Locally Biased Transformers Better Align with Human Reading Times

Andrea Gregor de Varda
University of Milano – Bicocca
a.devarda@campus.unimib.it

Marco Marelli
University of Milano – Bicocca
marco.marelli@unimib.it

Abstract

Recent psycholinguistic theories emphasize the interdependence between linguistic expectations and memory limitations in human language processing. We modify the self-attention mechanism of a transformer model to simulate a lossy context representation, biasing the model's predictions to give additional weight to the local linguistic context. We show that surprisal estimates from our locally-biased model generally provide a better fit to human psychometric data, underscoring the sensitivity of the human parser to local linguistic information.

1 Introduction

In recent years, transformer models (Vaswani et al., 2017) have gained prominence in psycholinguistics due to their impressive predictive performance in forecasting psychometric measurements such as reading times (Hao et al., 2020; Merkx and Frank, 2021; de Varda et al., 2023; Hoover et al., 2023; Oh and Schuler, 2022, 2023, *inter alia*). These models excel at capturing complex linguistic dependencies, making them valuable tools in analyzing human language-processing behaviors. Much of the work that relates probabilistic estimates from transformer models with human processing has been conducted within the framework of surprisal theory, which posits that the difficulty experienced during processing is proportional to the negative logarithm of the probability of a word given its preceding context. In this context, transformer models, which are able to generate highly accurate probabilistic predictions for sequences of text, have been instrumental in providing empirical support for surprisal theory (Wilcox et al., 2020; Shain et al., 2022; de Varda and Marelli, 2022, 2023). However, despite their substantial predictive power, transformer models exhibit some design features that lack cognitive plausibility. One significant departure from human language processing is their ability to access in parallel the entire linguistic context within their input size. Unlike these models, human language comprehension is inherently incremental (Smith and Levy, 2013). Humans eagerly integrate in their representation of the context each linguistic unit as soon as it is encountered, and they cannot typically store in working memory the whole linguistic context. Thus, the transformer model's *all-at-once* approach to processing information starkly contrasts with the sequential and resource-constrained manner in which humans receive and interpret linguistic input, suggesting a need for models that more closely mirror human cognitive limitations and processing strategies.

To address these limitations, we introduce a modification to the self-attention mechanism of the transformer model, aimed at simulating a lossy memory representation, where linguistic units that are further away from the current word are assigned exponentially decaying attention scores. By doing so, our model aims to replicate the kind of linguistic processing that characterizes the human language parser, where recent information plays a significant role (Goodkind and Bicknell, 2021).

The evaluation of our locally-biased transformer model involves the employment of the surprisal – i.e., negative log probability – it assigns to words in context to predict human psychometric data, considering five large-scale datasets of eye movements and self-paced reading times in English. We show that our locally-biased model provides surprisal estimates that align more closely with human psychometric data than a standard pre-trained model.

2 Related work

Models that can explain the cognitive cost associated with sentence processing can be broadly divided into expectation- and memory-based theories. Expectation-based theories (such as surprisal theory; Levy, 2008; Hale, 2001) emphasize the role of contextual predictability as a core determinant of processing demands. Support for such theo-

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 30–36
August 15, 2024 ©2024 Association for Computational Linguistics

ries has come from several studies demonstrating reduced cognitive load in response to predictable words (e.g., Frank and Thompson, 2012; Frank et al., 2015; Wilcox et al., 2020). Memory-based theories, in contrast, are based on the idea that integrating the upcoming words into the context representation depends on the retrieval (Lewis and Vasishth, 2005) and storage (Gibson, 1998, 2000) of previous words in working memory. Support for memory-based theories comes from the difficulty in integrating words that are linearly distant in a sentence (dependency locality effects; Grodner and Gibson, 2005; Fedorenko et al., 2013).

In recent years, there have been proposals to reconcile expectation- and memory-based approaches into unified models. While the first combined theories posited limited (Demberg and Keller, 2008, 2009) or no interaction between memory and predictability (Rasmussen and Schuler, 2018; see Futrell et al., 2020), some recently developed frameworks account for complex interactions between the two (Futrell et al., 2020; Hahn et al., 2022). In particular, *lossy-context surprisal theory* (henceforth LCST; Futrell et al., 2020) holds that the processing difficulty associated with a word is proportional its surprisal, conditioned by a lossy (i.e., noisy) memory representation of the context. Hahn et al. (2022) presented a computationally-specified model of LCST (*resource-rational LCST*) that computes retention probabilities for each word in the context, based on the word's identity and position in the sentence. Similarly, Kuribayashi et al. (2022) have shown that reducing the number of words in input to language models improves the fit of the surprisal estimates to human reading times. Our modelling approach is reminiscent of LCST in that it assumes that the processing cost associated with a word is proportional to its surprisal, conditioned by the previous context where linearly distant words contribute less to its prediction.

In our modelling effort, we modify the attention scores of a transformer model to mimic the human difficulty in retrieving distant linguistic elements. We are not the first in drawing a parallelism between the self-attention mechanism and (cue-based) memory retrieval (Merkx and Frank, 2021; Hyun et al., 2022; Oh and Schuler, 2022; Timkey and Linzen, 2023). Indeed, like the self-attention mechanism scores the weights to assign to the words in input based on the compatibility between keys and queries, cue-based retrieval theories posit that items in working memory are accessed by comparing the retrieval cues of the current word with the features of the items in working memory (Timkey and Linzen, 2023). Our choice to bias the transformer model's retrieval process towards the recent linguistic context is supported by vast evidence in psycholinguistics showing that local information holds a privileged role in human language processing, with a chief example being word frequency. Indeed, it is well-known that the frequency of a word (which is proportional to its unigram probability) influences its reading times above and beyond its contextual predictability (Rayner, 1998; Shain, 2023). Furthermore, other studies have detected an effect of N-gram surprisal beyond the effect of surprisal as calculated from larger sentential contexts (Goodkind and Bicknell, 2021). Note that the idea of assigning reduced weights to distal elements has a long-standing tradition in psychologically-oriented computational models of semantic memory: one of the first distributional semantic models, the Hyperspace Analogue to Language (HAL; Lund and Burgess, 1996), weighs word-by-word co-occurrences as a function of their linear distance. It is noteworthy that the idea that human language processing privileges recent information was explicitly implemented in early work in computational semantics, and lost with later developments in the field.

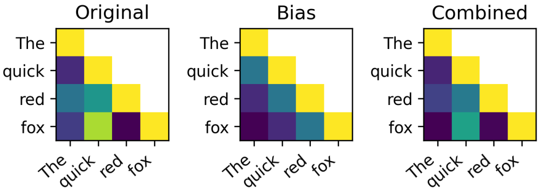

Figure 1: Visual example of our custom modification of the model's attention pattern. The original attention scores (left) and the exponentially decaying bias (center) are summed to derive the combined attention scores (right).

3 Methods and materials

We modified the pre-trained GPT-2 model (Radford et al., 2019) to increase the attention weights associated to the nearby words. We conducted our analyses on the smallest GPT-2 variant, as it has been proven to be particularly effective at modelling human reading times (Shain et al., 2022).

All the code supporting our analyses is publicly available on GitHub.[1]

[1] https://github.com/Andrea-de-Varda/local_attention_reading_times

3.1 Locally biased attention

Attention weights are initially computed using the standard dot-product attention, involving the multiplication of the query matrix with the transpose of the key matrix ($W = QK^T$). Then, an exponential decay bias matrix B is computed using an exponential decay function based on the absolute differences between positions in the sequence, scaled by a decay rate. Thus, the bias is computed as $B_{i,j} = e^{-\lambda|i-j|}$, where $i, j \in \{0, 1, \ldots, n-1\}$ indicate the position of two tokens in the sequence, n specifies the sequence length, and λ is the decay rate. As a final step, we blend together the original attention weights with the exponential decay bias with a weighted sum to obtain the final attention weights $A = (1 - \alpha) \cdot W + \alpha \cdot B$. As a last step, the softmax function is applied to A. A visual summary of this procedure is provided in Figure 1. Note that both α and λ serve as free parameters in our modified attention mechanism. To identify the optimal values for these parameters, we employed hyperparameter tuning techniques as detailed in §3.4.

3.2 Data

The analyses were run on three eye-tracking and three self-paced reading datasets. The eye-tracking resources we considered were the Provo corpus (Luke and Christianson, 2018; $N = 2659$[2]), the English portion of the MECO corpus (Siegelman et al., 2022; $N = 2096$), and the UCL$_{\text{ET}}$ corpus (Frank et al., 2013, $N = 1726$). The three self-paced reading datasets were the UCL$_{\text{SPR}}$ dataset (Frank et al., 2013, $N = 1726$), the Brown corpus (Smith and Levy, 2013; $N = 5862$), and the Natural Stories reading times corpus (NatStor, Futrell et al., 2021; $N = 8779$). In our analysis of the eye-tracking data, we focused on first-pass gaze duration times, in accordance with previous research in computational psycholinguistics (see for instance Aurnhammer and Frank, 2019; Goodkind and Bicknell, 2018; Smith and Levy, 2013; Wilcox et al., 2020). For words that did not receive any fixation, we assigned a gaze duration time of zero. We excluded words located at the beginning of sentences from our analyses. Beyond this exclusion, we did not implement any further filtering criteria. To obtain word-level gaze duration times, we calculated the average word reading times across all participants. Likewise, for the self-paced reading tasks, we calculated the average reaction times on the target word across participants.

3.3 Analyses

In our analyses, the dependent variable of interest (either gaze duration or self-paced reading times) was predicted with a linear model including surprisal, subtitle-based log-frequency (Brysbaert and New, 2009), and orthographic length as regressors.[3] Surprisal values obtained with our locally biased transformer model (s_{loc}) were compared with the estimates produced by the original GPT-2 model (s_{orig}). For each psychometric dataset, we identified the best model with the Akaike Information Criterion (AIC; Akaike, 1998). In particular, we subtracted the AIC$_{loc}$ obtained with s_{loc} to the AIC$_{orig}$ obtained with s_{orig} to obtain a ΔAIC. In interpreting the ΔAIC scores, we refer to the guidelines offered by Burnham and Anderson (2004), which indicate that if two models have a $\Delta\text{AIC} \leq 2$, they both have substantial support; if $4 \leq \Delta\text{AIC} \leq 7$, the best model has considerably more support, and if $\Delta\text{AIC} \geq 10$, the worse model has essentially no support[4].

3.4 Hyperparameter tuning

To identify the best values for the parameters α and λ (see §3.1), we iteratively sampled from the hyperparameter space – restricted to $\lambda \in (0, 100)$ and $\alpha \in (0, 1)$ – using a Tree-structured Parzen Estimator algorithm. For each (λ, α) pair, we specified a locally-biased GPT-2 model with such hyperparameters, and derived surprisal values for the sentences in the Provo corpus. Then, we fit a linear model predicting the reading times in the Provo corpus from the obtained surprisal values, log-frequency, and word length; through hyperparameter tuning we seeked to minimize the negative log likelihood of the model ($N_{trials} = 100$). As a result of this procedure, we identified $\lambda = 82.86$ and $\alpha = 0.37$ as the optimal values for the two parameters. The parameters obtained in the Provo corpus were transferred to the other behavioral datasets without further tuning.

[2] N is the number of datapoints after data aggregation.

[3] The exact linear model specification was DV $\sim$ LENGTH(w$_i$) + FREQUENCY(w$_i$) + SURPRISAL(w$_i$)

[4] In terms of relative likelihood, if $\Delta\text{AIC} \leq 2$ the worse model is 0.3678 times as probable as the best model to minimize the information loss; with $4 \leq \Delta\text{AIC} \leq 7$, this probability is in the range $(0.0302, 0.1353)$, and with $\Delta\text{AIC} \geq 10$ the probability is lower than 0.0067.

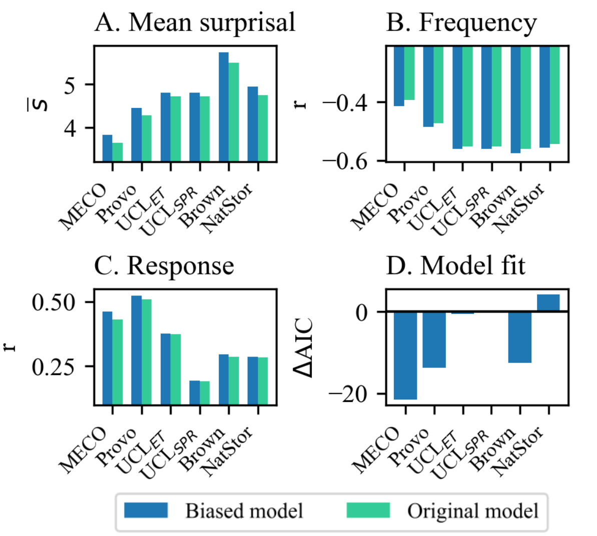

Figure 2: From the top left, clockwise: **A.** Average per-word surprisal as computed with the biased model and the baseline; **B.** Correlation of the obtained surprisal estimates with log frequency; **C.** Correlation of the surprisal estimates with the psychometric measurements; **D.** ΔAIC between the locally biased model and the original model.

4 Results

The results of our analyses are reported in Figure 2. Our locally biased transformer assigned higher average per-word surprisal values to the input texts across all datasets (A), showing a reduced autoregressive accuracy with respect to its unbiased counterpart. The obtained surprisal estimates correlated more strongly with log frequency values (B) and with the behavioral responses considered (C). Furthermore, our comparison between the biased and the original GPT-2 model (D) revealed that our modification of the attention mechanism caused a substantial increase in predictive performance in three behavioral datasets, encompassing both eye-tracking (MECO, $\Delta\text{AIC} = -21.55$; Provo, $\Delta\text{AIC} = -13.72$) and self-paced reading (Brown, $\Delta\text{AIC} = -12.55$). Our manipulation had no effect in the UCL corpus (UCL_{ET}, $\Delta\text{AIC} = -0.59$; UCL_{SPR}, $\Delta\text{AIC} = -0.06$) and resulted in a poorer fit in the NatStor dataset ($\Delta\text{AIC} = 4.31$).

5 Discussion

In this study, we have demonstrated that a modification of the GPT-2 model to emphasize local context via a locally-biased attention mechanism results in surprisal estimates that are more strongly correlated with human reading times, and generally display a better fit to human psychometric data. An exception to this second observation is offered by the NatStor and UCL corpora; in Appendix A, we report tentative evidence that the model improvement seems to be related to the average sentence length in the corpus. In particular, our locally-biased attention seems to be particularly beneficial in cases where the sentences are longer. This finding is compatible with the idea of a lossy representation of the context, where memory constraints become more marked for longer text sequences. Future approaches could consider dynamically manipulating the α parameter as a function of sentence lenght, adjusting the strength of the bias to cases where the human memory is taxed more strongly.

In LCST, the way memory representations degrade typically results in a word's contextual processing cost approaching its context-independent processing cost, as predicted by its standalone probability (Futrell et al., 2020). Essentially, as the fidelity of a listener's memory representations diminishes, their anticipations increasingly align with the prior, context-independent unigram probability. Our findings empirically demonstrate that this is the case, as the surprisal estimates from our locally-biased transformer tend to regress towards word frequency estimates (Figure 2, B). While our intervention on the attention mechanism is directly inspired by LCST, it should be noted that this implementation does not respect all the assumptions of the theory. In particular, LCST posits as an assumption the inaccessibility of the context (Claim 3); here, the context is always available to the model, albeit reduced attention weights are assigned to the elements that are linearly distant from each other.

Importantly, our modification of the attention mechanism resulted in models that were less performant in next-word prediction (see Figure 2, A). This is of course to be expected, as the addition of the exponential decay bias to the attention scores produces final attention weights that deviate from the ones that have been optimized for autoregression. Nonetheless, our results show that a worse NLP model can constitute a better cognitive model in terms of fit to psychometric data. This result challenges the *quality-power hypothesis* (QP; Wilcox et al., 2023), which posits that more accurate language models (i.e., models whose surprisal estimate better approximate the values from the data-generating distribution) should provide surprisal estimates that better fit behavioral data. However, QP does not hold if the probabilistic information that humans deploy in real time is systematically biased with respect to the data-generating distribution. One example of this systematic deviation is offered by the sensitivity of the human parser to local word co-occurrence statistics (Goodkind and Bicknell, 2021), which is exactly what we model in the present paper. Thus, our results show that human-like language processing might inherently involve biases and limitations that deviate from optimal statistical models.

Limitations

This study, while providing insights into the integration of cognitive constraints in transformer models, is not without limitations. The approach assumes a fixed attention decay rate, a simplification that might not fully capture the dynamic nature of human memory in language processing. Furthermore, while we consider more psychometric datasets than most studies in computational psycholinguistics, the fact that we have only five corpora does not allow us to draw conclusive inferences on the impact of average sentence length on the relative performance of our locally biased models.

References

Hirotogu Akaike. 1998. Information theory and an extension of the maximum likelihood principle. *Selected papers of hirotugu akaike*, pages 199–213.

Christoph Aurnhammer and Stefan L Frank. 2019. Comparing gated and simple recurrent neural network architectures as models of human sentence processing.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Kenneth P Burnham and David R Anderson. 2004. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304.

Andrea de Varda and Marco Marelli. 2023. Scaling in cognitive modelling: A multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–149.

Andrea Gregor de Varda and Marco Marelli. 2022. The effects of surprisal across languages: Results from native and non-native reading. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 138–144.

Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. 2023. Cloze probability, predictability ratings, and computational estimates for 205 english sentences, aligned with existing eeg and reading time data. *Behavior Research Methods*, pages 1–24.

Vera Demberg and Frank Keller. 2008. A psycholinguistically motivated version of tag. In *Proceedings of the Ninth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+ 9)*, pages 25–32.

Vera Demberg and Frank Keller. 2009. A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of the annual meeting of the cognitive science society*, volume 31.

Evelina Fedorenko, Rebecca Woodbury, and Edward Gibson. 2013. Direct evidence of memory retrieval as a source of difficulty in non-local dependencies in language. *Cognitive science*, 37(2):378–394.

Stefan Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The erp response to the amount of information conveyed by words in sentences. *Brain and language*, 140:1–11.

Stefan Frank and Robin Thompson. 2012. Early effects of word surprisal on pupil size during reading. In *Proceedings of the annual meeting of the cognitive science society*, volume 34.

Stefan L Frank, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior research methods*, 45(4):1182–1190.

Richard Futrell, Edward Gibson, and Roger P Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive science*, 44(3):e12814.

Richard Futrell, Edward Gibson, Harry J Tily, Idan Blank, Anastasia Vishnevetsky, Steven T 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.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000:95–126.

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 (CMCL 2018)*, pages 10–18.

Adam Goodkind and Klinton Bicknell. 2021. Local word statistics affect reading times independently of surprisal. *arXiv preprint arXiv:2103.04469*.

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive science*, 29(2):261–290.

Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. 2022. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

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 Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–86.

Jacob Louis Hoover, Morgan Sonderegger, Steven T Piantadosi, and Timothy J O'Donnell. 2023. The plausibility of sampling as an algorithmic theory of sentence processing. *Open Mind*, 7:350–391.

Ryu Soo Hyun et al. 2022. Using transformer language model to integrate surprisal, entropy, and working memory retrieval accounts of sentence processing. In *35th Annual Conference on Human Sentence Processing*.

Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context limitations make neural language models more human-like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10421–10436.

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

Richard L Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29:375–419.

Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2):203–208.

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.

Byung-Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from gpt-2 attention patterns predict reading times over and above gpt-2 surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9324–9334.

Byung-Doh Oh and William Schuler. 2023. 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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nathan E Rasmussen and William Schuler. 2018. Left-corner parsing with distributed associative memory produces surprisal and locality effects. *Cognitive science*, 42:1009–1042.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Cory Shain. 2023. Word frequency and predictability dissociate in naturalistic reading.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Philip Levy. 2022. Large-scale evidence for logarithmic effects of word predictability on reading time.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior research methods*, pages 1–21.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8705–8720.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ethan Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

A Sentence length

In our main analysis, we found that the surprisal estimates from our locally biased model were associated with a better fit to human psychometric data in the MECO, Provo, and Brown datasets, while our locally-biased model performed on par with the original GPT-2 model in the UCL datasets, and worse than its counterpart in the NatStor corpus. We noted that the relative performance of the locally biased model was particularly improved in datasets with long average sentence length (MECO, Provo, and Brown). Indeed, the Pearson correlation between mean sentence length (i.e., average number of word per sentence) and ΔAIC is $r = -0.77$ ($p = 0.07$). While the number of observations ($N = 6$) and the absence of statistical significance does not license strong conclusions on this regard, we remark that this trend is compatible with the idea of a lossy representation of the context, where memory constraints are more pronounced in processing longer text sequences.