Why Does Surprisal From Smaller GPT-2 Models Provide Better Fit to Human Reading Times?

Anonymous ACL submission

Abstract

This work presents an in-depth analysis of an observation that contradicts the findings of recent work in computational psycholinguistics, namely that smaller GPT-2 models that show higher test perplexity nonetheless generate surprisal estimates that are more predictive of human reading times. Analysis of the surprisal values shows that rare proper nouns, which are typically tokenized into multiple subword tokens, are systematically assigned lower surprisal values by the larger GPT-2 models. A comparison of residual errors from regression models fit to reading times reveals that regression models with surprisal predictors from smaller GPT-2 models have significantly lower mean absolute errors on words that are tokenized into multiple tokens, while this trend is not observed on words that are kept intact. These results indicate that the ability of larger GPT-2 models to predict internal pieces of rare words more accurately makes their surprisal estimates deviate from humanlike expectations that manifest in self-paced reading times and eye-gaze durations.

1 Introduction

Expectation-based theories of sentence processing (Hale, 2001; Levy, 2008) posit that processing difficulty is mainly driven by how predictable upcoming linguistic material is given its context. In support of this position, predictability quantified through information-theoretical surprisal (Shannon, 1948) has been shown to strongly correlate with behavioral and neural measures of processing difficulty (Demberg and Keller, 2008; Smith and Levy, 2013; Hale et al., 2018; Shain et al., 2020).

In previous studies, language models (LMs), which directly define a conditional probability distribution of a word given its context, have been evaluated as surprisal-based cognitive models of sentence processing. Surprisal estimates from several well-established types of LMs, including n-gram models, Simple Recurrent Networks (Elman, 1991), and Long Short-Term Memory networks (LSTM; Hochreiter and Schmidhuber, 1997), have been compared against behavioral measures of processing difficulty (e.g. Smith and Levy, 2013; Goodkind and Bicknell, 2018; Aurnhammer and Frank, 2019). More recently, as Transformer-based (Vaswani et al., 2017) models have dominated many NLP tasks, both large pretrained and smaller 'trained-from-scratch' Transformer-based LMs have been evaluated as models of processing difficulty (Wilcox et al., 2020; Hao et al., 2020; Merkx and Frank, 2021; Schrimpf et al., 2021).

A consistent finding that emerged out of these studies is that better language models are also better models of comprehension difficulty, or in other words, there is a negative correlation between language model perplexity and fit to human reading times. Goodkind and Bicknell (2018) compared surprisal estimates from a set of n-gram and LSTM LMs and observed a negative linear relationship between perplexity and regression model fit. Wilcox et al. (2020) evaluated n-gram, LSTM, Transformer, and RNNG (Dyer et al., 2016) models and replicated the negative relationship, although they note a more exponential relationship at certain intervals.[1]

2 Background

Recently, however, it was observed that when pretrained GPT-2 models (Radford et al., 2019) are used to generate surprisal estimates, surprisal from *GPT-2 Small*, which has the least number of parameters, makes the biggest contribution to regression model fit on self-paced reading times (Anonymous, under review). Using self-paced reading times from the Natural Stories Corpus (Futrell et al., 2021), the

[1] Although counterexamples to this trend have been noted, they were based on comparisons of LMs and incremental parsers that were trained on different data (Oh et al., 2021) or evaluation on a language with different syntactic head-directionality than English (Kuribayashi et al., 2021).

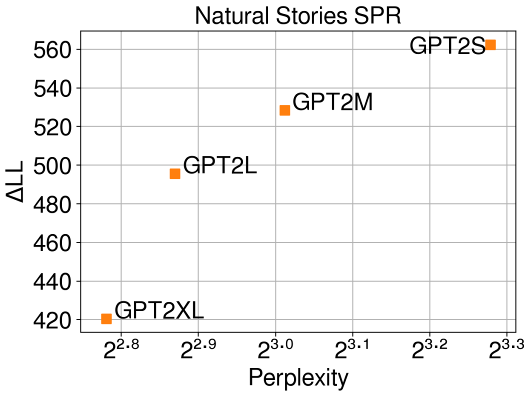

Figure 1: Perplexity measures from each GPT-2 model, and improvements in regression model log-likelihood from including each surprisal estimate on Natural Stories self-paced reading data.

authors calculated the increase in log-likelihood (ΔLL) to a baseline linear-mixed effects (LME) model as a result of including a surprisal predictor.[2] Their results in Figure 1 show a robust *positive* correlation between language model perplexity and predictive power of surprisal predictors from pretrained GPT-2 models of different sizes.[3] This effect was then replicated on the Dundee eye-tracking corpus (Kennedy et al., 2003).

As the different variants of pretrained GPT-2 models share the primary architecture (i.e. autoregressive Transformers) and training data, this offers an especially strong counterexample to recent works that observe a negative relationship between these two variables (Goodkind and Bicknell, 2018; Hao et al., 2020; Wilcox et al., 2020).

3 Methods

The current work attempts to provide an explanation for the positive correlation observed between language model perplexity and fit to self-paced reading times by reproducing these results and conducting an error analysis with the regression models.[4]

[2]The baseline regression model included predictors that capture low-level cognitive processing, such as word length measured in characters and index of word position within each sentence. All predictors were centered and scaled prior to model fitting, and the LME models included by-subject random slopes for all fixed effects and random intercepts for each word and subject-sentence interaction.

[3]The authors observe the same trend when unigram surprisal is included in the baseline and spillover effects are controlled for through the use of continuous-time deconvolutional regression (CDR; Shain and Schuler, 2021).

[4]All code used in this work is available at: `github.com/xxx/yyy`

3.1 Response Data

Following the results described in Section 2, we evaluated surprisal predictors on self-paced reading times from the Natural Stories Corpus (Futrell et al., 2021), which contains data from 181 subjects that read 10 naturalistic English stories consisting of 10,245 tokens. The data were filtered to exclude observations corresponding to sentence-initial and sentence-final words, observations from subjects who answered fewer than four comprehension questions correctly, and observations with durations shorter than 100 ms or longer than 3000 ms. This resulted in a total of 770,102 observations, which were subsequently partitioned into an exploratory set of 384,905 observations and a held-out set of 385,197 observations.[5] All observations were log-transformed prior to model fitting.

3.2 Predictors

The results in Section 2 used surprisal estimates calculated from four different variants of pretrained GPT-2 models[6] (Radford et al., 2019), which are decoder-only autogressive Transformer models that differ in their sizes:

- *GPT2S*: GPT-2 Small, which has 12 layers and ~124M parameters.
- *GPT2M*: GPT-2 Medium, which has 24 layers and ~355M parameters.
- *GPT2L*: GPT-2 Large, which has 36 layers and ~774M parameters.
- *GPT2XL*: GPT-2 XL, which has 48 layers and ~1558M parameters.

Each story of the Natural Stories Corpus was tokenized according GPT-2's byte-pair encoding (BPE; Sennrich et al., 2016) tokenizer and was provided to each pretrained GPT-2 model to calculate surprisal estimates. In cases where a single word w_t was tokenized into multiple subword tokens, negative log probabilities of subword tokens corresponding to w_t were added together to calculate $\mathsf{S}(w_t) = -\log \mathsf{P}(w_t \mid w_{1..t-1})$.

3.3 Regression Modeling and Error Analysis

Subsequently, four LME models that contain the baseline predictors (i.e. word length and word position) and each of the GPT-2 surprisal predictors

[5]The results in Figure 1 are from regression models fit on the held-out set.

[6]The pretrained models are publicly available at `https://github.com/openai/gpt-2`.

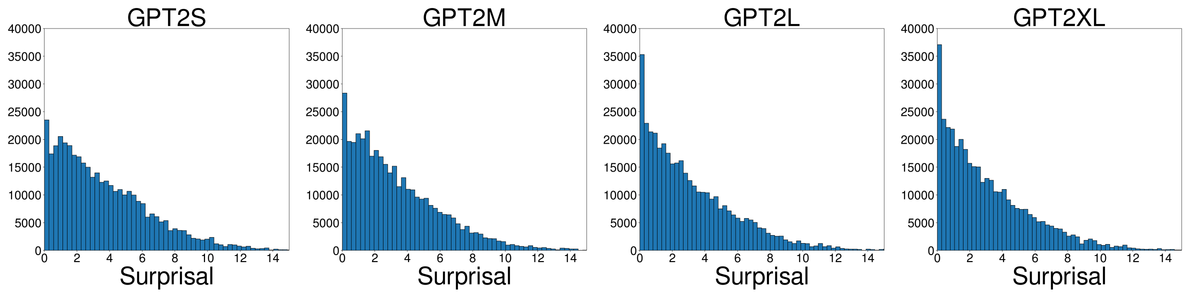

Figure 2: Histogram of word-level surprisal values on the held-out set of Natural Stories Corpus from different pretrained GPT-2 models.

Sentence #	Word #	Word	*GPT2Ssurp*	*GPT2Msurp*	*GPT2Lsurp*	*GPT2XLsurp*	# Subwords
382	6	Pflock,	16.9745	12.1140	6.2818	1.7086	4
362	13	Marcel,	11.7783	4.4075	0.4812	0.4383	2
1	19	jennies	13.1263	9.1347	4.6793	2.6570	3
379	26	Mogul,	11.1371	2.9520	1.0758	1.1000	3
451	26	coprolalia,	21.8774	14.2319	10.2438	11.8560	4
141	24	dollar	8.9853	1.0388	1.5773	0.1183	1
446	11	throat-clearing,	14.7768	9.8318	8.6016	6.3010	5
388	21	Provinces,	12.6217	9.6031	9.3428	4.3365	4
382	53	Agustin	7.8970	6.4648	1.7403	0.1384	3
362	9	Stanton	8.6183	6.3176	4.4433	0.9583	1

Table 1: Top 10 words with the biggest surprisal value differences between the *GPT2S* and *GPT2XL* models, and their corresponding surprisal values from the *GPT2M* and *GPT2L* models.

were fit to the held-out set of self-paced reading times using `lme4` (Bates et al., 2015). After the models were fitted, the predictions for all data points ($\hat{y}$) were generated in order to calculate the residual errors ($y - \hat{y}$) from each regression model. Additionally, surprisal values from the different pretrained GPT-2 models were analyzed in order to identify where they make the most divergent predictions.

4 Results

The histogram of surprisal values in Figure 2 shows that as the model size becomes larger, surprisal values of more words tend to be concentrated in the lowermost bin. This indicates that the larger pretrained models are indeed better language models in terms of next-word prediction, and is also consistent with the trend of perplexity measures reported in Figure 1. However, this may also be the reason that the surprisal estimates from the larger GPT-2 models lead to worse fit on self-paced reading times; since more data points are assigned near-zero surprisal, the regression model may not be able to accurately predict potentially high reading times at those points.

In order to identify the words that are assigned relatively low surprisal values by the larger models but relatively high surprisal values by the smaller models, the words were sorted according to the difference between the surprisal values from the *GPT2S* and *GPT2XL* models, which have the most divergent profiles. Table 1 presents the surprisal values for the top 10 words that show the biggest difference between the *GPT2S* and *GPT2XL* models. As can be seen, most of these words demonstrate a systematic decrease in their surprisal values as the model size increases, which indicates that these are the words that are partially responsible for the trend observed in Figure 2. Additionally, most of these words are rare proper nouns, and were therefore tokenized into multiple subword tokens by the GPT-2 models. Given these two observations, it was hypothesized that the better regression model fit observed for the smaller GPT-2 models is mainly driven by more accurate predictions of reading times for such multi-token words.

To test this hypothesis, the data points in the held-out set of Natural Stories Corpus were separated according to whether each word remained intact or was tokenized into multiple subword tokens by the GPT-2 model. This resulted in a *single-token* partition of 337,752 data points, and a *multiple-token* partition of 47,445 data points. Subsequently, the absolute errors ($|y - \hat{y}|$) from the four regression models were compared on each set. The above hypothesis would be supported if the absolute er-

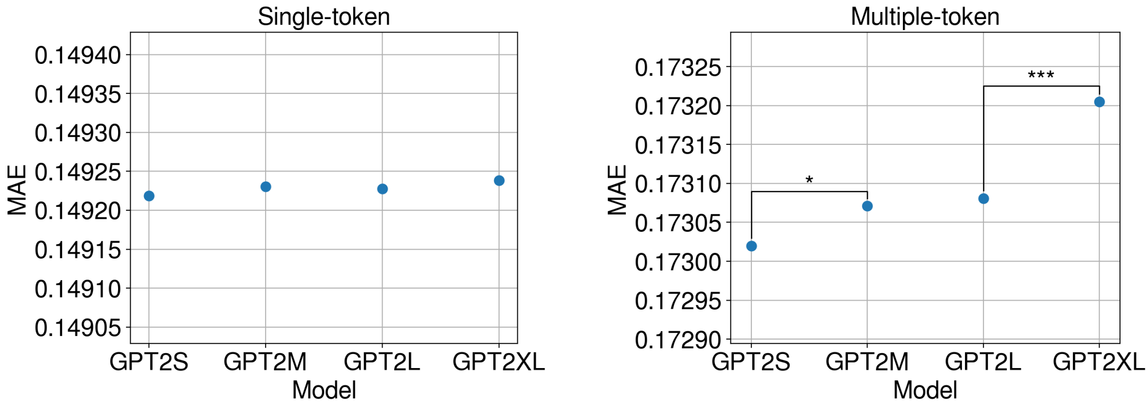

Figure 3: Mean absolute errors from each regression model on data points consisting of single-token words (left) and multiple-token words (right) from the Natural Stories Corpus. Statistical significance of the difference between means was determined by a paired permutation test at the event level (*: $p < 0.05$, ***: $p < 0.001$). Note that the figures share the scale of the y-axis.

rors were similar across regression models on the *single-token* partition, but not on the *multiple-token* partition.

The results in Figure 3 show that for all four regression models, the mean absolute errors are higher on words with multiple tokens, which indicates that all GPT-2 models tend to generate surprisal estimates that do not align well with self-paced reading times on these words. More importantly, on the *multiple-token* partition, mean absolute errors are lower for the regression models with surprisal estimates from the smaller GPT-2 models, which is consistent with the trend in ΔLL observed in Figure 1. Pairwise permutation tests with mean absolute errors between "neighboring" models show that the difference between *GPT2S* and *GPT2M* models, as well as that between the *GPT2L* and *GPT2XL* models is statistically significant. In contrast, this trend is not attested in the mean absolute errors on the *single-token* partition, where none of the difference in mean absolute errors between neighboring models are statistically significant. Taken together, these results indicate that the better fit to human reading times achieved by surprisal estimates from smaller GPT-2 models achieve is partly driven by their characteristic of assigning high surprisal values to multi-token words. In other words, the extra parameters of larger models may be improving transitions between subword units in a way that is beyond human ability.

5 Conclusion

This paper presents an in-depth analysis of an observation that contradicts the findings of recent work in computational psycholinguistics, namely that smaller pretrained GPT-2 models that perform *worse* in terms of next-word prediction (i.e. higher perplexity) nonetheless generate surprisal estimates that are *more predictive* of human reading times (i.e. higher contribution to regression model fit).

Analysis of the surprisal values from each of the GPT-2 models showed that as model size increases, more words are assigned near-zero surprisal, which confirms the ability of larger models to predict upcoming words more accurately. In order to examine whether this capability of larger models are responsible for the unexpected trend in fit to human reading times, words that show the biggest difference in surprisal values between the smallest and largest GPT-2 models were identified. This analysis revealed that rare proper nouns or words with punctuation marks, which are typically tokenized into multiple subword tokens, are systematically assigned lower surprisal values by the larger GPT-2 models. A subsequent comparison of residual errors from the regression models on reading times of words that are tokenized (i.e. *multiple-token*) showed that the regression models with surprisal estimates from smaller GPT-2 models have significantly lower mean absolute errors, while this trend was not observed on reading times of words that are kept intact (i.e. *single-token*).

These results indicate that the ability of larger GPT-2 models to predict internal pieces of rare words more accurately makes their surprisal estimates deviate from humanlike expectations that manifest in self-paced reading times.

6 Ethical Considerations

Experiments presented in this work used datasets from previously published research (Futrell et al., 2021), in which the procedures for data collection and validation are outlined.

References

Christoph Aurnhammer and Stefan L. Frank. 2019. Comparing gated and simple recurrent neural network architectures as models of human sentence processing. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, pages 112–118.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209.

Jeffrey L Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225.

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.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, pages 10–18.

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.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. 2018. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2727–2736.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the 10th Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–86.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The Dundee Corpus. In *Proceedings of the 12th European conference on eye movement*.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5203–5217.

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

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22, Online. Association for Computational Linguistics.

Byung-Doh Oh, Christian Clark, and William Schuler. 2021. Surprisal estimators for human reading times need character models. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3746–3757.

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

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *bioRXiv*.

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, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. 2020. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138.

Cory Shain and William Schuler. 2021. Continuous-Time Deconvolutional Regression for Psycholinguistic Modeling. *Cognition*, 215.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 1707–1713.