Context Limitations Make Neural Language Models More Human-Like

Tatsuki Kuribayashi[1,2]   **Yohei Oseki**[3,4]   **Ana Brassard**[1,4]   **Kentaro Inui**[1,4]
[1]Tohoku University   [2]Langsmith Inc.   [3]University of Tokyo   [4]RIKEN
{kuribayashi, inui}@tohoku.ac.jp
oseki@g.ecc.u-tokyo.ac.jp   ana.brassard@riken.jp

Abstract

Language models (LMs) have been used in cognitive modeling as well as engineering studies—they compute information-theoretic complexity metrics that simulate humans' cognitive load during reading. This study highlights a limitation of modern neural LMs as the model of choice for this purpose: there is a discrepancy between their context access capacities and that of humans. Our results showed that *constraining* the LMs' context access improved their simulation of human reading behavior. We also showed that LM-human gaps in context access were associated with specific syntactic constructions; incorporating syntactic biases into LMs' context access might enhance their cognitive plausibility.[1]

1 Introduction

In computational psycholinguistics, human reading behavior has been compared with various complexity metrics to understand human sentence processing (Crocker, 2007). Having historically started from simple measures such as word length, surprisal $(-\log p(\text{word}|\text{context}))$ computed by language models (LMs) has become a common choice (Levy, 2008; Smith and Levy, 2013). On top of this, the next question arises—which model implementation and/or algorithm can compute surprisal that successfully simulates human behavior? In this line of research, modern neural LMs such as Transformer (Vaswani et al., 2017) have been analyzed with respect to their cognitive plausibility (Wilcox et al., 2020; Merkx and Frank, 2021; Kuribayashi et al., 2021).

Despite their use in cognitive modeling, such modern LM architectures (e.g., self-attention) are, arguably, an unnatural choice when it comes to human cognitive constraints; modern LM architectures assume powerful, parallel access to a vast
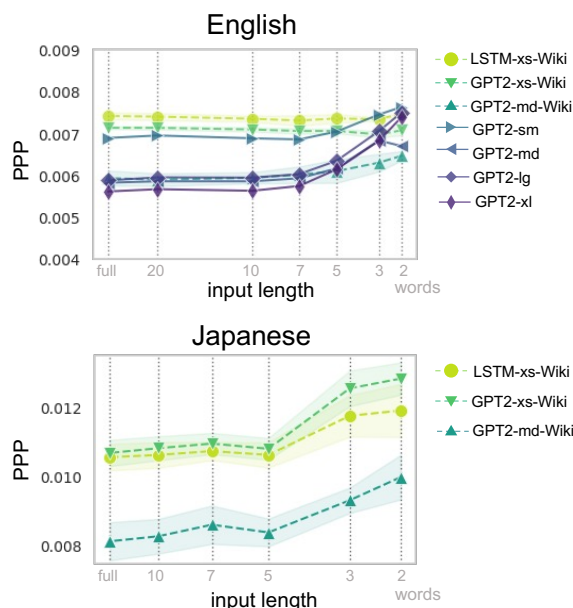

Figure 1: Relationship between psychometric predictive power (PPP) of language models (LMs) and their context access constraints. LMs with less context access better simulate human reading behavior (higher PPP). The marker color/shape indicates LM settings; colored areas present one standard deviation of PPP.

number of context tokens, while humans might have limited and selective context access (Hawkins, 1994; Gibson, 1998, 2000; Lewis et al., 2006). Searching for a computational model that better simulates human sentence processing than previously examined ones, we hypothesized that introducing such context limitations can improve LMs' estimation of human cognitive load.

Specifically, as a starting point, we applied an n-gram-ification trick to neural LMs mimicking loading for long context access (*locality effects*) and compared their surprisal with human reading behavior data. Despite the simple context limitation design, our experiments with 280 settings (40 LM settings×7 noise patterns) showed that the advantage of a shorter context was consistent among

---

Neural model architecture	Data size	Training steps	Tokenization	Syntactic supervision	Context length
Goodkind and Bicknell (2018); Aurnhammer and Frank (2019); Wilcox et al. (2020); Hao et al. (2020); Merkx and Frank (2021); Oh et al. (2021); Kuribayashi et al. (2021)	Wilcox et al. (2020); Kuribayashi et al. (2021)	Kuribayashi et al. (2021)	Wilcox et al. (2020); Oh et al. (2021)	Hale et al. (2018); Yoshida et al. (2021)	**This work**

Table 1: Related studies exploring psychometric predictive power of neural models while *separately* controlling a specific factor of their configuration.

LMs and typologically different languages (Figure 1). This showed that *constraining* the modern LM's context access is key to increasing their similarity to the model of human reading.

Furthermore, expecting that humans' context limitations might be more complex than simple distance-based erasure, we conducted exploratory analysis of in which constructions longer/shorter contexts were beneficial. We found that the context limitation (dis)advantages were allocated in specific syntactic constructions, suggesting that, to build more cognitively plausible LMs, adding syntactic biases in their context access could be beneficial. From a psycholinguistic view, our results empirically highlight the memory account of human sentence processing during naturalistic reading (Futrell et al., 2020a).

2 Background

2.1 Human sentence processing

Humans incrementally process text and exhibit different processing costs (e.g., reading times) for different tokens. Psycholinguistic theories on such processing costs are divided between expectation-based and memory-based perspectives.

Expectation-based theories claim that humans predict upcoming words during incremental sentence processing (Clark, 2013). Recent studies have extensively analyzed this expectation-based aspect by comparing surprisal, $-\log p(\text{word}|\text{context})$, to human reading behavior (Hale, 2001; Levy, 2008; Wilcox et al., 2020).

On the other hand, *memory-based* theories have asserted that human sentence processing is constrained by a limited cognitive resource (Gibson, 2000; Lewis and Vasishth, 2005; Lewis et al., 2006). Cross-linguistic studies have reported that different languages incur different memory decay during reading (Konieczny, 2000; Vasishth et al., 2010; Husain et al., 2014; Frank et al., 2016). Notably, memory efficiency is also considered in exploring the design (e.g., word order) and evolution of nat-

ural language (Greenberg, 1963; Chomsky, 2005; Gibson et al., 2019; Hahn et al., 2020a,b).

Recently, Futrell and Levy (2017) and Futrell et al. (2020a) have proposed integrating the two theories through the concept of *lossy-context surprisal*—next-word probabilities calculated with *noisy* context should better predict human reading behavior than with complete context. These studies have focused on its theoretical aspects and explaining a specific phenomenon (e.g., verb forgetting); on top of this, our study demonstrates the theory's broad benefit in modeling naturalistic reading data.

Notably, such a simulation of human cognitive load also contributes to achieving text readability assessment (Ambati et al., 2016). Furthermore, human-like agents are necessary in *in silico* simulation studies on language evolution (Galke et al., 2022; Rita et al., 2020; Ueda and Washio, 2021).

2.2 Cognitive plausibility of LMs

Surprisal from certain LMs could predict human reading behavior well; thus, what *type* of LM does better simulate human reading behavior? LM-based analyses have typically explored inductive biases, such as LM architecture (Table 1). We focus on context limitation as an alternative factor.

Studies comparing the cognitive plausibility of LM architectures also addressed, albeit implicitly, context access abilities (Aurnhammer and Frank, 2019; Merkx and Frank, 2021). For example, simple recurrent neural networks assume relatively weak context access, whereas Transformer LMs (Vaswani et al., 2017) assume stronger access than those (Michaelov et al., 2021; Merkx and Frank, 2021). In addition, studies contrasted count-based n-gram LMs and neural LMs (Wilcox et al., 2020; Hao et al., 2020; Goodkind and Bicknell, 2018); however, (i) count-based versus neural-based estimation and (ii) partial versus full context access were not distinguished. By contrast, we fixed the architectures and investigated the exact effect of context access with input deletion.

3 Methods

We investigate how human-like neural LMs become with more or less context at their input. Specifically, we measured the psychometric predictive power (PPP) of lossy-context LM surprisal for gaze duration modeling. In the following sections, we describe each measure in detail.

3.1 Psychometric predictive power

In this study, the cognitive plausibility of a model θ is measured via the similarity between its surprisal and human gaze duration across words based on surprisal theory (Smith and Levy, 2013; Levy, 2008). Here, the surprisal of a word computed by a model θ, $-\log p_\theta(\text{word}|\text{context})$, is compared with the corresponding word's gaze duration.

Specifically, we measured the **psychometric predictive power (PPP)** of surprisal values by fitting two nested linear mixed-effects regression models that predict gaze duration, one with surprisal features and the other without. Here, the per-token difference in their log-likelihoods (ΔLogLik; LogLik with surprisal minus LogLik without surprisal) denotes PPP, following Goodkind and Bicknell (2018). The larger the PPP (ΔLogLik), the more useful the surprisal for modeling gaze duration, i.e., the model computes surprisal well correlating with human behavior. See Appendix A for detailed features used in regression modeling.

3.2 Lossy-context surprisal

Instead of the full-context surprisal, we investigate the PPP of surprisal conditioned by limited context $-\log p_\theta(\text{word}|\text{lossy\_context})$ to explore the cognitive plausibility of context-limited LMs (Futrell et al., 2020a). The **lossy-context surprisal** of the symbol w_i given its preceding context $c_{<i} = [w_0, \cdots, w_{i-1}]$ is defined as follows:

$$I_{\text{lossy}}(w_i, c_{<i}, f) \\ = -\log p_\theta(w_i|\text{<s>} \circ f([w_0, \cdots, w_{i-1}])) \ , \quad (1)$$

where θ denotes left-to-right LMs, <s> denotes the beginning of a sequence, $\circ$ is a concatenation function, and f represents a noise function. The noise function controls the LMs' access to contextual information by deleting the input of LMs with a particular pattern. For example, if f is leaving only the last two symbols, I_{lossy} corresponds to surprisal from 3-gram LMs, and if f is an identity function, I_{lossy} corresponds to unmodified surprisal.

Data	sentLen	contextLen	wordLen
DC	17.8±11.7	13.0±10.1	1.3±0.7
BE	7.9±5.3	6.0±4.6	3.4±2.3

Table 2: Statistics of all the sentences in each corpus. The values present mean±standard deviation; sentLen denotes the number of words in a sentence, contextLen denotes the number of preceding words within the same sentence for each word, and wordLen denotes the number of subwords in each word.

Gaze duration is typically annotated in larger spans such as words, while LMs' input is at the smaller levels (i.e., subwords). The lossy-context surprisal of a span $s = [w_l, w_{l+1}, \cdots, w_m]$ ($0 \leqq l < m$) was calculated as the cumulative surprisal of its constituent subwords:

$$I_{\text{lossy}}(s, c_{<l}, f) = \sum_{j=l}^{m} I_{\text{lossy}}(w_j, c_{<j}, f) \ . \quad (2)$$

N**-gram surprisal.** As a starting point, based on the assumption about human working memory that distant context is hard to access (Lewis et al., 2006), we explored surprisal given by LMs conditioned on $n-1$ preceding words (not subwords); henceforth, this surprisal is referred to as n-gram surprisal (a special case of lossy-context surprisal). In Appendix B, we also explored a probabilistic version of the noise inspired by Futrell et al. (2020a), yielding consistent conclusions with our experiments using n-gram surprisal.

3.3 Gaze duration

Gaze duration data were modeled by lossy-context surprisal. To explore the cross-linguistic consistency of our results, we used two typologically different languages, English and Japanese; their difference is introduced in the later paragraph.

Data. For English, we used the Dundee Corpus (DC) (Kennedy et al., 2003). As its Japanese counterpart, we used BCCWJ-EyeTrack (BE) (Asahara et al., 2016). In both corpora, first-pass gaze duration information was used. The average sentence/context lengths are shown in Table 2. Note that while the English gaze duration annotation is typically attached to space-separated words, Japanese gaze duration annotation is attached to each phrasal unit (*bunsetsu*; henceforth, "word"); Japanese "words" contain more subwords than English words. Following Goodkind and Bicknell

(2018), we excluded outliers such as words with special characters (details in Appendix C). We used 212,649 data points from DC and 9,217 from BE.

Cross-linguistic analysis. English and Japanese sentence structures differ in their branching directions; while English word order (SVO) has mixed directionalities of head-initial and head-final dependencies, Japanese word order (SOV) strongly prefers head-final, left-branching constructions. The dependency structures of the sentence "*the dog wagging its tail eats fish on the desk.*" in English and Japanese are contrasted below:[2]

(1) The dog wagging its tail ate fish on the desk.

(2) 尻尾を 振る 犬が 机の 上で 魚を 食べた。
     tail wagging dog on desk fish ate
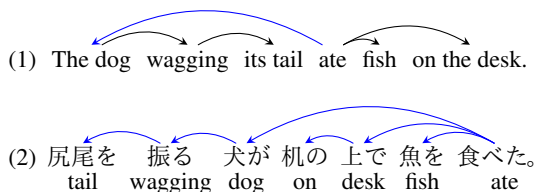
Such an asymmetry of structure has been reported to incur different memory biases for sentence processing (Konieczny, 2000; Nakatani and Gibson, 2008; Vasishth et al., 2010; Futrell et al., 2020a). Thus, we included typologically different languages in our experiments.

4 Language models

We used two types of neural LMs for lossy-context (n-gram) surprisal computation: (i) Wiki-LMs and (ii) pretrained OpenAI GPT-2s (Radford et al., 2019). Their hyperparameters are shown in Appendix D. Notably, using *neural* LMs makes the comparison of long-context and short-context LMs computationally tractable.[3]

4.1 Wiki-LMs

Model settings. We used three variants of unidirectional neural LM architectures: LSTM-xs-Wiki (27M parameters) (Hochreiter and Schmidhuber, 1997), GPT2-xs-Wiki (29M), and GPT2-md-Wiki (335M) (Vaswani et al., 2017). We trained each LM with three different random seeds using the `Fairseq` toolkit (Ott et al., 2019).

In both English and Japanese settings, the input is split into subwords with byte-pair encoding (Sen-

---

... \<b> _was _also _the _first _hotel _in _Westchester _County . \<b> _on _4 _March _1990 _with _a _concert _performed _by _Ell a _Fitzgerald _at _the _Royal _Albert _Hal \<b> _the _Har row \<b> _On _the _night _of _the _31 _May _/ _1 _June _1941 _he \<b> ...

Table 3: An example of the modified training data, where sub-sequences (with the same color) sampled from the original corpus were randomly patched. The special token (\<b>) indicates the break of contextual dependence between before and after.

---

nrich et al., 2016).[4] Specifically, for the Japanese data, we adopted two-stage segmentation to ensure that multiple subwords compose a Japanese word defined in a commonly used corpus (e.g., BE).[5] That is, text was segmented in advance into morphemes (Maekawa et al., 2014), and then a subword tokenizer was applied to the morpheme-separated texts. Details are in Appendix D.

Training data. For English, the training data were approximately 4M sentences from the WikiText-103 dataset (Stephen et al., 2016), and for Japanese, the data were 4M sentences from Wikipedia and news articles (approximately 0.5GB data size in both English and Japanese). The sentence order was shuffled, and duplicated sentences were excluded.

Mitigating training-inference mismatches. During n-gram surprisal computation, LMs must predict the upcoming words with limited context from the middle of a sentence, while such a prediction is rarely enforced during ordinal document-level training. Such a training-inference mismatch could lead to confusion on whether our results stem from the LM-human gap or biases from the training/inference mismatch.

To handle such a potential mismatch, we modified the LM training data to make the language modeling task more like n-gram one. Specifically, we randomly split original sentences into smaller chunks of various lengths and randomly patched them by inserting a special token \<b> in between the chunks (see Appendix E for the detailed process). Table 3 shows an example. In this modi-

---

[2] For simplicity, some functional words (e.g., "the," "を") are merged into a single node.

[3] For example, if we use *count-based* LMs (Heafield et al., 2013), even a single 5-gram Japanese LM took 27GB in model size.

[4] Note that there is some debate on the cognitive plausibility of subwords (Oh et al., 2021; Anonymous, 2022); we consider this issue to be out of the scope of this study.

[5] This procedure is common practice in Japanese NLP. See https://github.com/himkt/awesome-bert-japanese. We used Mecab (Kudo, 2006) with a unidic dictionary (https://unidic.ninjal.ac.jp/) for morphological analysis.

Lang.	Model	Input length							Δ
		full	20	10	7	5	3	2	
En	GPT2-xl	5.6	5.7	5.6	5.8	6.2	6.8	**7.4**†	1.8
	GPT2-lg	5.9	6.0	6.0	6.0	6.4	7.1	**7.5**†	1.6
	GPT2-md	5.8	5.9	5.9	5.9	6.2	**6.8**	6.7†	0.9
	GPT2-sm	6.9	7.0	6.9	6.9	7.1	7.5	**7.6**†	0.7
	GPT2-md-Wiki	5.9	5.9	5.9	6.0	6.1	6.3	**6.5**†	0.6
	GPT2-xs-Wiki	**7.1**	**7.1**	**7.1**	**7.1**	**7.1**	7.0	**7.1**	0.0
	LSTM-xs-Wiki	7.4	7.4	7.4	7.3	7.4	7.4	**7.5**	0.1
Ja	GPT2-md-Wiki	8.1	8.1	8.3	8.6	8.4	9.3	**10.0**†	1.9
	GPT2-xs-Wiki	10.7	10.7	10.8	11.0	10.8	12.5	**12.8**†	2.1
	LSTM-xs-Wiki	10.6	10.6	10.6	10.7	10.6	11.8	**11.9**†	1.3

Table 4: Average PPP of n-gram surprisal; for example, the input length of 2 corresponds to the PPP of surprisal computed by the neural LMs that take only the 2-gram context as input. For readability, values are multiplied by 1000. The 2-gram PPP with † is significantly higher than its corresponding full-context PPP. The Δ column shows the PPP gain from the full context to 2-gram context surprisal in each LM setting.

fied corpus, LMs must predict upcoming words by severely limited usable context especially in the data points immediately after the special tokens. When computing n-gram surprisal, the \<b\> token is set instead of \<s\> in Eq. 1. Note that this modification does not change the total corpus size.

We trained the Wiki-LMs using this modified data. In Section 5.1, we ablated the effect of this training modification and showed that such careful training makes the short-context advantage clearer.

4.2 Pretrained GPT-2s

To investigate large-scale LMs typically developed in NLP, we additionally used four variants of pretrained English OpenAI GPT-2s (Radford et al., 2019): GPT2-sm (117M params.), GPT2-md (345M), GPT2-lg (774M), and GPT2-xl (1558M). The input was split into subwords by their pretrained tokenizer with a vocabulary size of 50,257. The training data were 40GB of web texts. The potential training-inference mismatch is not handled in the GPT-2 experiments due to the high retraining cost; this point is partially addressed in Section 5.1. Note that we did not use Japanese versions of pretrained GPT-2s since available models[6] have a tokenizer that is inconsistent with the BE annotation; 16.4% of word boundaries in the BE were not separated by their pretrained tokenizer.

[6] https://huggingface.co/rinna/japanese-gpt2-small

5 Experiments

Our experiments demonstrate how limiting context access improved the PPP in LMs, i.e., their surprisal becomes a more effective predictor for human gaze duration (Section 5.1). As described in Section 3.2, we applied distance-based noise to the input (i.e., computing n-gram surprisal). A potential training-inference mismatch bias is handled (Section 5.2). Furthermore, we explored the connection to existing studies (Section 7.2).

Settings. We measured the PPP of 40 LMs ({3 Wiki-LMs}×{3 seeds}×{2 languages}×{2 training settings}+{4 OpenAI GPT-2s}) with seven noise patterns. Specifically, we explored the {2, 3, 5, 7, 10, 20}-grams and full settings for each LM, where full refers to using the entire sentence (w_0 to w_{i-1}) as context. Note that only intrasentential context is used as the main focus of this study is sentence-level syntactic processing.

5.1 PPP and input length

Shorter context improved or did not decrease human likeness. The PPP of n-gram surprisal in relation to input length n is shown in Figure 1 and Table 4. The English results show that using a shorter context improved (OpenAI GPT-2s) or did not hurt (Wiki-LMs) their human likeness. Notably, we also conducted experiments using probabilistic versions of the noise in Appendix B, yielding consistent results.

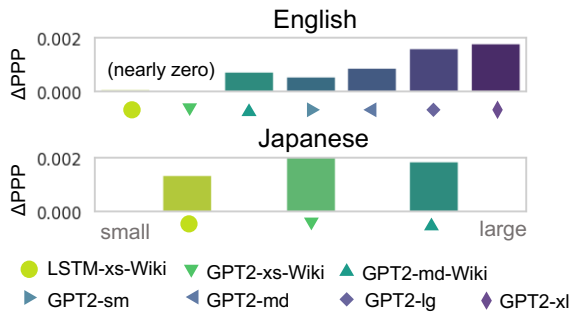Note that the tininess of the values in Figure 1 and Table 4 (e.g., 0.0074 v.s. 0.0056 in GPT2-

Figure 2: Increase in PPP (from the full-gram to 2-gram settings) in each model type (ordered by their parameter size). The bar colors correspond to those in Figure 1.

xl) does not imply that the difference is valueless, but this is just because the score is divided by the number of data points (e.g., 212,649 in the Dundee corpus) to facilitate inter-corpora comparison. As a statistical test, we compared the by-token squared residual errors from 2-gram models with those from full-context models using paired permutation tests (p=0.05). The short context, 2-gram models had significantly smaller fitting errors than the full context models ($p < 0.001$) in using relatively large LMs (GPT2-md-Wiki, GPT2-sm, GPT2-md, GPT2-lg, and GPT2-xl); smaller LMs (LSTM-xs-Wiki, and GPT2-xs-Wiki) have no significant differences ($p \sim 0.4$).

Notably, we also observed that **larger** GPT-2s have **less** human-like behavior in the full setting (right-most column in Table 4). This trend was weakened by introducing our context limitation.

Cross-linguistic consistency. Figure 1 and Table 4 also show the cross-linguistic generality of the short context advantage. The short context was more clearly favored in Japanese than in English. Using the same method as the English experiments, we performed the significance tests; 2-gram models exhibited smaller fitting errors ($p \sim 0.001$) in all the Japanese LM settings. The language-dependent differences are further investigated in Section 6.

The larger the LM, the greater the increase in PPP when limiting context access. Figure 2 shows the PPP increase in each LM class by context limitation (PPP at 2-gram minus PPP at full-gram). The bars were ordered by the model parameter size (small → large). We found a clear trend that larger LMs become human-like by a larger margin because of context limitation; larger full-context LMs deviate more from human-like context access.

We statistically tested whether the gain by context limitation (full-context v.s. bigram) was larger in the largest LMs (GPT2-md in Japanese and GPT2-xl in English) than in the smallest LMs (LSTM-xs). Specifically, we compared the by-token decrease in squared residual errors; the large model exhibited a larger error decrease than the small model ($p = 0.024 < 0.05$ in Japanese, and $p < 0.001$ in English). In addition, the rank correlation between model size and PPP gain by context limitation was 0.50 in Japanese and 0.96 in English.

General effectiveness of surprisal. Note that, in all the LMs, the PPP scores (equivalent to ΔlogLik) were significantly higher than 0 with the chi-square test ($p < 10^{-31}$ even in the worst case); surprisal was an effective factor as existing studies reported. On top of this, we newly showed that their effect size differs due to the context limitation levels.

5.2 Does the potential training-inference mismatch bias our results?

Vanilla LMs slightly underestimate the short-context advantage. We additionally trained Wiki-LMs (LSTM-xs-Wiki, GPT2-xs-Wiki, and GPT2-sm-Wiki) without the data modification handling the training-inference gap (Section 4.1) (henceforth; vanilla LMs). Figure 3 shows the results of the models with and without the training modification. The vanilla LMs slightly underestimated the short-context advantage; the PPP of 2-gram surprisal improved when we adopted the modified training. That is, mitigating the train-inference gap made clearer the trend that *context limitation increases PPP*. Carefully training n-gram neural LMs could be a way to create more human-like computational model.

6 Analyses

In our experiments, we merely deleted distant contexts regardless of linguistic factors. However, this design is somewhat counter-intuitive in the sense that humans are assumed to completely forget the distant context during reading. To gain insights into a more sophisticated noise design to fill the LM-human gap, we observed *in which constructions longer/shorter contexts improved simulation of human gaze duration*.

6.1 Settings

Quantifying long context effect. To quantify the long context advantage for each data point, we
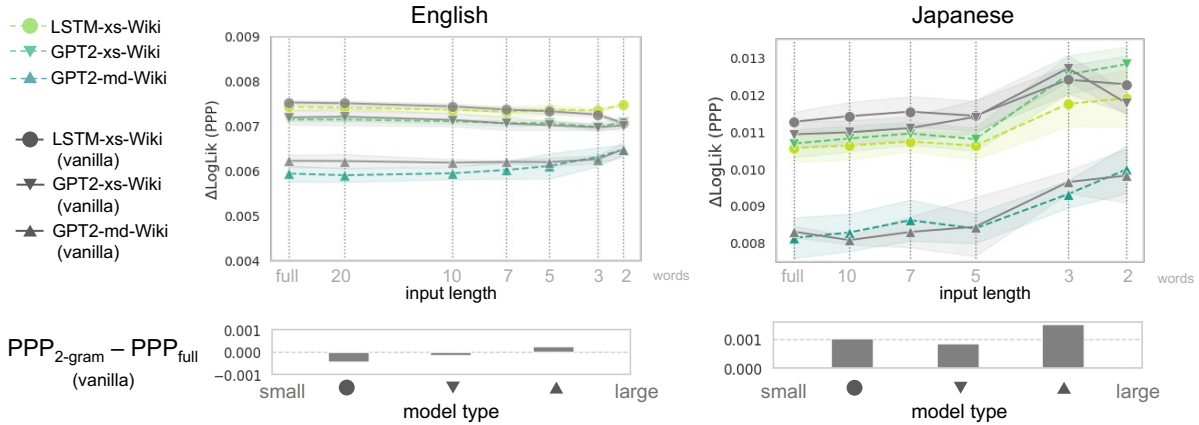
Figure 3: Reproduction of Figures 1 and 2 using LMs without training setting modifications (Section 5.1). The results from Wiki-LMs with the modification (colored) and without the modification (gray) are overlayed. In the line charts, the X-axis indicates input length, and the Y-axis indicates PPP. The bottom bar charts show the increase in PPP (from full-gram to 2-gram setting) of the modified LMs.

compared the squared residual (fitting) errors of the regression models we used to compute PPP in Section 5. Note that the larger the squared residual error is, the worse the model fit with the target variable (gaze duration).

Specifically, we contrasted the two regression models with different context access: (i) the model with 2-gram surprisal, and (ii) the model with full context surprisal. For each data point d, we measured the effectiveness of long context (ELC) in explaining gaze duration. Specifically, the difference between the squared residual errors by the regression models with 2-gram surprisal $r_2(d)$ and full surprisal $r_{\text{full}}(d)$ was computed:

$$\text{ELC}(d) = r_2(d) - r_{\text{full}}(d) \ . \tag{3}$$

Here, a **high** ELC value indicates that reading times on d were **better** simulated with **long** context ($r_{\text{full}}(d) \downarrow$); **worse** simulated with **short** context ($r_2(d) \uparrow$). The aim of this section is to find the data points with a high ELC value. In the following analyses, we used all the models from Section 5.1, and ELC scores for each data point were averaged across all the LMs.

Dependency structure. Human context access has typically been discussed with respect to syntactic structure (Gibson, 1998; Demberg and Keller, 2008); we first explored the interactions between context limitation advantage and syntactic dependencies. We analyzed two syntactic factors: (i) **dependency locality** and (ii) **dependency type**, where the dependency locality of a token denotes how far its syntactically related preceding items

(i.e., with a direct dependency) are placed on average. An example is as follows:

(3) The boy over there had a cap.
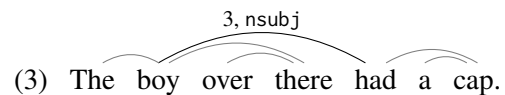
Here, the dependency locality of "had" is three; note that the dependency direction was disregarded.

In the following analyses, we only used data points with potential long context access, i.e., those in the latter part of a sentence.[7] After this filtering, the average dependency locality score was 2.5 and 2.6 in the DC and BE, respectively. Manual linguistic annotations were used in our analyses (Barrett et al., 2015; Omura and Asahara, 2018).

6.2 Results

Dependency locality. We first grouped the data points by their dependency locality and calculated the average ELC scores for each group. Figure 4a shows the results. Surprisingly, in the English data, there is no advantage in considering the long context for tokens with long dependencies. By contrast, in the Japanese data, long context access contributed to simulating reading time for tokens with a moderate (two or three) dependency length, but not for long dependency locality. These imply that **the solution is more complex than simply using long context for words with long dependency.**

---

[7]13th/7th or later words of a sentence in the DC/BE data (20,554 words/4,051 words) were used, based on the median of the word position in sentences (12 and 6 in the DC and BE corpus, respectively).

Dependency type. Does long context matter in specific syntactic constructions? We categorized the data points by their dependency type to their preceding syntactically related items and calculated the averaged ELC score for each group.[8]

Figure 4b shows that different dependency types are associated with different ELC scores. For example, the discourse type in English have relatively larger ELC scores; long context input is necessary to simulate its gaze duration. Figure 4b also suggests that such context-favoring (with high ELC) dependency types are different between English and Japanese. These findings imply that **the LM-human context access gap occurred in specific syntactic constructions in each language.**

One-way ANOVA revealed that the average ELC scores for each dependency type significantly varied ($p = 0.029 < 0.05$ in English, $p = 0.038 < 0.05$ in Japanese), suggesting that the variation of the ELC score is related to certain constructions. More specifically, we compared the ELC distribution between the categories with the highest and lowest average ELC scores (discourse vs. cop in English, and advcl vs. obl) using an unpaired t-test. The test exhibited a significant difference ($p = 0.012 < 0.05$ in English, and $p = 0.019 < 0.05$ in Japanese). Note that if the test is repeated for other dependency-length/type pairs, multiple comparison problems should occur; some counteractions, such as Bonferroni correction should be applied, and a more conservative conclusion can be reached.

7 Discussion

### 7.1 Interpretations of the main results

We observed that simply deleting distant context improved LMs' PPP—*as context decreased, LMs became more human-like*. We finally discuss several potential interpretations of our results.

One interpretation is that our results supported **the dominance of short context access in human sentence processing**. In this sense, our findings emphasized that explicitly incorporating principles from the memory-based account of human sentence processing is still necessary for simulating human sentence processing despite the success of modern LMs in cognitive modeling (Wilcox et al., 2020; Schrimpf et al., 2020). Notably, there are several other theories on human working memory; sparse
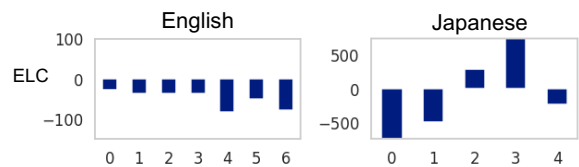
(a) Relationship between dependency locality and the ELC scores. The X-axis corresponds to dependency locality (e.g., the group "3" denotes the data points with the locality score of three). The Y-axis denotes the ELC score for each group.
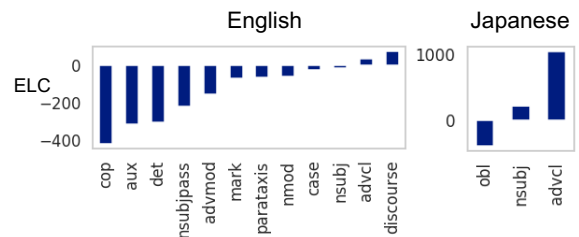
(b) Relationship between dependency type and the ELC scores. The X-axis corresponds to the dependency type. The Y-axis denotes the average ELC score for each group. Dependency types for which there are more than 100 long dependencies (locality>4) were included.

Figure 4: Relationships between syntactic factors and the ELC scores.

allocation of elements incurring memory load (Gibson, 2000), hierarchical memory operations (van Schijndel et al., 2013), and cue-based memory retrieval (Lewis and Vasishth, 2005). Incorporating these perspectives into context-limited LMs could be an interesting future direction.

Another possibility is that **identifying the cause of the LM-human gap as context limitations is over-claiming**; our study alone did not rule out some potentially confounding factors. For example, increasing the softmax temperature when LMs compute the next-word distribution may induce a similar effect to our context limitation with respect to that both modifications make LMs less confident about the upcoming word (if temperature matters, the linear relationship between surprisal and cognitive load may be doubted first). Further exploring such factors will be an important investigation.

There is also a possibility that **the eye movement data only reflected local, shallow aspects in human sentence processing**. Similarly, Gauthier and Levy (2019) obtained somewhat counterintuitive results implying that word order is not important information in sentence processing—bag-of-words (i.e., not word-order-aware) models fit fMRI data surprisingly well. They concluded that their results may stem from shortcomings of the

---

[8]Here, we only focus on the dependencies with more than four distances to explore potential long context access.

measurement method along with the possibility of humans' good-enough processing. Exploring the advantage of context limitation in various types of reading behavior data and/or using other text materials (e.g., including more complex constructions) is also a line of future research.

Is the 2-gram advantage counter-intuitive? If the dominance of short context access in human reading is accepted once, some readers might be confused that the 2-gram context access sounds too severe. Again, as a strong word frequency effect does not deny context-dependent processing, our results also did not decline long context access and did not claim that the human language processing model is 2-gram LMs. Exploring the *interactions* of short- and long-context effects should be an interesting investigation.

Nevertheless, such severe memory limitations during reading might be consistent with the memory-based explanation for the linguistic universals in sentence structures such as the preferences toward consistent head directions, specific word order (e.g., short-before-long order), and projective structures (Futrell et al., 2020b). Such phenomena are typically explained by the humans' preferences toward short dependencies; here, those are sometimes a matter of severe choices, such as the preference for an average dependency length of 1 over 2 (Intuitively, Example (4) is preferred over Examples (5) and (6)):

(4) A B C D E   avg. dep len.=1.25

(5) A B C D E   avg. dep len.=2.25

(6) A B C D E   avg. dep len.=1.50
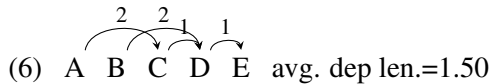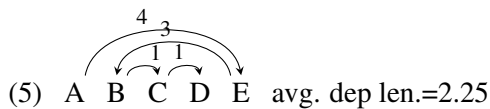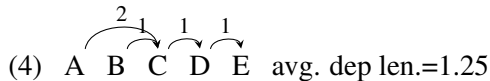
If one reasons these principles to the constraints of humans' cognitive resources, perhaps it makes sense that humans conduct syntactic processing with such a severe working memory that the immediately preceding word/phrase highly explains the cognitive load to the upcoming word.

7.2 PPP and next-word prediction accuracy

Lastly, we discuss the connection to reports on cognitive modeling with LMs—better next-word prediction ability of LMs indicates their better PPP (Fossum and Levy, 2012; Goodkind and Bick-
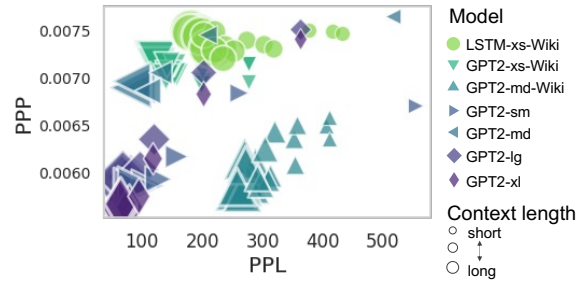

Figure 5: Relationship between PPP and perplexity (PPL) drawn using the English LMs targeted in Section 5.1. Each point corresponds to each configuration of the n-gram surprisal computation; marker color and shape present the LM architectures, and larger markers correspond to longer context access.

nell, 2018; Wilcox et al., 2020).[9] Our results in Section 5.1 might be conflicting with them; LMs with relatively **worse** prediction accuracy (less context access) exhibit **better** PPP. See Appendix F for next-word prediction accuracy of LMs.

Results. In fact, there is no clear relationship between PPP and next-word prediction accuracy (perplexity; PPL) of the LMs used in Section 5.1 (Figure 5; Appendix G exhibits Japanese results). The results show that LMs with nearly the same next-word prediction accuracy could show different PPP values. Furthermore, Pearson's correlation between PPP and PPL was even positive ($r = 0.15$). These observations corroborated the conclusion that the **PPL alone is not a good indicator of PPP**; different means of controlling PPL (e.g., context length vs. other factors existing studies focused on) could show different PPP-PPL relationships.

8 Conclusions

There has been little investigation of the cognitive plausibility of context-limited modern LMs. Our experiments using the input-controlled neural LMs have shown that short context LMs simulate human reading behavior surprisingly well, emphasizing the LM-human gap in context access. Further analysis has shown that the gap could be associated with specific syntactic constructions; injecting syntactic bias into LMs' context access could be one way to make LMs more human-like. This study has also asserted that the use of a modern LM popular in NLP as-is is not always a natural choice in cognitive modeling.

---

[9]There are also some counter-arguments (Hao et al., 2020; Oh et al., 2021; Kuribayashi et al., 2021; Anonymous, 2022).

Limitations

As discussed in Section 7, this study alone could not comprehensively explain the cause of the LM-human discrepancies. Nevertheless, our observation itself could advance the step toward understanding the relationship between human sentence processing and computational models typically developed in NLP, which is a central theme in the long history of artificial intelligence and the cognitive science of language.

This study was scientifically motivated to understand humans and language; this could sound like less impact on engineering-oriented efforts (e.g., solving real-world problems accurately). However, simulating human cognitive load during reading is directly associated with automatic text readability assessment. In addition, our study implies that human sentence processing could be performed with more efficient context access than modern LMs. This encourages the development of language processing models with increased efficiency; this is related to the sustainability issues, such as the environmental impact of creating gigantic NLP models.

Ethical considerations

This study explored the relationship between the LM-computed complexity measures and human reading behaviors. Human subjects' privacy information in the eye-tracking data was anonymized. We did not find any other ethical concerns; as a somewhat minor point, the LMs used in our experiments might be biased by the data we used (i.e., Wikipedia and Web data), although these follow the commonly used settings in the NLP research.

Acknowledgements

References

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental ccg parser. In *Proceedings of NAACL*, pages 1051–1057.

Anonymous. 2022. Why does surprisal from smaller GPT-2 models provide better fit to human reading times? *OpenReview*.

Masayuki Asahara, Hajime Ono, and Edson T Miyamoto. 2016. Reading-Time Annotations for "Balanced Corpus of Contemporary Written Japanese". In *Proceedings of COLING*, pages 684–694.

C Aurnhammer and S L Frank. 2019. Comparing Gated and Simple Recurrent Neural Network Architectures as Models of Human Sentence Processing. In *Proceedings of CogSci*, pages 112–118.

Maria Barrett, Željko Agi, and Anders Søgaard. 2015. The Dundee Treebank. In *Fourteenth International Workshop on Treebanks and Linguistic Theories*, pages 242–248.

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

Noam Chomsky. 2005. Three factors in language design. *Linguist. Inq.*, 36(1):1–22.

Andy Clark. 2013. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.*, 36(3):181–204.

Matthew W Crocker. 2007. Computational psycholinguistics. *The Handbook of Computational Linguistics and Natural Language Processing*.

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

Victoria Fossum and Roger Levy. 2012. Sequential vs. Hierarchical Syntactic Models of Human Incremental Sentence Processing. In *Proceedings of CMCL*, pages 61–69.

Stefan L Frank, Thijs Trompenaars, and Shravan Vasishth. 2016. Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cogn. Sci.*, 40(3):554–578.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020a. Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Journal of Cognitive Science*.

Richard Futrell and Roger Levy. 2017. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of EACL*, pages 688–698.

Richard Futrell, Roger P. Levy, and Edward Gibson. 2020b. Dependency locality as an explanatory principle for word order. *Journal of Language*.

Lukas Galke, Yoav Ram, and Limor Raviv. 2022. Emergent communication for understanding human language evolution: What's missing? *Emergent Communication Workshop at ICLR 2022*.

Jon Gauthier and Roger Levy. 2019. Linking artificial and human neural representations of language. In *Proceedings of EMNLP-IJCNLP*, pages 529–539.

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

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

Edward Gibson, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends Cogn. Sci.*, 23(5):389–407.

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 CMCL*, pages 10–18.

Joseph H Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. http://tushik.org/wp-content/uploads/GRE-order.pdf. Accessed: 2022-10-18.

Michael Hahn, Judith Degen, and Richard Futrell. 2020a. Modeling word and morpheme order in natural language as an efficient tradeoff of memory and surprisal. *Psychological Review*.

Michael Hahn, Dan Jurafsky, and Richard Futrell. 2020b. Universals of word order reflect optimization of grammars for efficient communication. *Proc. Natl. Acad. Sci. U. S. A.*, 117(5):2347–2353.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of NAACL*, pages 159–166.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan R. Brennan. 2018. Finding Syntax in Human Encephalography with Beam Search. In *Proceedings of ACL*, 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 CMCL*, pages 75–86, Online.

John A Hawkins. 1994. *A performance theory of order and constituency*. Cambridge University Press.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of ACL*, pages 690–696.

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

Samar Husain, Shravan Vasishth, and Narayanan Srinivasan. 2014. Strong expectations cancel locality effects: evidence from hindi. *PLoS One*, 9(7):e100986.

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

Lars Konieczny. 2000. Locality and Parsing Complexity. *Journal of Psycholinguistic Research*, 29(6):627–645.

Taku Kudo. 2006. MeCab: Yet Another Part-of-speech and Morphological Analyzer. *http://mecab. sourceforge. jp*.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of ACL-IJCNLP*, pages 5203–5217.

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

Richard L Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.*, 29(3):375–419.

Richard L Lewis, Shravan Vasishth, and Julie A Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.*, 10(10):447–454.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written japanese. *Language Resources and Evaluation*, 48(2):345–371.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of CMCL*, pages 12–22.

James A Michaelov, Megan D Bardolph, Seana Coulson, and Benjamin Bergen. 2021. Different kinds of cognitive plausibility: why are transformers better than RNNs at predicting N400 amplitude? *Proceedings of CogSci*, 43(43).

Kentaro Nakatani and Edward Gibson. 2008. Distinguishing theories of syntactic expectation cost in sentence comprehension: Evidence from Japanese. *Journal of Linguistics*, 46(1):63–87.

Byung-Doh Oh, Christian Clark, and William Schuler. 2021. Surprisal estimators for human reading times need character models. In *Proceedings of ACL-IJCNLP*, pages 3746–3757.

Mai Omura and Masayuki Asahara. 2018. UD-Japanese BCCWJ: Universal Dependencies annotation for the Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 117–125.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL (Demonstrations)*, pages 48–53.

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

Mathieu Rita, Rahma Chaabouni, and Emmanuel Dupoux. 2020. "lazimpa": Lazy and impatient neural agents learn to communicate efficiently. In *Proceedings of CoNLL*, pages 335–343.

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. 2020. The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *bioRxiv*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of ACL*, pages 1715–1725.

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

Merity Stephen, Xiong Caiming, Bradbury James, Socher Richard, et al. 2016. Pointer sentinel mixture models. In *Proceedings of ICLR*.

Ryo Ueda and Koki Washio. 2021. On the relationship between zipf's law of abbreviation and interfering noise in emergent languages. In *Proceedings of ACL (SRW)*, pages 60–70.

Marten van Schijndel, Andy Exley, and William Schuler. 2013. A model of language processing as hierarchic sequential prediction. *Top. Cogn. Sci.*, 5(3):522–540.

Shravan Vasishth, Katja Suckow, Richard L Lewis, and Sabine Kern. 2010. Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Lang. Cogn. Process.*, 25(4):533–567.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *NIPS*, pages 5998–6008.

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. In *Proceedings of CogSci*, pages 1707–1713.

Ryo Yoshida, Hiroshi Noji, and Yohei Oseki. 2021. Modeling human sentence processing with left-corner recurrent neural network grammars. In *Proceedings of EMNLP*, pages 2964–2973.

A Psychometric predictive power and regression models

Psychometric predictive power refers to the similarity between (lossy-context) surprisal and human gaze duration, calculated using a linear mixed-effects regression (Bates et al., 2015). First, gaze duration (GD) is modeled by the following formula:

$$
\begin{aligned}
\text{GD} \sim\ & \text{surprisal} + \text{surprisal\_prev\_1} \\
& + \text{surprisal\_prev\_2} + \text{freq} * \text{length} \\
& + \text{freq\_prev\_1} * \text{length\_prev\_1} \quad\quad (4) \\
& + \text{screenN} + \text{lineN} + \text{segmentN} \\
& + (1|\text{article}) + (1|\text{subj}) \ .
\end{aligned}
$$

Table 5 shows the descriptions for the factors used in the above formulation. Then, a baseline regression model without the `surprisal`, `surprisal_prev_1`, and `surprisal_prev_2` terms from Eq. 4 is trained additionally. We calculated the per-token average of the log-likelihood difference (ΔLogLik) between the two regression models.

B Probabilistic erasure noise

Futrell and Levy (2017) and Futrell et al. (2020a) suggested that a linear probabilistic erasure noise (LPEN), where more distant items are more likely to disappear as opposed to a constant cutoff point with n-grams, might be a plausible design of input limitations. We examined whether such a probabilistic nature of noise design substantially affects our conclusions. Within our experimental settings, there is no substantial difference in the results regardless of the probabilistic nature of noise.

Methods. To implement LPEN, we erased the j-th nearest word in the context with a probability of $\min(j * a, 1)$, here $a > 0$. We initially observed that erasing too close context hindered human-like behavior; we also introduced an always-present portion of the context (l nearest words) and applied noise only on farther words. That is, the probabilistic erasure noise is only applied to $[w_0, \cdots, w_{i-l-1}]$. Assuming $a = 0.25$, w_{i-l-1} is then erased with a probability of 0.25, w_{i-l-2} is erased with a probability of 0.5, and so on, while the l nearest words to the target are left intact. We compared the PPP of surprisal with $l \in \{2, 3, 5, 7, 10, 20\}$ and $a \in \{0.5, 0.25, 0.125, 0.0625\}$.

Context limitation did not change or improved PPP. The results are shown in Figure 6. The

Factor	Type	Description
surprisal	num	(lossy-context) surprisal calculated by LMs
GD	num	reading time (first pass gaze duration)
article	factor	article ID
screenN	int	screen display order
lineN	int	the serial number of line the segment is displayed
segmentN	int	the serial number of segment in a screen
sentN	int	the serial number of sentence the segment belongs to
tokenN	int	the position of segment in sentence
length	int	number of characters
freq	num	geometric mean of the frequencies of subword constituents in a segment
subj	factor	participant ID

Table 5: Factor names and their description.

trends were similar to those using discrete context noise (Figure 1): (i) context limitation did not change or improved PPP and (ii) larger LMs have larger PPP gain due to context limitation.

C Exclusion criteria for eye movement data

We excluded outliers following Goodkind and Bicknell (2018). Specifically, we excluded the data points meeting any of the following criteria in the English experiments, and those meeting (a), (c), or (e) in the Japanese experiments:

(a) has zero gaze duration or beyond three standard deviations

(b) contains punctuation

(c) contains numeric characters

(d) the next segment has punctuation or numeric characters

(e) is the first segment in a line

(f) is the last segment in a line

We included data points meeting (b) and (f) in the Japanese data out of concern that excluding them disregards the data points for the main verb, regarding the verb-final Japanese construction (punctuation is included in a *bunsetsu*). Note that in the Japanese data, the first/end word in a line correspond to first/end word in a sentence (sentences are presented line by line.). Similarly, (d) substantially reduces the Japanese data points, and the inter-segment-level influence of special symbols
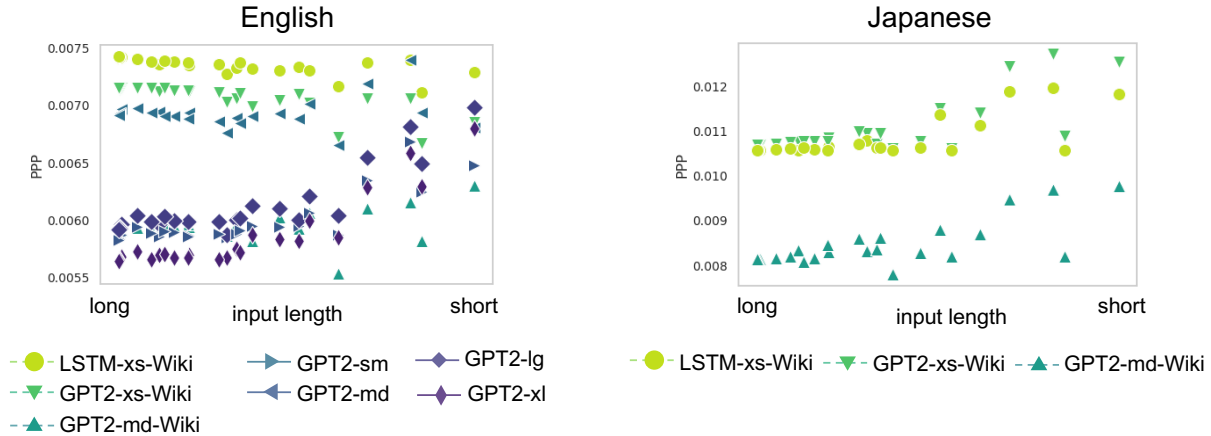
Figure 6: Relationship between the strength of input noise (X-axis) and PPP (Y-axis) under the probabilistic erasure noise settings. The results with various noise settings ($L \times A$ where $L = \{2, 3, 5, 7, 10, 20\}$ and $A = \{0.5, 0.25, 0.125, 0.0625\}$) are summarized with respect to average input length in each setting. The marker color and shape correspond to LM architectures.

would be less likely than in English considering that *bunsetsu* is a relatively larger unit than the English word.

D Hyperparameters of LMs

Table 6 shows the hyperparameters of Wiki-LMs. Training each LM took approximately three days on four GPUs (NVIDIA V100). The hyperparameters of OpenAI GPT-2s could be shown in https://huggingface.co/docs/transformers/model_doc/gpt2.

As for the subword tokenization used in Wiki-LMs, we set character coverage to 0.9995. Vocabulary size was set to 32,000 in English and 100,000 in Japanese, taking the rich characters and morphemes in Japanese into consideration. Note that this difference results in Japanese Wiki-LMs having more parameters than English LMs. In the Japanese settings, LSTM-xs-Wiki has 54M (27M in English), GPT2-xs-Wiki has 55M (29M in English), and GPT2-md-Wiki has 404M (335M in English) parameters.

E Mitigating training-inference gaps

As introduced in Section 4.1, we modified the training data to augment the data points, where LMs must predict the upcoming tokens from the middle of a sentence with severely limited context. Specifically, we first split each (i-th) sentence, $s^i = [w_0^i, w_1^i, \cdots, w_n^i]$, in training data into two sub-sequences: [<s>, $w_0^i, \cdots, w_{k-1}^i$] and [<b>, $w_k^i, \cdots, w_n^i$]. Here, the breakpoint k^i for the i-th sentence is sampled from the uniform distri-

bution $U(0, |s^i|)$. When $k = 0$, the former sub-sequence is [<s>]. Then, the sub-sequences obtained from the whole corpus were randomly concatenated to create the modified training data (e.g., $[\cdots, \text{<b>}, w_{k^c+1}^c, \cdots, w_n^c, \text{<b>}, w_{k+1}^a, \cdots, w_n^a, \text{<s>}, w_0^l, \cdots, w_k^l, \cdots]$). This modified data has two characteristics: (i) there is no dependency between before and after the special tokens (<s> and <b>), and (ii) to the <b> token, uniform prior about the token position within the sentence is set. It is expected that training LMs with this data and computing the next word prediction (e.g., with the query of [<b>, w]) enables them to adequately compute the upcoming token (i) with severely limited context and (ii) without any token position prior for <b>.

Concretely, the lossy-context surprisal is computed by the modified LMs as follows:

$$I_{\text{lossy}}(w_i, c_{<i}, f) = -\log p_\theta(w_i | \text{<b>} \circ f([w_0, \cdots, w_{i-1}])) , \quad (5)$$

Only when the <b> token corresponds to the sentence initial position, <s> was set instead of <b>.

F Perplexity of n-gram LMs

Perplexity (PPL): Perplexity (PPL), the inverse geometric mean of next-word probabilities $p(w_i | w_{<i})$ in a text that consists of N symbols $(w_1, w_2, \cdots, w_N)$, is a typical evaluation metric for the next-word prediction accuracy. Given a LM θ and context noise design f, the perplexity under lossy-context settings is calculated as follows:
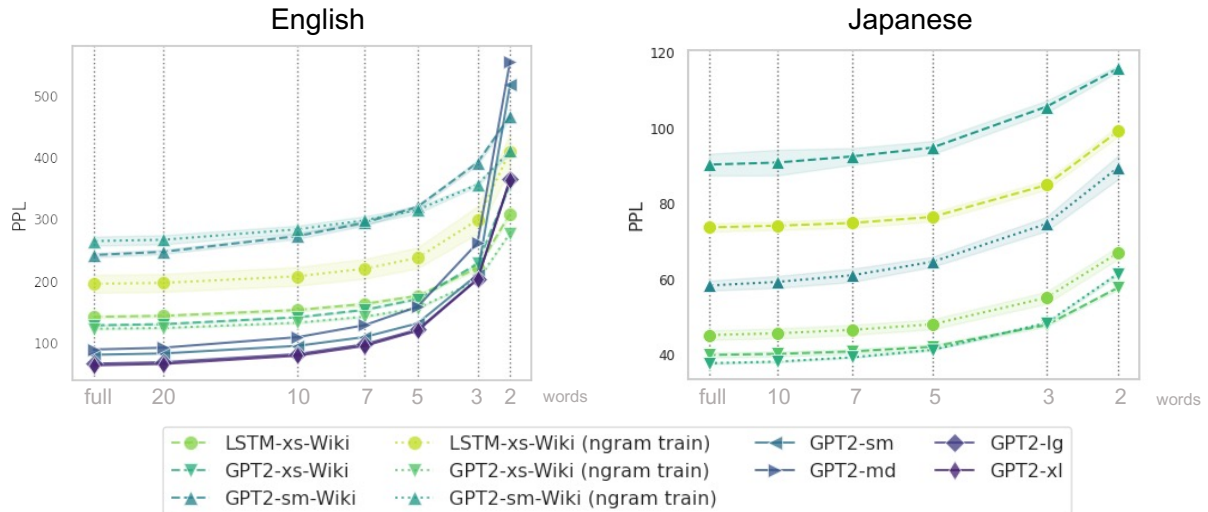
Figure 7: Relationship between the perplexity of n-gram LMs and input length. A monotonic relationship, the longer context LMs use, the lower perplexity LMs exhibit, is observed. The colored areas show a 95% confidence interval. The PPL was computed at the subword level; here, directly comparing the scale of Y-axis across languages is non-sense due to their different segmentation (e.g., vocabulary size).

$$ \text{PPL} = \prod_{i=0}^{N} p_\theta(w_i | \text{BOS} \circ f(c_{<i}))^{-\frac{1}{N}} \ . \qquad (6) $$

Low PPL indicates that the model and the context noise yielded accurate predictions about the upcoming signal.

Relationship between perplexity and context length: Figure 7 shows the relationship between the perplexity of n-gram LMs and their average context length. The PPL values are computed with the texts in the eye movement data. A monotonic relationship, the longer context LMs use, the lower perplexity LMs exhibit, is observed. This ensures that LMs with long context actually exploit the information in the added context to accurately predict the upcoming symbols.

G Next-word prediction accuracy and PPP in Japanese

Figure 8 shows the relationship between PPL and PPP in the Japanese experiments. Similar to the results of Section 5.1, LMs with a similar PPL value exhibited different PPP (e.g., results around PPL=60).
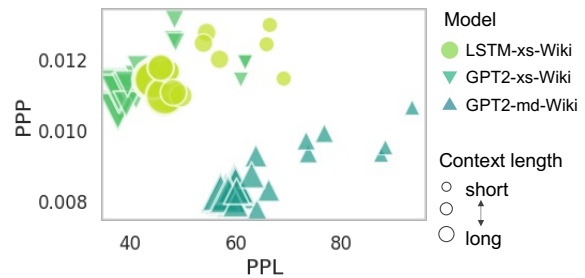

Figure 8: Relationship between PPP and perplexity (PPL) drawn using the Japanese LMs targeted in Section 5.1. Each point corresponds to each configuration of the n-gram surprisal computation; marker color and shape present the LM architectures, and larger markers correspond to longer context access.

Fairseq model	architecture	transformer_lm _gpt2_small
	adaptive softmax cut off	50,000, 140,000
	share-decoder-input-output-embed	True
	embed_dim	1,024
	ffn_embed_dim	4,096
	layers	24
	heads	16
	dropout	0.1
	attention_dropout	0.1
Optimizer	algorithm	AdamW
	learning rates	5e-4
	betas	(0.9, 0.98)
	weight decay	0.01
	clip norm	0.0
Learning rate scheduler	type	inverse_sqrt
	warmup updates	4,000
	warmup init lrarning rate	1e-7
Training	batch size	61,440 tokens
	sample-break-mode	none

(a) GPT2-md-Wiki.

Fairseq model	architecture	transformer_lm_gpt
	adaptive softmax cut off	50,000, 140,000
	share-decoder-input-output-embed	True
	embed_dim	384
	ffn_embed_dim	2,048
	layers	8
	heads	6
	dropout	0.1
	attention_dropout	0.1
Optimizer	algorithm	AdamW
	learning rates	5e-4
	betas	(0.9, 0.98)
	weight decay	0.01
	clip norm	0.0
Learning rate scheduler	type	inverse_sqrt
	warmup updates	4,000
	warmup init learning rate	1e-7
Training	batch size	61,440 tokens
	sample-break-mode	none

(b) GPT2-xs-Wiki.

Fairseq model	architecture	lstm_lm
	adaptive softmax cut off	50,000, 140,000
	share-decoder-input-output-embed	True
	embed_dim	400
	hiden_size	1,024
	layers	2
	dropout	0.1
Optimizer	algorithm	AdamW
	learning rates	1e-3
	betas	(0.9, 0.98)
	weight decay	0.01
	clip norm	0.0
Learning rate scheduler	type	inverse_sqrt
	warmup updates	4,000
	warmup init learning rate	1e-7
Training	batch size	20,480 tokens
	sample-break-mode	none

(c) LSTM-xs-Wiki.

Table 6: Hyperparameters of the LMs.