Correcting Language Model Word Probabilities Reveals a Greater Divergence Between Surprisal and Human Reading Times

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

Key Takeaways

1. Most English language models (LMs) build whitespaces directly into the front of their token
2. P(*mat* | *I was a*) is often calculated as P(␣mat | <s> I ␣was ␣a), resulting in inconsistent word probabilities
3. Correcting this to P(mat ␣ | <s> I ␣was ␣a ␣) reveals a larger divergence of LM surprisal from reading times

More in Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. *Proc. EMNLP.*

Word probabilities in psycholinguistics

Word probabilities allow us to evaluate what LMs learn, and study real-time processing difficulty in humans

Word	*If*	*you*	*were*	*to*
Reading Time	360 ms	304 ms	270 ms	292 ms
LM1 Surprisal	7.76	0.81	5.42	2.09
LM2 Surprisal	6.71	0.78	5.22	2.30
LM3 Surprisal	7.10	0.56	5.15	2.39

Problem: Inconsistent word probabilities

Most English tokenizers [10] have *leading whitespaces*, which has resulted in the common practice of:

$$P(\textit{mat} \mid \textit{I was a}) = P(\text{␣mat} \mid \text{<s> I ␣was ␣a}) \quad (1)$$

$$P(\textit{matron} \mid \textit{I was a}) = P(\text{␣mat ron} \mid \text{<s> I ␣was ␣a}) = P(\text{␣mat} \mid \text{<s> I ␣was ␣a}) \cdot P(\text{ron} \mid \text{<s> I ␣was ␣a ␣mat}) \quad (2)$$

Under this practice,
P(*mat* | *I was a*) ≥ P(*matron* | *I was a*), and
P(*mat* | *I was a*) + P(*matron* | *I was a*) can exceed 1

Solution: Whitespace-trailing decoding

The probability of the *trailing whitespace* should be accounted for as part of the word probability instead:

$$P(\textit{mat} \mid \textit{I was a}) = P(\text{mat ␣} \mid \text{<s> I ␣was ␣a ␣}) = P(\text{␣mat} \mid \text{<s> I ␣was ␣a}) \cdot \frac{P(\text{␣} \mid \text{<s> I ␣was ␣a ␣mat})}{\underbrace{P(\text{␣} \mid \text{<s> I ␣was ␣a})}_{\text{sum over probabilities of ␣ tokens}}} \quad (3)$$

$$P(\textit{matron} \mid \textit{I was a}) = P(\text{mat ron ␣} \mid \text{<s> I ␣was ␣a ␣}) = P(\text{␣mat} \mid \text{<s> I ␣was ␣a}) \cdot P(\text{ron} \mid \text{<s> I ␣was ␣a ␣mat}) \cdot \frac{P(\text{␣} \mid \text{<s> I ␣was ␣a ␣mat ron})}{\underbrace{P(\text{␣} \mid \text{<s> I ␣was ␣a})}_{\text{sum over probabilities of ␣ tokens}}} \quad (4)$$

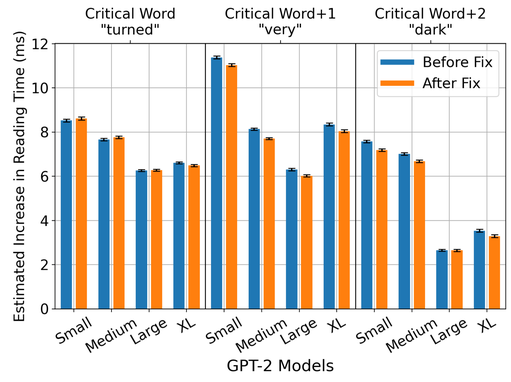
Now, P(*mat* | *I was a*) and P(*matron* | *I was a*) compete for probability, and the sum of all word probabilities equals 1 (proof in Oh and Schuler [8])

Exp. 1: Revisiting LMs' garden path effects

After the doctor left the room turned very dark ...
After the doctor left, the room turned very dark ...

LMs severely underpredict the difficulty at turned [4]

Increase in reading time across conditions estimated with GPT-2 LMs [9], following Huang et al. [4]

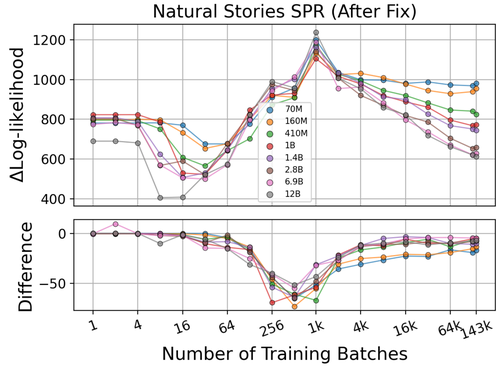


cf. Human effect sizes: ~120 ms, ~150 ms, ~65 ms

Exp. 2: Revisiting LMs' fit to naturalistic RT

After a certain point, deleterious effects of LM size and training data amount on fit to naturalistic RT [7]

Surprisal from Pythia LMs [1] fit to Natural Stories self-paced reading time data [3], following Oh and Schuler [7]



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

Key Takeaways

1. Increased LM size and training data help accurate predictions of low-frequency words. Therefore,
2. Consequence 1: Difference in fit to reading times is the largest on the subset of low-frequency words
3. Consequence 2: Word frequency shows differential fits to reading times depending on baseline surprisal

More in Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. *Proc. EACL.*; Dissociable frequency effects attenuate as large language model surprisal predictors improve. *Under review at JML.*

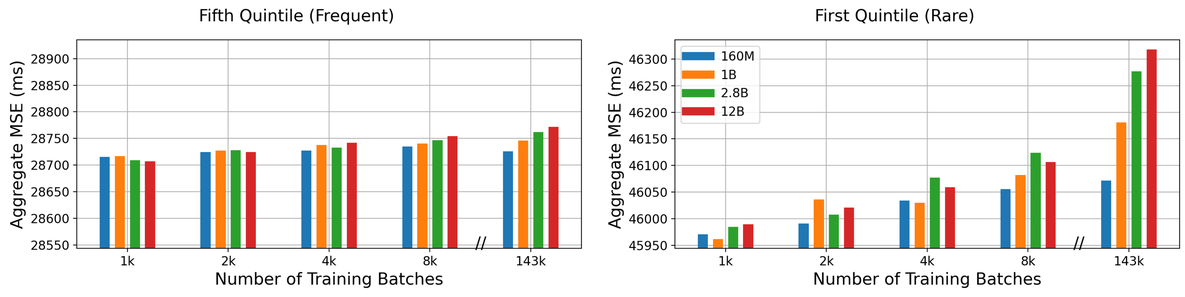
Word frequency has a strong influence on LM probabilities

Larger LMs boost probabilities more given the same training data [13], assigning lower surprisal to low-frequency words

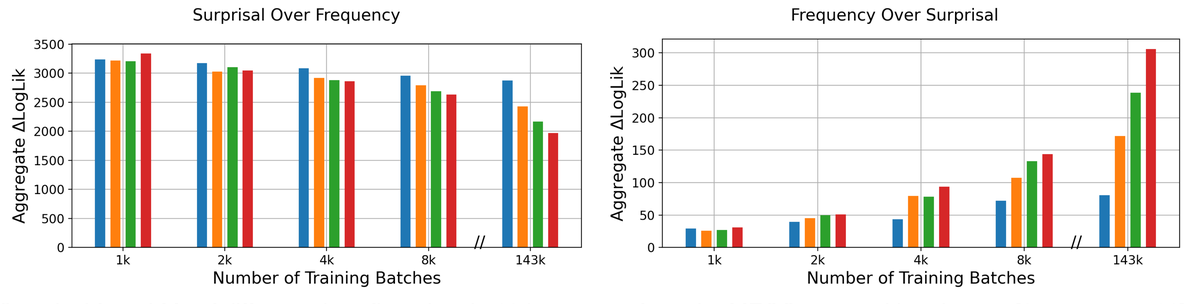
This leads to two consequences when modeling reading times using LM surprisal and word frequency

Consequence 1: Difference in errors is the largest on low-frequency words

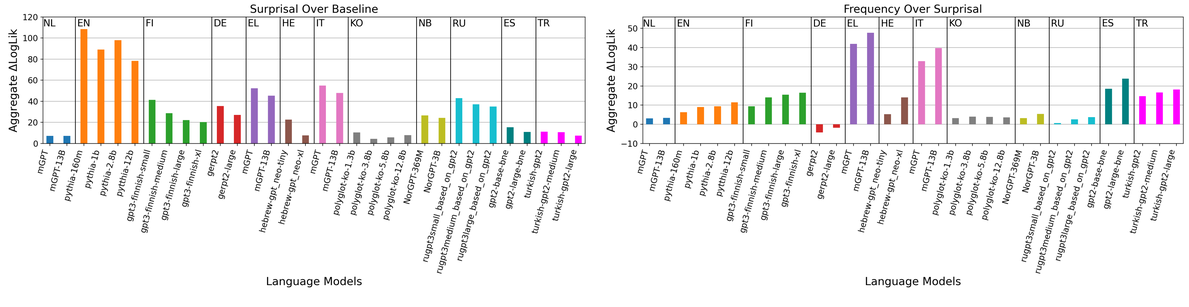
Surprisal from Pythia LMs [1] fit to five self-paced reading and eye-tracking datasets [10 measures; 2, 3, 5, 6, 12]



Consequence 2: Word frequency compensates for LM surprisal



Surprisal from LMs of different sizes fit to data in 12 languages from the MECO eye-tracking dataset [3 measures; 11]



[1] Biderman, S., Schoelkopf, H., Anthony, Q. G., et al. 2023. Pythia: A suite for analyzing large language models across training and scaling.

[2] Cop, U., Dirix, N., Drieghe, D., et al. 2017. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading.

[3] 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.

[4] Huang, K.-J., Arehalli, S., Kugemoto, M., et al. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty.

[5] Kennedy, A., Hill, R., & Pynte, J. 2003. The Dundee Corpus.

[6] Luke, S. G., & Christianson, K. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms.

[7] Oh, B.-D., & Schuler, W. 2023. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens.

[8] Oh, B.-D., & Schuler, W. 2024. Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities.

[9] Radford, A., Wu, J., Child, R., et al. 2019. Language models are unsupervised multitask learners.

[10] Sennrich, R., Haddow, B., & Birch, A. 2016. Neural machine translation of rare words with subword units.

[11] Siegelman, N., Schroeder, S., Acartürk, C., et al. 2022. Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO).

[12] Smith, N. J., & Levy, R. 2013. The effect of word predictability on reading time is logarithmic.

[13] Tirumala, K., Markosyan, A., Zettlemoyer, L., et al. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models.