Word Frequency Modulates the Effects of Model Size and Training Data Amount on Language Model Surprisal

Byung-Doh Oh[1] (oh.b@nyu.edu) and William Schuler[2] [1]NYU, [2]OSU

In cognitive modeling, word-by-word surprisal is often used as a predictor of processing difficulty, under a theoretical framework that emphasizes the predictive aspect of real-time language processing [5, 7]. In recent years, neural network-based language models (LMs) have been used to calculate and evaluate surprisal against human reading times [9, 17], which has opened possibilities for refining them as computational models of language processing and using them to study how predictive processing interacts with other cognitive processes.

Within this line of research, studies using surprisal from Transformer-based LMs have revealed a strong inverse correlation between the size of LMs and the fit of their surprisal to naturalistic reading times [11, 14]. Large amounts of training data have also been shown to play a detrimental role, with surprisal's fit to reading times starting to degrade after LMs see about two billion tokens of training data [10]. This degradation due to large amounts of training data seems to be alleviated by limitations in model size, with surprisal from smaller models degrading less severely compared to surprisal from their larger counterparts.

Motivated by recent studies that highlight the role of data frequency on the probabilities learned by LMs [16, 18], we propose word frequency as a unifying explanation for the inverse correlation between LMs' size, training data amount, and surprisal's fit to reading times. That is, after LMs learn to predict frequent words accurately during early training, large amounts of training data help them predict rare words excessively accurately, which drives the adverse effect of training data amount. However, during this later stage of training, limitations in model size constrain this excessive accuracy, thereby preventing surprisal from smaller LMs from degrading as quickly as surprisal from larger LMs. For regression models fit to human reading times, this makes two concrete predictions. The first is that the difference in fit to reading times will be the greatest on the subset of the rarest words across LMs of different sizes and training data amounts, where they make the most divergent predictions. The second is that the same predictor that operationalizes word frequency will demonstrate differential fits to reading times depending on which LM surprisal is included in the regression model.

We tested these two predictions using Pythia English LMs [2] of four different sizes (160M, 1B, 2.8B, 12B parameters) at five different points in later training (after 1,000, 2,000, 4,000, 8,000, 143,000 training batches of $\sim$2M tokens each) and five English self-paced reading and eye-tracking corpora [3, 4, 6, 8, 15]. For each of the ten reading time measures,[1] linear mixed-effects (LME) models were fit to approximately half of the data points using LM surprisal, unigram surprisal, and standard baseline predictors with maximal random effects supported by the data [1, Table 1]. Subsequently, held-out errors were calculated on the exploratory sets containing about 25% of the data points and further separated according to quintiles defined by unigram frequency to test the first prediction. Additionally, a new set of LME models without unigram surprisal was fit to the same data, and the increase in held-out log-likelihood (ΔLogLik) attributable to unigram surprisal was calculated on the exploratory sets to test the second prediction.

The regression modeling results provide support for both predictions, with mean squared errors on the subset of the rarest words showing the largest influence of LM size and training data amount (Figure 1a), and word frequency providing stronger fits to reading times over surprisal from larger LMs trained on more data (Figure 1c). These results provide insights into the factors that shape LM surprisal, and have important implications for using it to study whether frequency effects dissociate from predictability effects in naturalistic reading [12, 13].

[1]Self-paced reading times [4, 15]; first-pass, go-past durations [3, 6, 8]; scan path durations [6, 8]

Datasets	LME Formula
Self-paced [4, 15]	`RT ~ LMsurp + LMsurp_prev + Unisurp + length + index + (LMsurp + LMsurp_prev + length + index + 1 | subject)`
Eye-tracking [3, 6, 8]	`RT ~ LMsurp + LMsurp_prev + Unisurp + length + index + pfix + (LMsurp + index + 1 | subject)`

Table 1: LME formulae used in the experiments. `index`: position of the word within the sentence, `pfix`: whether the previous word was fixated. All predictors were z-transformed.

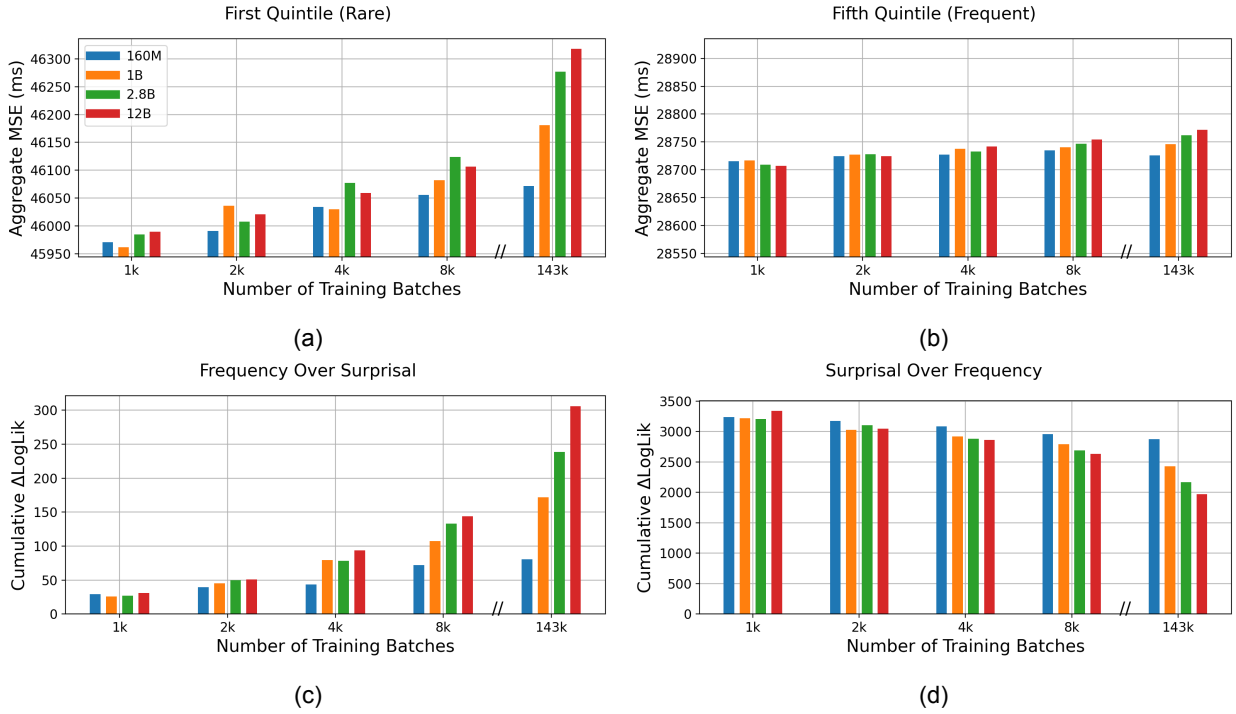

Figure 1: Held-out mean squared errors from LME models aggregated over ten reading time datasets, on the first and fifth word frequency quintiles (a and b; zoomed to equal scale on y-axis for clarity), increase in held-out log-likelihood due to unigram surprisal over surprisal from different LMs (c), increase in held-out log-likelihood due to surprisal from different LMs over unigram surprisal (d). The effects of model size and training data amount on MSE of data points in the first quintile (a) is significant at $p < 0.001$ level by a permutation test that permutes quintile membership, and the interaction term between model size and number of training batches is a significant predictor of cumulative ΔLogLik at $p < 0.001$ level (c).

[1] Barr, D. J., Levy, R., Scheepers, C., et al. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal.
[2] Biderman, S., Schoelkopf, H., Anthony, Q. G., et al. 2023. Pythia: A suite for analyzing large language models across training and scaling.
[3] Cop, U., Dirix, N., Drieghe, D., et al. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading.
[4] Futrell, R., Gibson, E., Tily, H. J., et al. 2021. The Natural Stories Corpus: A reading-time corpus of English texts containing rare syntactic constructions.
[5] Hale, J. 2001. A probabilistic Earley parser as a psycholinguistic model.
[6] Kennedy, A., Hill, R., & Pynte, J. 2003. The Dundee Corpus.
[7] Levy, R. 2008. Expectation-based syntactic comprehension.
[8] Luke, S. G., & Christianson, K. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms.
[9] Merkx, D., & Frank, S. L. 2021. Human sentence processing: Recurrence or attention?.
[10] Oh, B.-D., & Schuler, W. 2023. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens.
[11] Oh, B.-D., & Schuler, W. 2023. Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times?.
[12] Shain, C. 2019. A large-scale study of the effects of word frequency and predictability in naturalistic reading.
[13] Shain, C. 2024. Word frequency and predictability dissociate in naturalistic reading.
[14] Shain, C., Meister, C., Pimentel, T., et al. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time.
[15] Smith, N. J., & Levy, R. 2013. The effect of word predictability on reading time is logarithmic.
[16] Tirumala, K., Markosyan, A., Zettlemoyer, L., et al. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models.
[17] Wilcox, E. G., Gauthier, J., Hu, J., et al. 2020. On the predictive power of neural language models for human real-time comprehension behavior.
[18] Xia, M., Artetxe, M., Zhou, C., et al. 2023. Training trajectories of language models across scales.