A Better Way to Do Masked Language Model Scoring

Carina Kauf
Massachusetts Institute of Technology
ckauf@mit.edu

Anna A. Ivanova
Massachusetts Institute of Technology
annaiv@mit.edu

Abstract

Estimating the log-likelihood of a given sentence under an autoregressive language model is straightforward: one can simply apply the chain rule and sum the log-likelihood values for each successive token. However, for masked language models (MLMs), there is no direct way to estimate the log-likelihood of a sentence. To address this issue, Salazar et al. (2020) propose to estimate sentence pseudo-log-likelihood (PLL) scores, computed by successively masking each sentence token, retrieving its score using the rest of the sentence as context, and summing the resulting values. Here, we demonstrate that the original PLL method yields inflated scores for out-of-vocabulary words and propose an adapted metric, in which we mask not only the target token, but also all within-word tokens to the right of the target. We show that our adapted metric (`PLL-word-l2r`) outperforms both the original PLL metric and a PLL metric in which all within-word tokens are masked. In particular, it better satisfies theoretical desiderata and better correlates with scores from autoregressive models. Finally, we show that the choice of metric affects even tightly controlled, minimal pair evaluation benchmarks (such as BLiMP), underscoring the importance of selecting an appropriate scoring metric for evaluating MLM properties.[1]

1 Introduction

Most state-of-the-art transformer-based large language models (LLMs) fall into two classes: unidirectional (or autoregressive) models, where each token is generated based on its left context (e.g., GPT models; Radford et al., 2019), and bidirectional models, where a token is predicted from both left and right context tokens, some of which may be masked (e.g., BERT; Devlin et al., 2018). Often, it is beneficial to compare these models' performance on controlled sentence generation benchmarks. Whereas unidirectional architectures offer a natural way of calculating sentence log-likelihood (summing the log-likelihood scores of each sentence token given its left context), there is no direct way of estimating sentence log-likelihood for a bidirectional model.

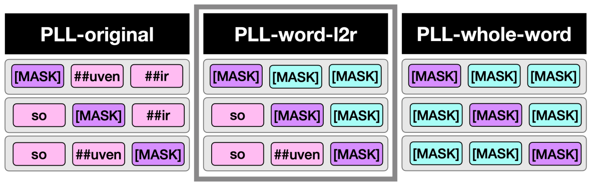

Figure 1: Three different ways to compute the PLL score of a multi-token word (e.g., `souvenir`) during masked language modeling. *Purple*: target token, *pink*: within-word tokens that are available during inference, *turquoise*: within-word tokens that are masked during inference. Sentence tokens that do not belong to the current word are always available during inference.

So far, the best available method to score a sentence under a bidirectional LLM has been the pseudo-log-likelihood (PLL) scoring approach described by Salazar et al. (2020) (and initially used by Shin et al., 2019; Wang and Cho, 2019). The PLL of a sentence is calculated as the sum of PLL scores for each token given all other sentence tokens, thus providing a comparable metric to unidirectional models' log-likelihood (LL) sentence scoring. The PLL metric is extremely popular; it is used extensively in LLM studies tackling topics as diverse as effects of training data (Sinha et al., 2021; Zhang et al., 2021), model fluency (Laban et al., 2021), syntactic and conceptual knowledge (Sinclair et al., 2022; Bhatia and Richie, 2022), social biases (Nangia et al., 2020), and others. Some of these studies have already accrued dozens of citations.

Here, we show that the metric proposed by Salazar et al. (`PLL-original`) has important shortcomings that limit its utility. Specifically, `PLL-original` overestimates the PLL of out-of-vocabulary (OOV) words, which LLM tokenizers split into multiple tokens. As a result, `PLL-original` scores fail on several theoretically

[1] Our results and code are available at https://github.com/carina-kauf/better-mlm-scoring.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 925–935
July 9-14, 2023 ©2023 Association for Computational Linguistics

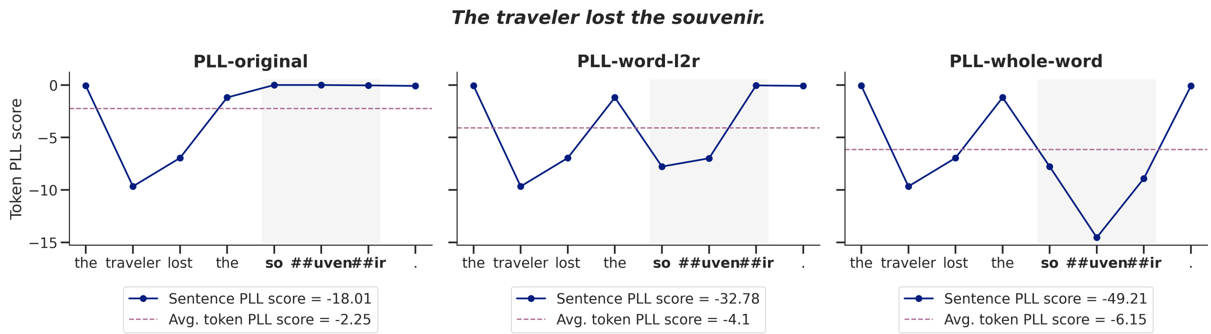

Figure 2: The `PLL-original` metric inflates scores of multi-token words, such as *souvenir*; the adjusted metrics, `PLL-word-l2r` and `PLL-whole-word`, mitigate this issue. Example generated using the `bert-base-cased` model.

desired property tests: a robust inverse relationship between sentence length and sentence PLL (Section 4.1), a robust positive correlation between a word's frequency and its PLL score (4.2), and a positive correlation between unidirectional and bidirectional model scores for the same sentences (Section 5). To remedy these issues, we propose an adjusted PLL metric, `PLL-word-l2r` (l2r: left-to-right), which estimates token PLL when future within-word tokens are also masked (Figure 1). We show that the `PLL-word-l2r` metric outperforms both `PLL-original` and alternative PLL-based metrics. We therefore recommend to use the `PLL-word-l2r` metric when estimating sentence PLL under a bidirectional LLM.

2 Motivation: score inflation for multi-token words

The `PLL-original` metric grossly overestimates the probability of OOV lexical items, such as *souvenir* (Figure 2). This is because OOV words are tokenized into subword tokens (e.g., *so ##uven ##ir*), and each subword token is predicted using the token's bidirectional context, which crucially includes the remaining tokens that make up the OOV word. Thus, even though the OOV word itself may be surprising given the sentence context, the individual parts of the OOV word are not surprising to a bidirectional model given a sentence context that includes all other subtokens of that word (e.g., it is easy to predict *so* given *##uven ##ir*; see Appendix A for additional examples).

To mitigate this bias, we adjust the PLL sentence scoring algorithm such that the model cannot access future within-word tokens (`PLL-word-l2r`) or any within-word tokens (`PLL-whole-word`) when predicting the target.

Below, we conduct a rigorous investigation of our modified metrics to determine whether this intuitive benefit holds quantitatively.

3 Methods

For our analysis, we adapt the `scorer` module of the `minicons` library (Misra, 2022), an open-source wrapper library around HuggingFace `transformers` (Wolf et al., 2020) that enables efficient extraction of word- and sentence-level probabilities from LLMs. The MLM scoring procedure of the `minicons` library follows the procedure originally proposed by Salazar et al. (2020). For details on sentence preprocessing, see Appendix B.

3.1 PLL metrics

`PLL-original`. In this metric, each sentence token s_t of a sentence S with n tokens is consecutively replaced with a `[MASK]` and is predicted using all past and future tokens, irrespective of whether the context tokens belong to the same or a different word than the target token. Thus, inference is conditioned on the context $S_{\setminus t} := (s_1, \ldots, s_{t-1}, s_{t+1}, \ldots, s_n)$. The final sentence score is obtained as the sum of the log probabilities of each sentence token given its context:

$$\mathrm{PLL}_{\mathrm{orig}}(S) := \sum_{t=1}^{n} \log P_{\mathrm{MLM}}(s_t \mid S_{\setminus t}) \qquad (1)$$

`PLL-word-l2r`. In this metric, a `[MASK]` is placed not only over the current target token (now: s_{w_t}), but also over all future sentence tokens that belong to the same word s_w as the target. Inference is then conditioned on a context that includes all preceding sentence tokens (including those belonging to the current word) and all sentence tokens from future words. The final score of a sentence S is obtained as the sum of the log probabilities of each of the $|w|$ tokens in each of the $|S|$ words:

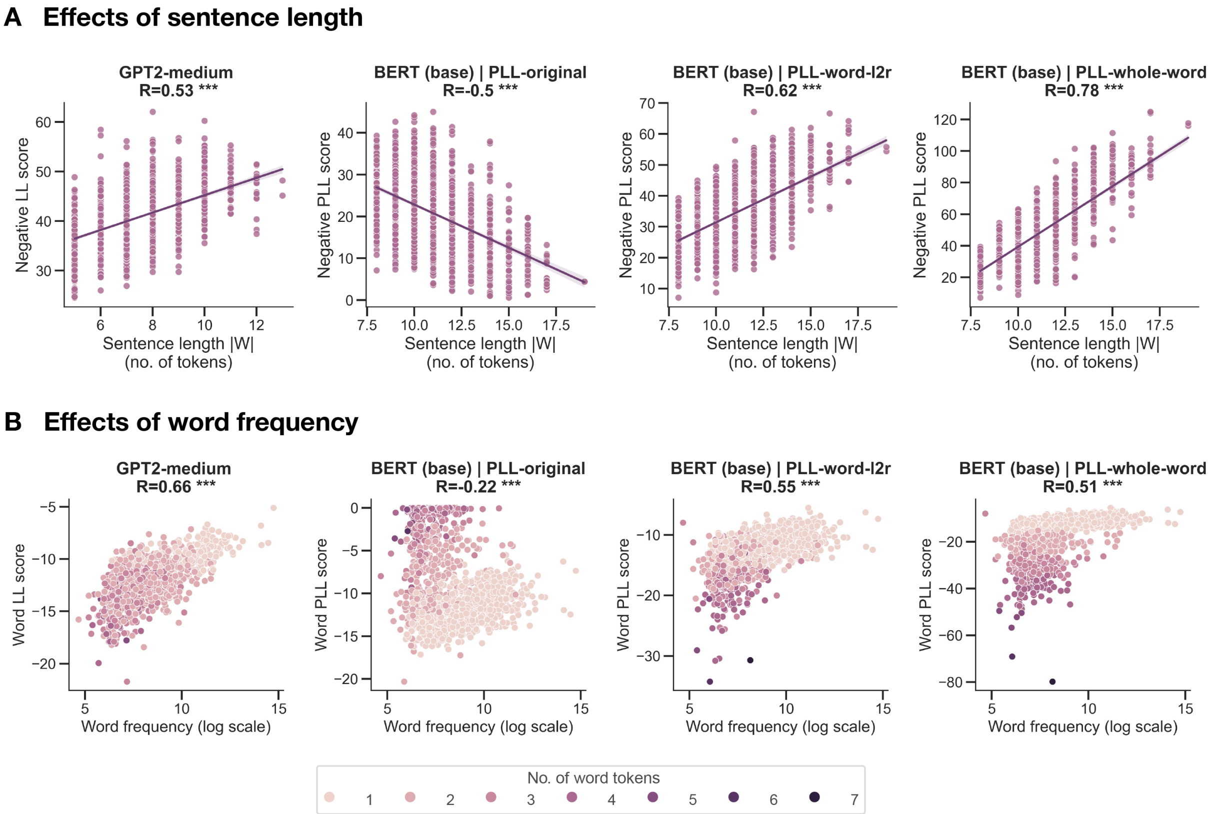

Figure 3: Out of all PLL metrics, `PLL-word-l2r` best satisfies theoretical desiderata: **(A)** an inverse relationship between negative sentence PLL (a measure of model surprisal) and sentence length and **(B)** a positive correlation between word PLL and word log frequency. In (A), each dot is a sentence; in (B), each dot is a unique word from the dataset. Here and elsewhere, reported correlations are Pearson correlations.

$$\mathrm{PLL}_{\mathrm{l2r}}(S) := \sum_{w=1}^{|S|} \sum_{t=1}^{|w|} \log P_{\mathrm{MLM}}(s_{w_t} \mid S_{\backslash s_{w_{t'\geq t}}}) \tag{2}$$

`PLL-whole-word`. This metric is similar to `PLL-word-l2r` and differs from it only in that a `[MASK]` is placed over *all* sentence tokens that belong to the same word s_w as the target (both preceding and future). Inference is then conditioned on a context that includes all sentence tokens except those belonging to the current word. The final score of a sentence S is obtained as the sum of the log probabilities of each of the $|w|$ tokens in each of the $|S|$ words in S given the token's context:

$$\mathrm{PLL}_{\mathrm{ww}}(S) := \sum_{w=1}^{|S|} \sum_{t=1}^{|w|} \log P_{\mathrm{MLM}}(s_{w_t} \mid S_{\backslash s_w}) \tag{3}$$

In Appendix G, we also report results for a PLL metric where not only future within-word tokens, but *all* sentence tokens to the right of the target context are masked (`PLL-sentence-l2r`). Although this method is most similar to autoregressive LL scoring, `sentence-l2r` masking for BERT is known to produce poor quality generations (Wang and Cho, 2019); we therefore refrain from including this metric in the main text.

3.2 Models

We report results for `bert-base-cased` (and `gpt2-medium` for comparison) unless stated otherwise. Results for larger models are provided in Appendices D-F.

3.3 Datasets

For our main analyses, we use the EventsAdapt dataset (Kauf et al., 2022, based on Fedorenko et al., 2020). It contains a curated set of 782 syntactically simple sentence pairs that describe plausible or implausible agent-patient interactions in active or passive voice (e.g., *The traveler lost the souvenir*). Sentences in this dataset are 5-7 words long (mean: 6.1, std: 1.05), with an average word log frequency of 10.95. We use this dataset because it

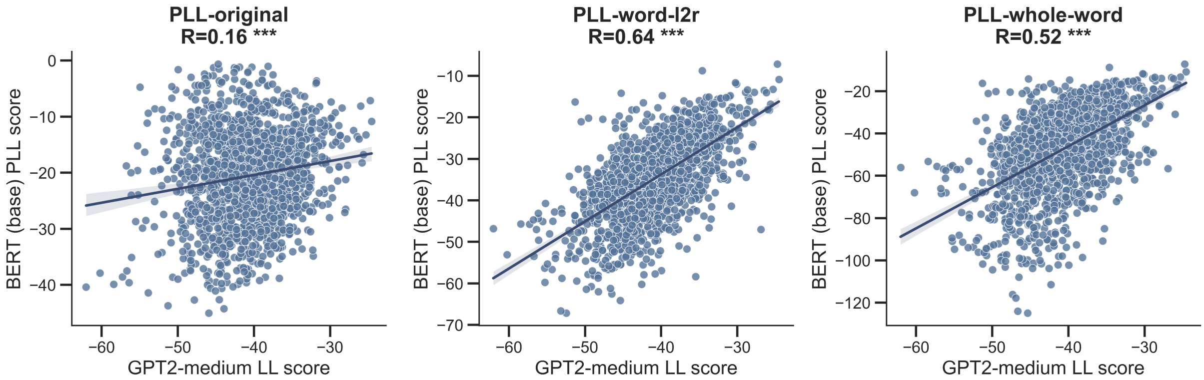

Figure 4: Correlation between bidirectional model PLL scores and unidirectional model LL scores. Each dot is a sentence.

contains a high number of OOV words (19.6% for BERT and 40.3% for GPT-2; see also Appendix C). In Appendices D-F, we show that our results generalize to two larger and more diverse corpora: the Brown corpus (Francis and Kucera, 1979) and the reference sentence set from the LibriSpeech corpus (Panayotov et al., 2015). We also apply our PLL metrics to score the sentences in the Benchmark of Linguistic Minimal Pairs (BLiMP) (Warstadt et al., 2020), a challenge set of 67k sentence pairs which target specific aspects of linguistic knowledge.

4 Evaluating PLL metric properties

4.1 Effects of sentence length

Like Salazar et al. (2020), we expect that models should, on average, assign lower probability to longer sentences. Thus, negative PLL (which reflects model surprisal) should be positively correlated with sentence length. However, the `PLL-original` metric violates this expectation in our test sentence set, which shows a negative correlation between the number of tokens and negative PLL. In contrast, `PLL-word-l2r` and `PLL-whole-word` metrics exhibit a positive correlation between the number of sentence tokens and negative PLL, just as the negative LL scores for a unidirectional model, GPT2-medium (Figure 3A).

4.2 Effects of word frequency

An appropriate (P)LL metric should reflect the fact that LLMs are sensitive to distributional patterns in training text corpora. In particular, we expect more frequent words to have higher (P)LL scores in the absence of contextual effects. This is indeed the case for GPT2-medium; however, the score inflation for multi-token words means that the `PLL-original` metric grossly overestimates the scores for low-frequency words (Figure 3B). `PLL-word-l2r` scores restore this relationship: their correlation with word frequency is much higher than for `PLL-original`. `PLL-whole-word` also performs well, although its correlation with word frequency is lower than for `PLL-word-l2r`, suggesting that it excessively penalizes OOV words.

5 Correlation with GPT-2 scores

We expect that PLL scores for bidirectional models should be at least somewhat consistent with LL scores for unidirectional models: both metrics are designed to serve are a proxy for sentence probability. Here, we show that the GPT-2/BERT score correlation for the `PLL-original` metric is very low, whereas correlation scores for `PLL-word-l2r` and `PLL-whole-word` are much higher (Figure 4), indicating the validity of this metric for cross-model comparison. As in Section 4.2, `PLL-word-l2r` slightly outperforms `PLL-whole-word`, likely because it does not penalize OOV words as severely.

See Appendices D-F for evidence that all three trends hold for larger models and for other datasets (although the effects in other datasets are attenuated due to a lower OOV ratio).

6 Effects on benchmarking

Here, we show that the choice of PLL metric affects benchmarking results for a popular, highly controlled, minimal pair linguistic benchmark: BLiMP. Despite the fact that the comparisons are highly controlled, different metrics yield different BLiMP scores. For all four tested models, `PLL-word-l2r` achieves the best overall BLiMP score (Table 1).

Model	Metric	Overall score
BERT (base)	PLL-original	84.2
	PLL-word-l2r	**84.7**
	PLL-whole-word	83.1
BERT (large)	PLL-original	84.8
	PLL-word-l2r	**85.0**
	PLL-whole-word	82.6
RoBERTa (base)	PLL-original	85.4
	PLL-word-l2r	**86.7**
	PLL-whole-word	85.4
RoBERTa (large)	PLL-original	86.5
	PLL-word-l2r	**87.5**
	PLL-whole-word	85.9

Table 1: Bidirectional model performance on the BLiMP benchmark using different PLL metrics.

See Appendix H for detailed scores.

7 Conclusion

We have shown that `PLL-word-l2r` is the preferred metric for evaluating sentence PLL under a masked language model, such as BERT. Although the results from studies using the `PLL-original` metric can still be informative, they become harder to interpret if the proportion of OOV words in their test set is high. Therefore, we recommend using `PLL-word-l2r` in future works.

Limitations

The proposed `PLL-word-l2r` metric has the same practical limitations as previous LL/PLL approaches. Most importantly, these scores can be influenced by many superfluous factors, such as the number of available synonyms (*computer* vs. *laptop*; Holtzman et al., 2021). We therefore expect our method to be most useful in highly controlled minimal pair or multiple choice setups.

Even more accurate metrics may emerge in the future. For instance, our approach pre-specifies the number of tokens in a word, thus limiting the space of possible alternatives. Future approaches might investigate a way to normalize the PLL score distribution over words with a varying number of tokens. Further, it would be interesting to attempt to estimate the joint probability of all tokens in a word instead of predicting them left-to-right (as in `PLL-word-l2r`) or without any other within-word contextual information (as in `PLL-whole-word`).

Finally, we test our approach on English text corpora; our results might not generalize to agglutinative languages (due to a high number of tokens per word and, therefore, increased uncertainty) and are of less relevance to isolating languages (where, if enough training data are available, most word-level items can be represented as single tokens).

Ethics Statement

In our proposed metric, word tokens are masked from left to right following the writing tradition in English; however, for speakers of languages such as Arabic, a "right to left" notation would be more intuitive. Note, however, that this is primarily a denotational difference that does not affect the score itself (LLMs do not discriminate left and right, only beginning and end). We do not anticipate any specific harms that would be intrinsically associated with the techniques described in this paper.

Acknowledgements

We thank Jacob Andreas, Evan Hernandez, and the anonymous ACL reviewers for their insightful feedback. CK was supported by the K. Lisa Yang Integrative Computational Neuroscience (ICoN) Center at MIT. AI was supported by MIT Quest for Intelligence.

References

Sudeep Bhatia and Russell Richie. 2022. Transformer networks of human conceptual knowledge. *Psychological Review*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Evelina Fedorenko, Idan Asher Blank, Matthew Siegelman, and Zachary Mineroff. 2020. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, 203:104348.

W Nelson Francis and Henry Kucera. 1979. Brown corpus manual. *Letters to the Editor*, 5(2):7.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. Syntaxgym: An online platform for targeted evaluation of language models. Association for Computational Linguistics (ACL).

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan S She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2022. Event knowledge in large language models: the gap between the impossible and the unlikely. *arXiv preprint arXiv:2212.01488*.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6365–6378, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712.

Joonbo Shin, Yoonhyung Lee, and Kyomin Jung. 2019. Effective sentence scoring method using bert for speech recognition. In *Asian Conference on Machine Learning*, pages 1081–1093. PMLR.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

Appendix

A Additional examples of score inflation

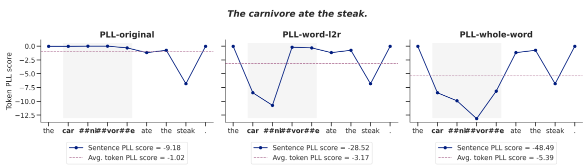

Figure 5: The `PLL-original` metric inflates the score of the word *carnivore*. `PLL-word-l2r` mitigate this issue, whereas `PLL-whole-word` overly penalizes the word. Model: `bert-base-cased`.

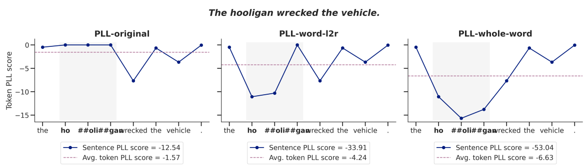

Figure 6: The `PLL-original` metric inflates the score of the word *hooligan*. `PLL-word-l2r` mitigate this issue, whereas `PLL-whole-word` overly penalizes the word. Model: `bert-base-cased`.

B Text preprocessing for (P)LL computation

The `minicons` library borrows the MLM preprocessing algorithm from Salazar et al. (2020): `[CLS]` and `[SEP]` tokens are prepended and appended to the text, respectively, and are not masked during PLL computation. For CLMs, we minimally adjust the `minicons` scorer library default and necessarily prepend the beginning of sentence token, `<|endoftext|>`, to the text, which enables us to get a probability for the first actual sentence token (see also the `lm-zoo` library; Gauthier et al., 2020). The (P)LLs of all special tokens are not counted toward the final sentence/word score.

When calculating the (P)LL score of individual words (to estimate word frequency effects), we place them in a neutral context *My word is* _. To ensure that the same pattern of results holds across multiple neutral contexts, we additionally test the context *I opened the dictionary and randomly picked a word. It was* _, as well as a no-context setup. These additional results are reported in Appendix E.1.

Word frequency was operationalized as the log of the number of occurrences of the word in the 2012 Google NGram corpus. Laplace smoothing was applied prior to taking the logarithm.

C Quantification of out-of-vocabulary words per dataset

Dataset	Model class	OOV ratio
EventsAdapt	BERT	19.6%
	RoBERTa	40.3%
	GPT	40.3%
LibriSpeech	BERT	8%
	RoBERTa	24.3%
	GPT	24.3%
Brown	BERT	8%
	RoBERTa	25%
	GPT	25%

Table 2: The out-of-vocabulary (OOV) ratio per dataset, quantified as the number of words split into at least two tokens by a given model's tokenizer divided by the total number of words in the dataset. GPT and RoBERTa models use byte-level Byte-Pair-Encoding tokenizers (Radford et al., 2019; Liu et al., 2019); BERT models use WordPiece tokenization (Devlin et al., 2018).

D Effects of sentence length

D.1 Larger LLM versions

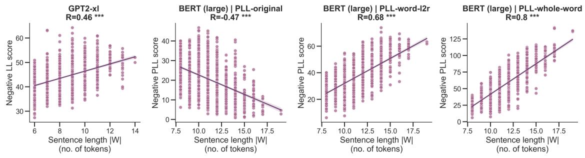

Figure 7: Sentence length effects for `gpt2-xl` and `bert-large-cased` on the EventsAdapt corpus.

D.2 Larger datasets

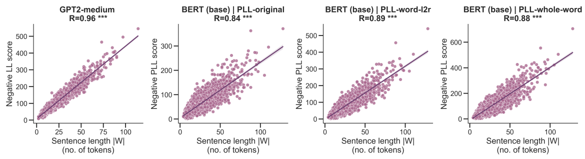

Figure 8: Sentence length effects for `gpt2-medium` and `bert-base-cased` on the LibriSpeech corpus.

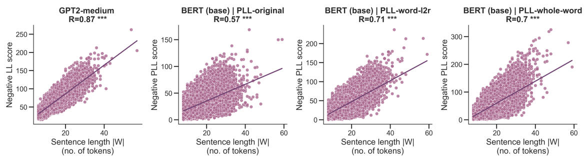

Figure 9: Sentence length effects for `gpt2-medium` and `bert-base-cased` on the Brown corpus.

E Effects of word frequency

E.1 Different word contexts

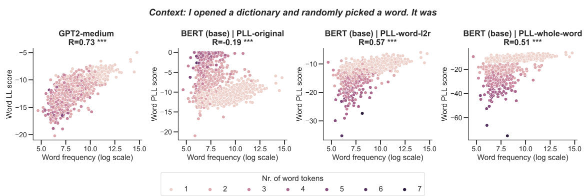

Figure 10: Word frequency effects for `bert-base-cased` on the EventsAdapt corpus. Word scores were retrieved with a neutral context: "`I opened a dictionary and randomly picked a word. It was` _".

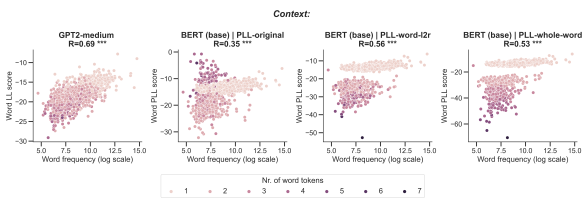

Figure 11: Word frequency effects for `bert-base-cased` on the EventsAdapt corpus. Word scores were retrieved without supporting context.

E.2 Different datasets

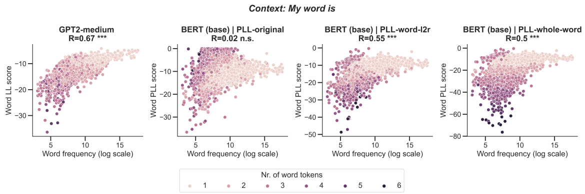

Figure 12: Word frequency effects for `bert-base-cased` on the LibriSpeech corpus. Word scores were retrieved with a neutral context: “`My word is` _”.

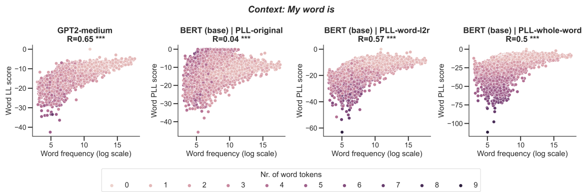

Figure 13: Word frequency effects for `bert-base-cased` on the Brown corpus. Word scores were retrieved with a neutral context: “`My word is` _”.

F Correlation with unidirectional models

F.1 Larger LLM versions

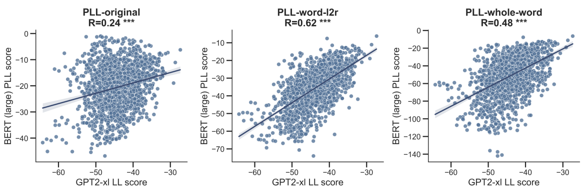

Figure 14: Correlation between `bert-large-cased` and `gpt2-xl` scores on the EventsAdapt corpus.

F.2 Larger datasets

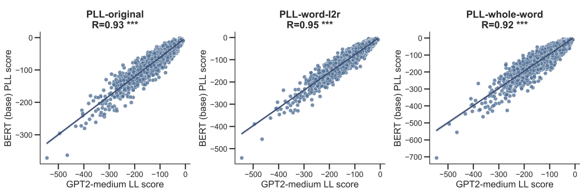

Figure 15: Correlation between `bert-base-cased` and `gpt2-medium` scores on the LibriSpeech corpus.

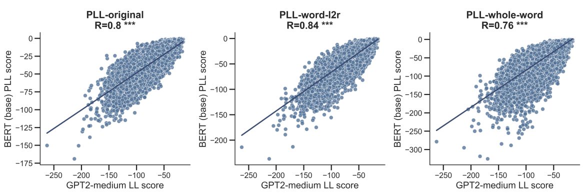

Figure 16: Correlation between `bert-base-cased` and `gpt2-medium` scores on the Brown corpus.

G Whole-sentence left-to-right token masking

Here, we report results for the scoring algorithm that masks the target token, s_t, and all sentence tokens to its right in a sentence S with n tokens (`PLL-sentence-l2r`). As in autoregressive language models, target token inference is thus conditioned solely on the token's leftward context: $P_{\text{MLM}}(s_t \mid S_{<t})$. The final sentence score is obtained as the sum of the log probabilities of each sentence token given its context:

$$\text{PLL}_{\text{sent}}(S) := \sum_{t=1}^{n} \log P_{\text{MLM}}(s_t \mid S_{<t}) \qquad (4)$$

Overall, the `PLL-sentence-l2r` metric satisfies the metric desiderata better than the `PLL-original` metric but worse than `PLL-word-l2r`. In addition, it is inferior to other metrics on the BLiMP evaluation benchmark (Appendix H), in line with previous reports of subpar generation quality (Wang and Cho, 2019).

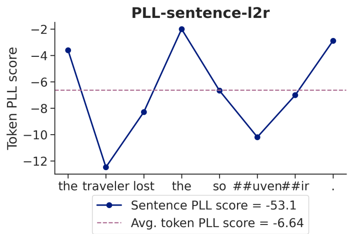

Figure 17: Scores for the motivating example computed with `PLL-sentence-l2r` (`bert-base-cased`).

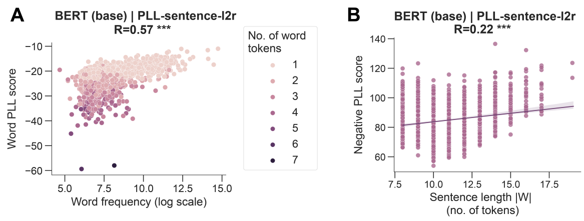

Figure 18: Word frequency **(A)** and sentence length **(B)** effects for scores computed with `PLL-sentence-l2r` on the EventsAdapt corpus (`bert-base-cased`)

.

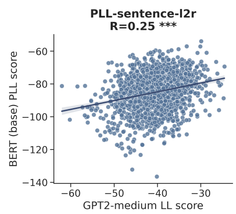

Figure 19: Correlation between `bert-base-cased` and `gpt2-medium` scores computed with `PLL-sentence-l2r` on the EventsAdapt corpus.

		Overall	ANA. AGR	ARG STR.	BINDING	CTRL. RAIS.	D-N AGR	ELLIPSIS	FILLER GAP	IRREGULAR	ISLAND	NPI	QUANTIFIERS	S-V AGR
BERT (base)	PLL-original	84.2	97.0	80.0	**82.3**	79.6	97.6	89.4	**83.1**	96.5	73.6	84.7	**71.2**	**92.4**
	PLL-word-l2r	**84.7**	**97.1**	**81.0**	**82.3**	**81.9**	**98.4**	**89.6**	83.0	96.5	**75.0**	**85.0**	69.8	92.1
	PLL-whole-word	83.1	96.6	76.5	81.5	80.5	96.9	87.1	82.5	**97.1**	74.9	83.8	69.2	88.5
	PLL-sentence-l2r	58.7	80.3	63.0	68.3	53.5	82.1	68.3	47.8	47.3	56.5	38.9	51.6	50.7
BERT (large)	PLL-original	84.8	**97.2**	**80.7**	**82.0**	82.7	97.6	**86.4**	**84.3**	**92.8**	77.0	83.4	**72.8**	**91.9**
	PLL-word-l2r	**85.0**	96.8	80.6	81.9	**84.8**	**97.8**	85.8	84.0	92.0	**78.8**	**83.6**	71.7	91.2
	PLL-whole-word	82.6	96.6	75.7	79.9	81.4	95.2	83.6	83.3	90.1	78.7	81.5	70.4	86.7
	PLL-sentence-l2r	59.8	61.5	63.0	71.3	60.5	71.8	58.3	58.5	63.0	50.2	42.8	51.9	63.0
RoBERTa (base)	PLL-original	85.4	97.3	83.5	77.8	81.9	97.0	91.4	**90.1**	**96.2**	80.7	81.0	**69.8**	91.9
	PLL-word-l2r	**86.7**	**97.8**	**84.8**	**78.7**	**84.9**	**98.3**	**91.6**	90.0	95.4	**81.0**	84.4	69.7	**94.0**
	PLL-whole-word	85.4	97.6	80.9	76.6	85.2	96.6	**91.6**	90.0	95.6	80.2	**84.7**	69.6	91.1
	PLL-sentence-l2r	79.3	97.0	79.9	71.2	78.4	95.0	84.8	82.6	85.0	68.2	80.6	58.4	81.6
RoBERTa (large)	PLL-original	86.5	97.8	84.6	79.1	84.1	96.8	**90.8**	88.9	**96.8**	**83.4**	85.5	70.2	91.4
	PLL-word-l2r	**87.5**	98.0	**85.0**	**80.0**	86.8	**98.3**	90.4	**89.1**	95.7	**83.4**	**88.0**	**70.3**	**93.2**
	PLL-whole-word	85.9	98.2	80.2	78.0	**87.1**	96.0	90.1	88.9	95.6	82.2	**88.0**	69.8	89.7
	PLL-sentence-l2r	80.4	**98.8**	82.5	71.8	80.4	95.1	82.0	80.8	91.6	73.0	76.6	57.8	86.0
Human		*88.6*	*97.5*	*90.0*	*87.3*	*83.9*	*92.2*	*85.0*	*86.9*	*97.0*	*84.9*	*88.1*	*86.6*	*90.9*

Table 3: Unsupervised performance (forced choice accuracy) on all BLiMP benchmark paradigms, using the original and adjusted PLL sentence scoring methods. `PLL-original` scores replicate those reported in Salazar et al. (2020). Human scores are taken from Warstadt et al. (2020).

H Detailed BLiMP benchmark results

Table 3 shows results for each sentence suite within the BLiMP benchmark (in addition to the overall scores reported in the main text). All models shown in Tables 1 and 3 are cased models. `PLL-original` scores replicate those reported in Salazar et al. (2020).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
8

☒ A2. Did you discuss any potential risks of your work?
we do not anticipate specific risks associated with our work

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

all

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
the models are available on huggingface, and the experiments are computationally light

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3 and Appendix A (no hyperparameter search was conducted though)

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
all results figures

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.