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

Word-by-word conditional probabilities from Transformer-based language models (LMs) are increasingly being used to model the incremental processing difficulty of human readers [9, 12]. However, due to the way many contemporary LMs tokenize raw strings into tokens that can be processed, there is a confound in calculating word probabilities correctly that has been overlooked.

To allow LMs to flexibly handle unseen word forms and to keep the vocabulary size manageable, it has become a standard practice in language modeling to use 'subword' vocabularies [11]. Such vocabularies are built based on corpus statistics such that they contain frequent character sequences (which may or may not correspond to words) as independent tokens. A common design choice in this process is to prepend the whitespace character to tokens, such that tokens have leading whitespaces. Consider how the GPT-2 English LM [10] tokenizes the following minimal pair of sentences:

(1) a. `I ␣was ␣a ␣mat ron ␣in ␣France .`
    b. `I ␣was ␣a ␣mat ␣in ␣France .`

Under this tokenization, a common way to calculate $\mathrm{P}(\textit{mat} \mid \textit{I was a})$ and $\mathrm{P}(\textit{matron} \mid \textit{I was a})$ is by calculating $\mathrm{P}(\texttt{␣mat} \mid \texttt{I ␣was ␣a})$ and $\mathrm{P}(\texttt{␣mat ron} \mid \texttt{I ␣was ␣a})$ respectively. However, as $\mathrm{P}(\texttt{␣mat ron} \mid \texttt{I ␣was ␣a}) = \mathrm{P}(\texttt{␣mat} \mid \texttt{I ␣was ␣a}) \cdot \mathrm{P}(\texttt{ron} \mid \texttt{I ␣was ␣a ␣mat})$, the sum of two word probabilities $\mathrm{P}(\textit{mat} \mid \textit{I was a}) + \mathrm{P}(\textit{matron} \mid \textit{I was a})$ can exceed one, which would violate the probability axiom that the probability of all outcomes equals one [6]. We propose a simple fix for calculating these word probabilities instead as $\mathrm{P}(\texttt{mat ␣} \mid \texttt{I ␣was ␣a ␣})$ and $\mathrm{P}(\texttt{mat ron ␣} \mid \texttt{I ␣was ␣a ␣})$ by reaccounting the probabilities of whitespaces. This correction results in word probabilities that sum to one and are more congruent with self-paced reading and eye-tracking paradigms where human subjects directly observe upcoming word boundaries.

We subsequently evaluated the impact of this confound on two previously reported psycholinguistic modeling experiments. The first experiment re-evaluated surprisal-based estimates of garden-path effects in English transitive/intransitive sentences [7, 3] from GPT-2 LMs [10], using the data and following the procedures of Huang et al. [4]. First, linear mixed-effects (LME) models were fit to self-paced reading times of filler items, which were used to generate predicted reading times for the critical word and two spillover words over 24 items. The difference in predicted reading times was then estimated through another LME model that includes an ambiguity condition as a main predictor:

(2) a. Ambiguous condition: *After the doctor left the room turned very dark ...*
    b. Unambiguous condition: *After the doctor left, the room turned very dark ...*

Figure 1 shows that the correction lowers the estimated magnitude of garden-path effects in the first and second spillover regions. Such lower estimates suggest that the underestimation of human-like garden-path effects by LM surprisal is more severe than previously reported.

The second experiment re-evaluated the fit of surprisal from Pythia English LMs [1] to naturalistic reading times of the English Natural Stories and Dundee corpora [2, 5], following the procedures of Oh and Schuler [8]. On each dataset, LME models including LM surprisal and standard baseline predictors were fit to approximately half of the data points, whose log-likelihoods were compared against that of the baseline LME model without LM surprisal to calculate the increase due to surprisal. Figure 2 shows that the correction results in poorer fits to naturalistic reading times, especially for LMs that have seen around 256 to 1,000 batches of training data. Taken together, these results suggest that part of the processing difficulty predicted by LM surprisal was spuriously due to the LMs' implicit prediction of word boundaries.

LME Formula
Filler [4]: `RT ~ surp + surp_prev + surp_prev2 + s(length) + freq + freq_prev + freq_prev2 + s(index) + (1 | subject) + (1 | item)`
Predicted RT [4]: `pred_RT ~ condition + (1 | subject) + (1 | item)`
Natural Stories [2]: `log(RT) ~ surp + length + index + (surp + length + index + 1 | subject) + (1 | subject:sentid)`
Dundee [5]: `log(GPD) ~ surp + length + index + slength + pfix + (surp + length + index + slength + pfix + 1 | subject) + (1 | sentid)`

Table 1: LME formulae used in the experiments. `GPD`: Go-past duration, `index`: position of the word within the sentence, `slength`: saccade length, `pfix`: whether the previous word was fixated, `sentid`: index of the sentence within each corpus. On the Natural Stories and Dundee corpora, all predictors were z-transformed.

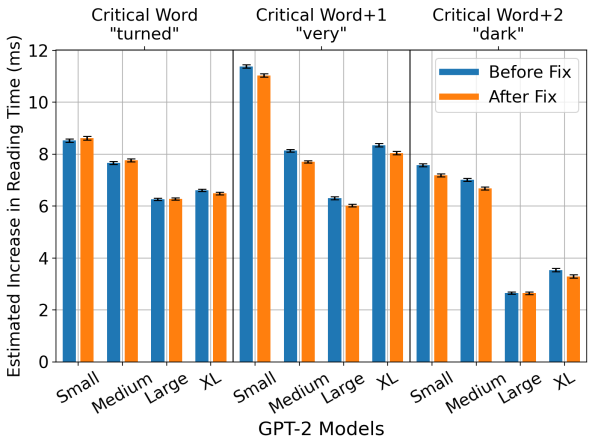

Figure 1: Estimated effects of interest at each region for the transitive/intransitive garden-path construction before and after probability correction. The difference in estimated effects of interest is significant at $p < 0.05$ level for all comparisons in the first and second spillover regions, except for GPT-2 Large in the second spillover region.

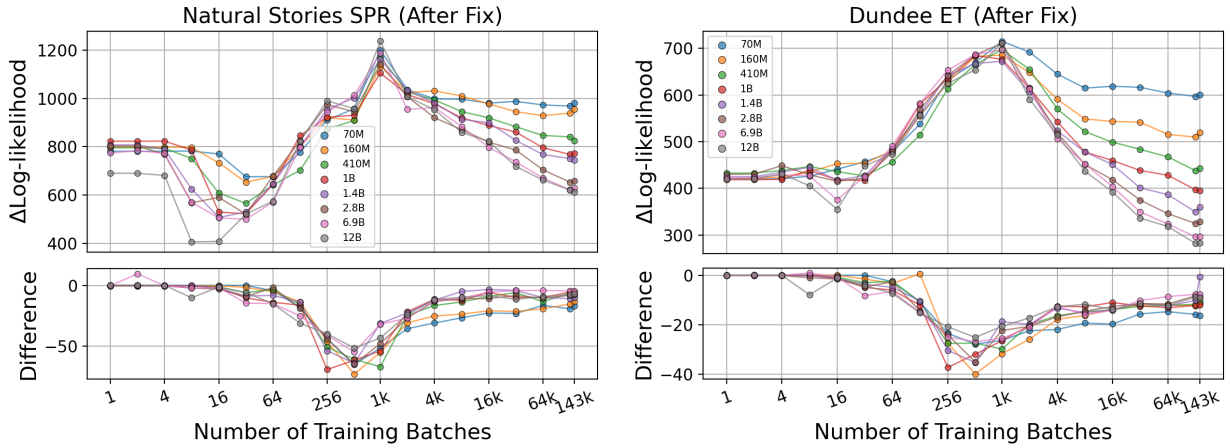

Figure 2: Increase in LME model log-likelihood due to including surprisal from Pythia LMs calculated with probability correction (top) and the resulting change in LME model log-likelihood (bottom) on the Natural Stories Corpus (left) and the Dundee Corpus (right).

[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] Kennedy, A., Hill, R., & Pynte, J. 2003. The Dundee Corpus.
[6] Kolmogorov, A. N. 1933. Foundations of the Theory of Probability (Berlin: Julius Springer)
[7] Mitchell, D. C. 1987. Lexical guidance in human parsing: Locus and processing characteristics.
[8] Oh, B.-D., & Schuler, W. 2023. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens.
[9] Oh, B.-D., & Schuler, W. 2023. Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times?.
[10] Radford, A., Wu, J., Child, R., et al. 2019. Language models are unsupervised multitask learners.
[11] Sennrich, R., Haddow, B., & Birch, A. 2016. Neural machine translation of rare words with subword units.
[12] Shain, C., Meister, C., Pimentel, T., et al. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time.