing learning curves during language model pre-training: Learning, forgetting, and stability. arXiv Preprint arXiv:2308.15419.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1724–1734).

Clark, C., Oh, B.-D., & Schuler, W. (2025). Linear recency bias during training improves transformers' fit to reading times. In *Proceedings of the 31st international conference on computational linguistics* (pp. 7735–7747).

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204.

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, *49*(2), 602–615.

de Varda, A. G., & Marelli, M. (2024). Locally biased transformers better align with human reading times. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 30–36).

Dyer, C., Kuncoro, A., Ballesteros, M., & 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* (pp. 199–209).

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

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., et al. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–225.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469–495.

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

Frank, S. L., Monsalve, I. F., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, *45*(4), 1182–1190.

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3).

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S., et al. (2021). The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, *55*, 63–77.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., et al. (2020). The Pile: An 800GB dataset of diverse text for language modeling. arXiv Preprint arXiv:2101.00027.

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* (pp. 95–126). Cambridge, MA: MIT Press.

Gokaslan, A., & Cohen, V. (2019). OpenWebText Corpus.

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

Goodkind, A., & Bicknell, K. (2021). Local word statistics affect reading times independently of surprisal. arXiv Preprint arXiv:2103.04469v2.

Gruteke Klein, K., Meiri, Y., Shubi, O., & Berzak, Y. (2024). The effect of surprisal on reading times in information seeking and repeated reading. In *Proceedings of the 28th conference on computational natural language learning* (pp. 219–230).

Hahn, M., Futrell, R., Gibson, E., & Levy, R. P. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43), Article e2122602119.

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., Dyer, C., Kuncoro, A., & Brennan, J. (2018). Finding syntax in human encephalography with beam search. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 2727–2736).

Hao, Y., Mendelsohn, S., Sterneck, R., Martinez, R., & Frank, R. (2020). Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the 10th workshop on cognitive modeling and computational linguistics* (pp. 75–86).

Hasson, U., & Honey, C. J. (2012). Future trends in neuroimaging: Neural processes as expressed within real-life contexts. *NeuroImage*, *62*(2), 1272–1278.

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 690–696).

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

Hofmann, M. J., Remus, S., Biemann, C., Radach, R., & Kuchinke, L. (2022). Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, *4*, Article 730570.

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

Jelassi, S., Brandfonbrener, D., Kakade, S. M., & Malach, E. (2024). Repeat after me: Transformers are better than state space models at copying. arXiv Preprint, arXiv:2402.01032v2.

Jin, L., & Schuler, W. (2020). Memory-bounded neural incremental parsing for psycholinguistic prediction. In *Proceedings of the 16th international conference on parsing technologies and the IWPT 2020 shared task on parsing into enhanced universal dependencies* (pp. 48–61).

Juhasz, B. J., & Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, *13*(7–8), 846–863.

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

Kennedy, A., Hill, R., & Pynte, J. (2003). The Dundee Corpus. In *Proceedings of the 12th European conference on eye movement*.

Kretzschmar, F., Schlesewsky, M., & Staub, A. (2015). Dissociating word frequency and predictability effects in reading: Evidence from coregistration of eye movements and EEG. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(6), 1648–1662.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59.

Kuribayashi, T., Oseki, Y., Brassard, A., & Inui, K. (2022). Context limitations make neural language models more human-like. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 10421–10436).

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

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

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

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*, 195–212.

Liu, T., Škrjanec, I., & Demberg, V. (2024). Temperature-scaling surprisal estimates improve fit to human reading times - But does it do so for the "right reasons"? In *ICLR 2024 workshop on representational alignment*.

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

McDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, *43*(16), 1735–1751.

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. *35*, In *Advances in neural information processing systems*.

Merkx, D., & Frank, S. L. (2021). Human sentence processing: Recurrence or attention? In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 12–22).

Michaelov, J. A., Arnett, C., & Bergen, B. K. (2024). Revenge of the fallen? Recurrent models match transformers at predicting human language comprehension metrics. arXiv Preprint arXiv:2404.19178.

Monaghan, P., Chang, Y.-N., Welbourne, S., & Brysbaert, M. (2017). Exploring the relations between word frequency, language exposure, and bilingualism in a computational model of reading. *Journal of Memory and Language*, *93*, 1–21.

Nair, S., & Resnik, P. (2023). Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship? In *Findings of the association for computational linguistics: EMNLP 2023* (pp. 11251–11260).

Oh, B.-D., Clark, C., & Schuler, W. (2021). Surprisal estimators for human reading times need character models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (pp. 3746–3757).

Oh, B.-D., Clark, C., & Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, *5*, Article 777963.

Oh, B.-D., & 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* (pp. 1915–1921).

Oh, B.-D., & 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., & Schuler, W. (2024a). The impact of token granularity on the predictive power of language model surprisal. arXiv Preprint arXiv:2412.11940.

Oh, B.-D., & Schuler, W. (2024b). Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 3464–3472).

Oh, B.-D., Yue, S., & 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* (pp. 2644–2663).

Ohio Supercomputer Center (1987). Ohio Supercomputer Center.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., et al. (2022). In-context learning and induction heads. *Transformer Circuits Thread*.

OpenAI (2023). GPT-4 technical report. *OpenAI Technical Report*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *35*, In *Advances in neural information processing systems* (pp. 27730–27744).

Pimentel, T., & Meister, C. (2024). How to compute the probability of a word. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 18358–18375).

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

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 358–374.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*(3), 191–201.

Reichle, E., Pollatsek, A., Fisher, D., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*, 125–157.

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 324–333).

Robyn Speer (2022). Rspeer/wordfreq: v3.0.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866.

Roland, D., Yun, H., Koenig, J.-P., & Mauner, G. (2012). Semantic similarity, predictability, and models of sentence processing. *Cognition*, *122*, 267–279.

van Schijndel, M., Exley, A., & Schuler, W. (2013). A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, *5*(3), 522–540.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics* (pp. 1715–1725).

Shain, C. (2019). A large-scale study of the effects of word frequency and predictability in naturalistic reading. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 4086–4094).

Shain, C. (2021). CDRNN: Discovering complex dynamics in human language processing. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (pp. 3718–3734).

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

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

Shain, C., & Schuler, W. (2021). Continuous-time deconvolutional regression for psycholinguistic modeling. *Cognition*, *215*, Article 104735.

Shain, C., & Schuler, W. (2024). A deep learning approach to analyzing continuous-time cognitive processes. *Open Mind*, *8*, 235–264.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Shliazhko, O., Fenogenova, A., Tikhonova, M., Kozlova, A., Mikhailov, V., & Shavrina, T. (2024). mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, *12*, 58–79.

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H.-D., Alexeeva, S., Amenta, S., et al. (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, *54*(4), 2843–2863.

Škrjanec, I., Broy, F. Y., & Demberg, V. (2023). Expert-adapted language models improve the fit to reading times. In *Proceedings of the 27th international conference on knowledge-based and intelligent information & engineering systems (KES 2023)* (pp. 3488–3497).

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

Smith, R., Snow, P., Serry, T., & Hammond, L. (2021). The role of background knowledge in reading comprehension: A critical review. *Reading Psychology*, *42*(3), 214–240.

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

Staub, A. (2011). The effect of lexical predictability on distribution of eye fixation durations. *Psychonomic Bulletin & Review*, *18*(2), 371–376.

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., & Benatar, A. (2013). Individual differences in fixation duration distributions in reading. *Psychonomic Bulletin & Review*, *20*(6), 1304–1311.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433.

Tirumala, K., Markosyan, A., Zettlemoyer, L., & Aghajanyan, A. (2022). Memorization without overfitting: Analyzing the training dynamics of large language models. *35*, In *Advances in neural information processing systems* (pp. 38274–38290).

Vaidya, A., Turek, J., & Huth, A. (2023). Humans and language models diverge when predicting repeating text. In *Proceedings of the 27th conference on computational natural language learning* (pp. 58–69).

Vasishth, S. (2006). On the proper treatment of spillover in real-time reading studies: Consequences for psycholinguistic theories. In *Proceedings of the international conference on linguistic evidence* (pp. 96–100).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *30*, In *Advances in neural information processing systems* (pp. 6000–6010).

Wang, S., Zhu, Y., Liu, H., Zheng, Z., Chen, C., & Li, J. (2023). Knowledge editing for large language models: A survey. arXiv Preprint arXiv:2310.16218.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. P. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd annual meeting of the cognitive science society* (pp. 1707–1713).

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.

Xia, M., Artetxe, M., Zhou, C., Lin, X. V., Pasunuru, R., Chen, D., et al. (2023). Training trajectories of language models across scales. In *Proceedings of the 61st annual meeting of the association for computational linguistics* (pp. 13711–13738).