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

Byung-Doh Oh
Center for Data Science
New York University
oh.b@nyu.edu

William Schuler
Department of Linguistics
The Ohio State University
schuler.77@osu.edu

Abstract

Predictions of word-by-word conditional probabilities from Transformer-based language models are often evaluated to model the incremental processing difficulty of human readers. In this paper, we argue that there is a confound posed by the most common method of aggregating subword probabilities of such language models into word probabilities. This is due to the fact that tokens in the subword vocabulary of most language models have leading whitespaces and therefore do not naturally define stop probabilities of words. We first prove that this can result in distributions over word probabilities that sum to more than one, thereby violating the axiom that $\mathsf{P}(\Omega) = 1$. This property results in a misallocation of word-by-word surprisal, where the unacceptability of the end of the current word is incorrectly carried over to the next word. Additionally, this implicit prediction of word boundaries incorrectly models psycholinguistic experiments where human subjects directly observe upcoming word boundaries. We present a simple decoding technique to reaccount the probability of the trailing whitespace into that of the current word, which resolves this confound. Experiments show that this correction reveals lower estimates of garden-path effects in transitive/intransitive sentences and poorer fits to naturalistic reading times.

1 Introduction

Language models (LMs), which are trained to make predictions about upcoming words, are at the core of many natural language processing (NLP) applications. While most contemporary applications involve generating text by sampling from the LMs' conditional probability distribution, the magnitudes of the probabilities they assign to each word in a given sentence have been important from two perspectives. The first is from the perspective of LM interpretability, which aims to study their predictions and the linguistic knowledge encoded in their representations. A well-established paradigm in this line of research is what has been dubbed "targeted syntactic evaluation" (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018), in which probabilities of critical words in minimal pairs (e.g. grammatical vs. ungrammatical sentences) are compared.

Moreover, in cognitive modeling, conditional probabilities from LMs are used to model the word-by-word reading times of human subjects, often under the theoretical link that the contextual predictability of a word determines its processing difficulty (Hale, 2001; Levy, 2008). Recent work in this line of research has evaluated surprisal estimates (i.e. negative log probabilities) from LMs and has shown that surprisal from larger Transformer-based model variants are less predictive of naturalistic reading times (Oh and Schuler, 2023b; Shain et al., 2024; Steuer et al., 2023) and that surprisal greatly underpredicts the processing difficulty of garden-path constructions (van Schijndel and Linzen, 2021; Arehalli et al., 2022; Huang et al., 2024).

As such, while the use of word-by-word probabilities from LMs is popular in computational linguistics research, we argue that there is a confound for calculating them correctly that has gone unaddressed. This confound is posed by subword tokenization schemes (e.g. byte-pair encoding; Sennrich et al., 2016) that are used to define the token-level vocabulary for training most contemporary LMs (e.g. AI@Meta, 2024; Google Gemini Team, 2024; Jiang et al., 2023). For languages that use whitespace orthography, these subword tokenization schemes often build the whitespace character directly into the front of the tokens, thereby resulting in *leading* whitespaces. As a consequence, the stop probability of a word (i.e. the probability of the *trailing* whitespace) is never explicitly calculated, and therefore the sum over the probabilities of all possible whitespace words can exceed one.

We propose a simple and efficient decoding

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 3464–3472
November 12-16, 2024 ©2024 Association for Computational Linguistics

method that reaccounts the probability of the trailing whitespace into that of the current word, which resolves this confound. Regression results show that this correction reveals significantly lower surprisal-based estimates of garden-path effects in transitive/intransitive sentences and poorer fits of LM surprisal to naturalistic reading times.

2 Confound From Leading Whitespaces and Whitespace-Trailing Decoding

This section provides a proof that the leading whitespaces of the LMs' subword vocabulary result in inconsistent word probabilities, describes a related confound, and proposes a simple decoding method for addressing it.

2.1 Proof of Inconsistent Word Probabilities

On languages that use whitespace orthography, the vocabulary V defined by the subword tokenization scheme consists of the set of tokens that begin with a whitespace V_B, and the set of tokens that do not begin with a whitespace V_I.[1] In the context of next-word prediction, the sample space of a whitespace-delimited word is $\Omega = \{x_{1..n} \mid x_1 \in V_B,\ x_{2..n} \in V_I,\ n \in \mathbb{N}\}$, where n is the total number of subword tokens in each whitespace word as determined by the tokenizer.

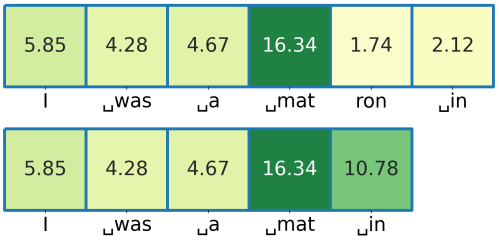
Theorem 1 Leading whitespaces of the LMs' subword vocabulary can result in word probabilities that violate the Kolmogorov (1933) axiom that $\mathsf{P}(\Omega) = 1$.

Existence Proof Let $\mathsf{P}(x_1{=}j_1) = 1$ and $\mathsf{P}(x_2{=}j_2 \mid x_1{=}j_1) = 1$, where $j_1 \in V_B$ and $j_2 \in V_I$. It follows from the chain rule of conditional probabilities that:

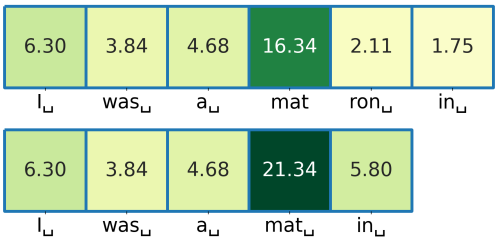
$$\begin{aligned}\mathsf{P}(x_1{=}j_1, x_2{=}j_2) &= \mathsf{P}(x_1{=}j_1) \cdot \mathsf{P}(x_2{=}j_2 \mid x_1{=}j_1) \\ &= 1 \cdot 1 = 1. \qquad (1)\end{aligned}$$

If word probabilities are simply defined as the product of the probabilities of the tokens within those words, then $\mathsf{P}(x_1{=}j_1) + \mathsf{P}(x_1{=}j_1, x_2{=}j_2) > 1$, $\mathsf{P}(\Omega) > 1$, and therefore the axiom is violated. ■

For example, given the minimal pair *I was a matron in France* and *I was a mat in France*, where *matron* is more likely than *mat*, the LM tokenizes the two sentences as follows and calculates the conditional probability of each token.[2]



(a) Surprisal values calculated with leading whitespaces.



(b) Surprisal values calculated with trailing whitespaces.

Figure 1: Surprisal values calculated for the partial sentences *I was a matron in* and *I was a mat in* using the GPT-2 XL LM (Radford et al., 2019), with leading whitespaces (top; standard practice) and trailing whitespaces (bottom; proposed in this work).

I ␣was ␣a ␣mat ron ␣in ␣France (2)

I ␣was ␣a ␣mat ␣in ␣France (3)

The presence of leading whitespaces results in an incorrect allocation of word-by-word surprisal. As can be seen in Example 2, due to this tokenization, $\mathsf{P}(\text{␣}mat\ ron \mid I\ \text{␣}was\ \text{␣}a)$ is factorized into $\mathsf{P}(\text{␣}mat \mid I\ \text{␣}was\ \text{␣}a) \cdot \mathsf{P}(ron \mid I\ \text{␣}was\ \text{␣}a\ \text{␣}mat)$, and therefore it follows that $\mathsf{P}(\text{␣}mat\ ron \mid I\ \text{␣}was\ \text{␣}a) \leq \mathsf{P}(\text{␣}mat \mid I\ \text{␣}was\ \text{␣}a)$, despite the fact that *matron* is more acceptable than *mat* in the above context. Instead, part of the 'unacceptability' of *mat* is incorrectly carried over to $\mathsf{P}(\text{␣}in \mid I\ \text{␣}was\ \text{␣}a\ \text{␣}mat)$, where *␣in* competes for probability mass against the highly likely *ron* (Figure 1a).

2.2 Confound: Incompatibility With Psycholinguistic Experimental Paradigms

Additionally, the presence of leading whitespaces in subword tokens makes the LM's predictions incompatible with the self-paced reading paradigm, in which human readers directly observe the upcoming word boundary.

The cat sat on the ***mat*** *...* (4)

[1] $V = V_B \cup V_I$, and $V_B \cap V_I = \varnothing$. The subscripts respectively represent the 'beginning' and 'inside' of a whitespace-delimited word.

[2] In the context of the LM's tokens, '␣' is used to denote the explicit whitespace character that is part of the token, and whitespace is used to delimit subword tokens.

In Example 4, when human readers see *mat*, they know that the next keystroke will reveal a new whitespace-delimited word (analogous to observing that the next token will be in V_B) and not transform it into e.g. *matron* (analogous to observing that the next token will be in V_I). In contrast, LMs define a probability distribution over both V_B and V_I after the token *mat* in the sequence *The cat sat on the mat*.

While this confound is more apparent in the self-paced reading paradigm, this is also a potential confound for studying data collected through the typical eye-tracking paradigm. This is because native speakers of languages with whitespace orthographies have been shown to be sensitive to the location of upcoming whitespaces through parafoveal processing and utilize this information to plan eye movements (Pollatsek and Rayner, 1982; Rayner et al., 1998; Perea and Acha, 2009). Therefore, although information about word boundaries is not directly built into the design of the paradigm, it can be argued that human subjects engaged in the eye-tracking paradigm also face little uncertainty about upcoming word boundaries.

2.3 Proposed Solution: Whitespace-Trailing Decoding

This inconsistency and confound can be resolved by reaccounting the probability of the *trailing* whitespace as part of the word's probability, in lieu of that of the *leading* whitespace as LMs currently do (Examples 2 and 3). To this end, we propose whitespace-trailing (WT) decoding. Given a word w_{t+1} that consists of subword tokens $x_{n_t+1..n_{t+1}}$, where n_t is the total number of subword tokens in the word sequence $w_{1..t}$, and $x_{n_t+1} \in V_B$, and $x_{n_t+2..n_{t+1}} \in V_I$, WT decoding reallocates the probability of the leading whitespace of each word to its previous word:[3]

$$\mathsf{P}(w'_{t+1} \mid w'_{1..t}) = \mathsf{P}(w_{t+1} \mid w_{1..t}) \cdot \frac{\mathsf{P}(x_{n_{t+1}+1} \in V_B \mid w_{1..t+1})}{\mathsf{P}(x_{n_t+1} \in V_B \mid w_{1..t})}. \quad (5)$$

[3]See Appendix A for the proof that WT decoding results in consistent word probabilities. However, we note that WT decoding does not resolve other issues with subword units that may be addressed by re-training LMs with different tokenization schemes (e.g. Nair and Resnik, 2023), which can nonetheless be expensive. Concurrent work by Pimentel and Meister (2024) points out this same issue and also proposes WT decoding.

For instance, applying Equation 5 to Example 3 yields:

$$\mathsf{P}(mat\text{␣} \mid I\text{␣}\, was\text{␣}\, a\text{␣}) = \mathsf{P}(\text{␣}mat \mid I\,\text{␣}was\,\text{␣}a) \cdot \frac{\mathsf{P}(\text{␣} \mid I\,\text{␣}was\,\text{␣}a\,\text{␣}mat)}{\mathsf{P}(\text{␣} \mid I\,\text{␣}was\,\text{␣}a)}. \quad (6)$$

As WT decoding simply involves the factorization of whitespace probabilities by marginalizing over tokens in V_B and rearranging them, it requires no modifications to the LM and minimal overhead. Additionally, the joint probability of the entire sequence, and therefore metrics like perplexity, changes minimally by a factor of the probability of the final trailing whitespace with WT decoding.

As can be seen in Figure 1b, incorporating the probabilities of trailing whitespaces correctly differentiates between *matron* and *mat* in this context, and removes the inherent relationship between the two probabilities that holds with leading whitespaces. Additionally, the 'unacceptability' of *mat* that was incorrectly carried over to ␣*in* in Example 3 is now reflected in $\mathsf{P}(mat\text{␣} \mid I\text{␣}\, was\text{␣}\, a\text{␣})$.

LM probabilities with trailing whitespaces are also better aligned with the self-paced reading paradigm where the upcoming word boundaries are directly observed. For example, the calculation of $\mathsf{P}(mat\text{␣} \mid The\text{␣}\, cat\text{␣}\, sat\text{␣}\, on\text{␣}\, the\text{␣})$ precludes the prediction of tokens in V_I directly after *mat*, which correctly reflects the fact that the next keystroke in Example 4 will reveal a new whitespace word.

3 Experiment 1: Surprisal-Based Estimates of Garden-Path Effects

Equation 6 shows that WT decoding will result in an increase (or decrease) in probability to the extent that the next token is likely to be in V_B proportional to the extent that the first token of the current word was likely to be in V_B. The first experiment demonstrates that the confound posed by leading whitespaces affects surprisal-based estimates of garden-path effects in transitive/intransitive sentences (Mitchell, 1987; Gorrell, 1991), which is caused by syntactic disambiguation that takes place at the critical word (highlighted in magenta).

After the doctor left the room turned very dark ... (7)

The same critical word in the control counterpart is thought to be easier to process, as the verb *left* is disambiguated by the comma.

After the doctor left, the room turned very dark ... (8)

3.1 Procedures

We estimated surprisal-based garden-path effects from GPT-2 model variants (Radford et al., 2019) with and without WT decoding, using the data and following the procedures of Huang et al. (2024). First, to estimate a linking function between LM surprisal and human reading times, linear mixed-effects regression (LMER) models with the following formula were fit to self-paced reading times ($n = 995,814$) of filler items from the Provo Corpus (Luke and Christianson, 2018) for each variant:

```
RT ~ surp + surp_prev1 + surp_prev2 + s(length) +
     freq + freq_prev1 + freq_prev2 + s(index) +
     (1 | subject) + (1 | item),
```

where `length` is word length in characters, `index` is the position of the word within the sentence, and the `frequency` predictors were log-transformed.

These modeling choices assume a linear relationship between surprisal and reading times (Shain et al., 2024; Wilcox et al., 2023), and that surprisal and log frequency from two previous words have a lingering influence on the current word (spillover effects). These LMER models were subsequently used to predict word-by-word reading times (in ms) for 24 items in the ambiguous condition (Example 7) and the unambiguous control condition (Example 8) of the transitive/intransitive construction, which were read by 2,000 subjects (n=15,915).

The increase in the predicted reading times of the disambiguating critical word and two subsequent words due to the increase in surprisal across conditions was estimated using LMER models with the following formula to quantify the magnitude of surprisal-based garden-path effects at each word:[4]

```
pred_RT ~ condition + (1 | subject) + (1 | item).
```

3.2 Results

The results in Figure 2 show addressing the confound posed by subword tokenization through WT decoding lowers the estimated magnitude of garden-path effects in the first and second spillover regions, the difference in which is significant at $p < 0.05$ level for all comparisons except GPT-2 Large in the second spillover region. This is due to the decrease in surprisal at the critical region of the ambiguous condition (*turned*), as the probability of its unlikely preceding whitespace[5] is reaccounted by the previous word (*room*). The resulting decrease in surprisal difference across conditions at the critical region is carried over to the two spillover regions. At the critical region itself, however, this decrease is not observed as the increase in surprisal difference of its previous word (*room*) cancels it out. Such lower estimates suggest that the underestimation of human-like garden-path effects is more severe than previously reported.

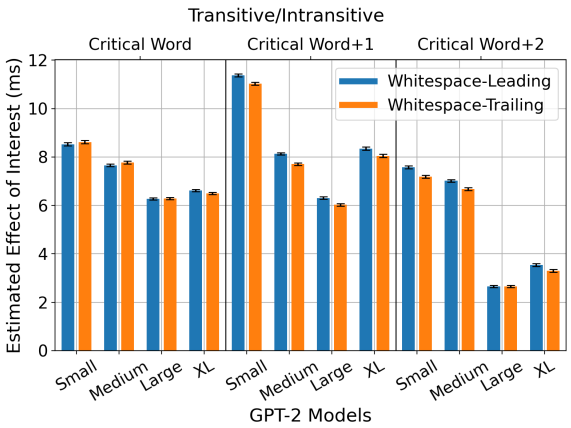


Figure 2: Estimated effects of interest at each region for the transitive/intransitive garden-path construction, using GPT-2 surprisal with and without WT decoding. Error bars represent 95% confidence intervals.

4 Experiment 2: Fit of Surprisal to Naturalistic Reading Times

The second experiment evaluates how addressing the confound posed by leading whitespaces affects the fit of LM surprisal to naturalistic reading times.

4.1 Procedures

The experimental procedures closely follow those of Oh and Schuler (2023a), who evaluated surprisal estimates from Pythia LM variants (Biderman et al., 2023) with different model sizes and training data amounts on self-paced reading (SPR) times from the Natural Stories Corpus (Futrell et al., 2021) and go-past durations (GPD) from the Dundee Corpus (Kennedy et al., 2003).[6] LMER models with the following formulae were respectively fit to the Natural Stories and Dundee corpora, whose likelihoods were then subtracted from those of the baseline LMER models without surprisal to calculate the increase in log-likelihood due to surprisal, or ΔLL:

```
log(SPR) ~ surp + length + index +
           (surp + length + index + 1 | subject) +
           (1 | subject:sentid)
```

[4]Both the 'filler item' and 'reading time increase' LMER models have been simplified from the specifications in Huang et al. (2024) due to convergence issues.

[5]The LMs strongly expect a comma right after *room*.

[6]See Appendix B for the data preprocessing procedures.

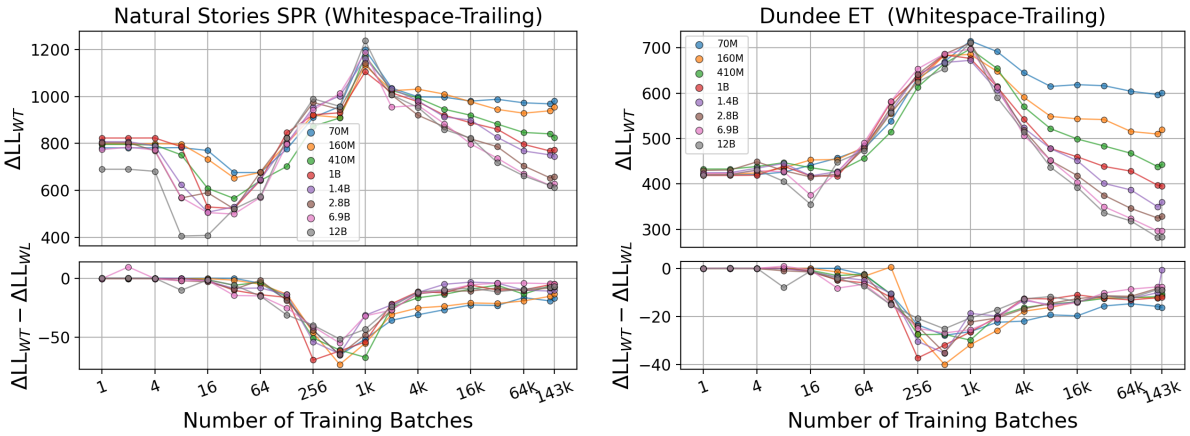


Figure 3: Increase in regression model log-likelihood due to including surprisal estimates from Pythia LM variants calculated with WT decoding (top) and the resulting change in regression model log-likelihood (bottom). See Appendix C for results from surprisal estimates calculated without WT decoding.

```
log(GPD) ~ surp + length + index + slength + pfix +
           (surp + length + index + slength + pfix +
            1 | subject) + (1 | sentid),
```

where `slength` is the saccade length, `pfix` is whether the previous word was fixated, and `sentid` is the index of the sentence within each corpus. These procedures were repeated with and without WT decoding to calculate ΔLL_{WT} and ΔLL_{WL} respectively, and the change in fit to reading times as a result of addressing the confound was calculated.

4.2 Results

Figure 3 shows that surprisal estimates calculated with WT decoding results in poorer fits to naturalistic reading times, especially for LMs that have seen around 256 to 1,000 batches of training data on both corpora. Nonetheless, the peak in ΔLL at around 1,000 training batches[7] and the adverse effect of model size at the end of LM training (Oh and Schuler, 2023a) are replicated. In contrast to these results, Pimentel and Meister (2024) report small improvements on the same two corpora as a result of applying WT decoding to fully trained Pythia LMs. We conjecture this is due to different regression modeling procedures involving different baseline predictors.

5 Conclusion

This work calls attention to an inconsistency and a confound that is inherent in word probabilities calculated from LMs trained with subword tokenization. These are posed by the fact that tokens have leading whitespaces in most models, meaning that the stop probability of a whitespace word is never explicitly calculated, which can result in word probability distributions whose sum exceeds one. We proposed WT decoding as a solution for these issues, and demonstrated that addressing them reveals lower surprisal-based estimates of transitive/intransitive garden-path effects and poorer fits of LM surprisal to naturalistic reading times. Other targeted syntactic constructions and naturalistic reading time corpora may similarly show systematic changes to word probabilities.

More generally, addressing these issues will have the biggest impact on probabilities of words neighboring low-probability whitespaces, such as those at potential phrasal/clausal boundaries where LMs will likely predict a punctuation mark. These issues will also be more pronounced for LMs that are not able to predict word boundaries accurately, such as those trained on smaller amounts of data. Therefore, future studies using LM word probabilities for interpretability and cognitive modeling research should control for them through WT decoding.[8]

Acknowledgments

We thank the ARR reviewers and the area chair for their helpful comments. This work was supported by the National Science Foundation (NSF) grant #1816891. All views expressed are those of the authors and do not necessarily reflect the views of the NSF. Computations for this work were partly run using the Ohio Supercomputer Center (1987).

[7] The change in log-likelihood ($\Delta\text{LL}_{WT} - \Delta\text{LL}_{WL}$) at 1,000 training batches is significant at $p < 0.001$ level on both corpora by a permutation test of aggregated squared errors.

[8] Code for implementing WT decoding is available at: https://github.com/byungdoh/wt_decoding.

Limitations

The confound in the connection between word-by-word conditional probabilities of Transformer-based language models and human reading times identified in this work is supported by experiments using language model variants trained on English text and data from human subjects that are native speakers of English. Therefore, the confound identified in this work may not generalize to other languages, in particular those that do not use whitespace orthography. Additionally, this work is concerned with the use of language models as cognitive models of human sentence processing, and therefore does not relate to their use in natural language processing applications, such as text generation, summarization, or question answering.

Ethics Statement

This work used data collected as part of previously published research (Huang et al., 2024; Luke and Christianson, 2018; Futrell et al., 2021; Kennedy et al., 2003). Readers are referred to the respective publications for more information on the data collection and validation procedures. As this work focuses on studying the connection between conditional probabilities of language models and human sentence processing, its potential negative impacts on society appear to be minimal.

References

AI@Meta. 2024. Llama 3 model card. *Accessed from GitHub*.

Suhas Arehalli, Brian Dillon, and Tal Linzen. 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*, pages 301–313.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 2397–2430.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2021. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55:63–77.

Google Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *arXiv preprint*, arXiv:2312.11805v3.

Paul Gorrell. 1991. Subcategorization and sentence processing. In R. C. Berwick, S. P. Abney, and C. Tenny, editors, *Principle-Based Parsing: Computation and Psycholinguistics*, pages 279–300. Springer, Dordrecht.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1195–1205.

John Hale. 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 on Language Technologies*, pages 1–8.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint*, arXiv:2310.06825.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The Dundee Corpus. In *Proceedings of the 12th European Conference on Eye Movement*.

Andrey Nikolaevich Kolmogorov. 1933. *Foundations of the Theory of Probability*. Julius Springer, Berlin.

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

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

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

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Don C. Mitchell. 1987. Lexical guidance in human parsing: Locus and processing characteristics. In M. Coltheart, editor, *Attention and Performance XII: The Psychology of Reading*, pages 601–618. Erlbaum, Hillsdale, NJ.

Sathvik Nair and Philip Resnik. 2023. Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11251–11260.

Byung-Doh Oh and William Schuler. 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*, pages 1915–1921.

Byung-Doh Oh and William Schuler. 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.

Ohio Supercomputer Center. 1987. Ohio Supercomputer Center.

Manuel Perea and Joana Acha. 2009. Space information is important for reading. *Vision Research*, 49(15):1994–2000.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. *arXiv preprint*, arXiv:2406.14561.

Alexander Pollatsek and Keith Rayner. 1982. Eye movement control in reading: The role of word boundaries. *Journal of Experimental Psychology: Human Perception and Performance*, 8(6):817–833.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Keith Rayner, Martin H. Fischer, and Alexander Pollatsek. 1998. Unspaced text interferes with both word identification and eye movement control. *Vision Research*, 38(8):1129–1144.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Julius Steuer, Marius Mosbach, and Dietrich Klakow. 2023. Large GPT-like models are bad babies: A closer look at the relationship between linguistic competence and psycholinguistic measures. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 142–157.

Marten van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988.

Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

A Proof of Consistent Word Probabilities With Whitespace-Trailing Decoding

Theorem 2 Applying whitespace-trailing decoding results in word probabilities that satisfy the Kolmogorov (1933) axiom that $\mathsf{P}(\Omega) = 1$.

Proof In the context of predicting w_{t+1} given $w_{1..t}$, the sample space is $\Omega = \{x_{n_t+1..n_{t+1}} \mid x_{n_t+1} \in V_B,\ x_{n_t+2..n_{t+1}} \in V_I,\ \{n_t, n_{t+1}\} \subset \mathbb{N},\ n_{t+1} > n_t\}$, where n_t is the total number of subword tokens in the word sequence $w_{1..t}$, and n_{t+1} is the total number of subword tokens in the word sequence $w_{1..t+1}$. Therefore, $\mathsf{P}(\Omega)$ is the total sum of word probabilities when $n_{t+1} - n_t = 1, 2, 3, \ldots$.

The sum of word probabilities according to Equation 5 when $n_{t+1} - n_t = 1$ is:

$$\sum_{j_1 \in V_B} \mathsf{P}(x_{n_t+1} = j_1 \mid w_{1..t}) \cdot \frac{\mathsf{P}(x_{n_t+2} \in V_B \mid x_{n_t+1} = j_1, w_{1..t})}{\mathsf{P}(x_{n_t+1} \in V_B \mid w_{1..t})} = \frac{\mathsf{P}(x_{n_t+1} \in V_B, x_{n_t+2} \in V_B \mid w_{1..t})}{\mathsf{P}(x_{n_t+1} \in V_B \mid w_{1..t})}$$
$$= \mathsf{P}(x_{n_t+2} \in V_B \mid x_{n_t+1} \in V_B, w_{1..t}). \quad (9)$$

More generally, the sum of word probabilities when $n_{t+1} - n_t \geq 2$ is:

$$\sum_{\substack{j_1 \in V_B \\ j_{2..(n_{t+1}-n_t)} \in V_I}} \mathsf{P}(x_{n_t+1..n_{t+1}} = j_{1..(n_{t+1}-n_t)} \mid w_{1..t}) \cdot \frac{\mathsf{P}(x_{n_{t+1}+1} \in V_B \mid x_{n_t+1..n_{t+1}} = j_{1..(n_{t+1}-n_t)}, w_{1..t})}{\mathsf{P}(x_{n_t+1} \in V_B \mid w_{1..t})}$$
$$= \mathsf{P}(x_{n_t+2..n_{t+1}} \in V_I, x_{n_{t+1}+1} \in V_B \mid x_{n_t+1} \in V_B, w_{1..t}). \quad (10)$$

$\mathsf{P}(\Omega)$ can then be calculated as the following series that sums over disjoint subspaces of Ω:

$$\begin{aligned}\mathsf{P}(\Omega) = {} & \mathsf{P}(x_{n_t+2} \in V_B \mid x_{n_t+1} \in V_B, w_{1..t}) + \\ & \mathsf{P}(x_{n_t+2} \in V_I, x_{n_t+3} \in V_B \mid x_{n_t+1} \in V_B, w_{1..t}) + \\ & \mathsf{P}(x_{n_t+2} \in V_I, x_{n_t+3} \in V_I, x_{n_t+4} \in V_B \mid x_{n_t+1} \in V_B, w_{1..t}) + \\ & \mathsf{P}(x_{n_t+2} \in V_I, x_{n_t+3} \in V_I, x_{n_t+4} \in V_I, x_{n_t+5} \in V_B \mid x_{n_t+1} \in V_B, w_{1..t}) + \\ & \ldots ,\end{aligned} \quad (11)$$

which approaches $\mathsf{P}(x_{n_t+2} \in V_B \mid x_{n_t+1} \in V_B, w_{1..t}) + \mathsf{P}(x_{n_t+2} \in V_I \mid x_{n_t+1} \in V_B, w_{1..t}) = 1$ in the limit. ■

B Preprocessing Procedures for Naturalistic Reading Time Corpora

The Natural Stories Corpus (Futrell et al., 2021) provides self-paced reading times from 181 subjects that read 10 English stories (10,256 words), which were filtered to exclude those shorter than 100 ms or longer than 3000 ms, those of sentence-initial and -final words, and those from subjects who answered fewer than four comprehension questions correctly. Approximately 50% of the observations (384,905 observations) selected based on the sum of the subject index and the sentence index was used to fit the LMER models and calculate ΔLL.

The Dundee Corpus (Kennedy et al., 2003) provides fixation durations from 10 subjects that read 67 English newspaper editorials (51,501 words), which were filtered to exclude those from unfixated words, those of words following saccades longer than four words, and those of sentence/document/line/screen-initial and -final words. Again, approximately 50% of the observations (98,115 observations) selected based on the sum of the subject index and the sentence index was used to fit the LMER models and calculate ΔLL.

C Increase in Regression Model Log-Likelihood Without WT Decoding

The increase in regression model log-likelihood due to including surprisal estimates from Pythia LM variants calculated without WT decoding can be found in Figure 4.

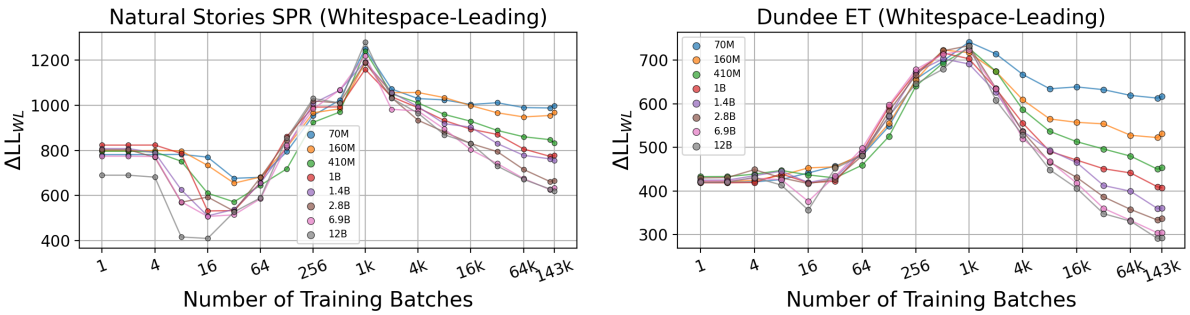


Figure 4: Increase in regression model log-likelihood due to including surprisal estimates from Pythia LM variants calculated without WT decoding.