Leading Whitespaces of Language Models' Subword Vocabulary Pose a Confound for Calculating Word Probabilities

Byung-Doh Oh William Schuler

oh.b@nyu.edu
github.com/byungdoh/wt_decoding

New York University The Ohio State University

PAPER

Key Takeaways

1. Most English subword tokenizers build whitespaces directly into the front of each token
2. So P(*mat* | *I was a*) is often calculated as P(␣mat | <s> I ␣was ␣a), which results in inconsistent word probabilities
3. Correcting this to P(mat ␣ | <s> I ␣was ␣a ␣) reveals a larger divergence of LMs from human-like language processing

Word probabilities in interpretability and cognitive modeling

Word probabilities are important for evaluating what LMs learn,

> P(*are* | *The keys to the cabinet*)
> P(*is* | *The keys to the cabinet*)

Linzen et al. [5]

and for studying what influences real-time processing difficulty in humans

Word	*If*	*you*	*were*	*to*	*journey*
Reading Time	360 ms	304 ms	270 ms	292 ms	304 ms
LM1 Surprisal	7.76	0.81	5.42	2.09	14.62
LM2 Surprisal	6.71	0.78	5.22	2.30	13.93
LM3 Surprisal	7.10	0.56	5.15	2.39	15.02

Oh and Schuler [7]

Problem: Inconsistent word probabilities

Most English subword tokenizers [e.g. 9] build whitespaces into the front of each token, which has resulted in the common practice of:

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

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

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 instead

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

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

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 paper)

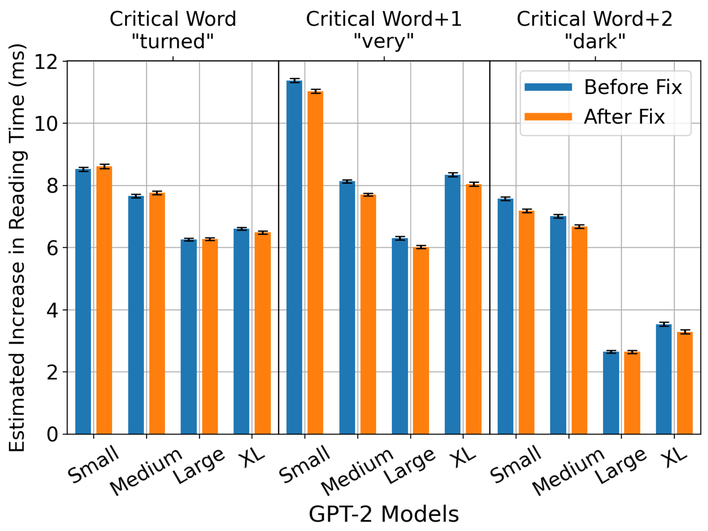
Experiment 1: Revisiting LMs' garden path effects

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

Mitchell [6], Gorrell [3]

People show high processing difficulty at turned due to syntactic disambiguation, but LMs severely underpredict this difficulty [4]

Increase in reading time across conditions estimated with GPT-2 LMs [8], using data and procedures of Huang et al. [4]

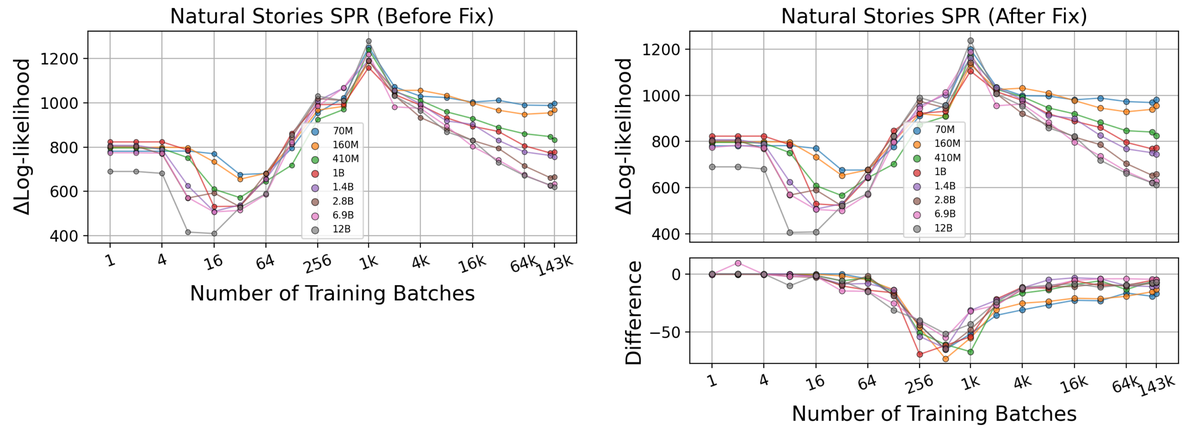


cf. People show effects of ~120 ms, ~150 ms, ~65 ms in the three regions

Experiment 2: Revisiting LMs' fit to naturalistic reading times

After a certain point, surprisal from LMs with more parameters and trained on more data yield poorer fits to naturalistic reading times [7]

Surprisal calculated using Pythia LMs [1] and regressed to Natural Stories reading time data [2], following the procedures of Oh and Schuler [7]



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

[2] 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.

[3] Gorrell, P. 1991. Subcategorization and sentence processing.

[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] Linzen, T., Dupoux, E., & Goldberg, Y. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies.

[6] Mitchell, D. C. 1987. Lexical guidance in human parsing: Locus and processing characteristics.

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

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

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