Information Integration in Large Language Models is Gated by Linguistic Structural Markers

Anonymous ACL submission

Abstract

Language comprehension relies on integrating information across both local words and broader context. We propose a method to quantify the information integration window of large language models (LLMs) and examine how sentence and clause boundaries constrain this window. Specifically, LLMs are required to predict a target word based on either a local window (local prediction) or the full context (global prediction), and we use Jensen-Shannon (JS) divergence to measure the information loss from relying solely on the local window, termed the local-prediction deficit. Results show that integration windows of both humans and LLMs are strongly modulated by sentence boundaries, and predictions primarily rely on words within the same sentence or clause: The local-prediction deficit follows a power-law decay as the window length increases and drops sharply at the sentence boundary. This boundary effect is primarily attributed to linguistic structural markers, e.g., punctuation, rather than implicit syntactic or semantic cues. Together, these results indicate that LLMs rely on explicit structural cues to guide their information integration strategy.

1 Introduction

Information in human language is hierarchically distributed across multiple scales, including words, sentences, and discourse (Chomsky, 1957; Phillips, 2003; Berwick et al., 2013). Evidence from cognitive science has demonstrated that information integration in human language processing is constrained by the multi-scale structure of language, which is thought to be central to hierarchical organization of the human brain (Hickok and Poeppel, 2007; Lerner et al., 2011; Friederici et al., 2017; Regev et al., 2024). How to integrate information across these time scales of language is also a central consideration when designing and evaluating large language models (LLMs). For instance, transformer-based LLMs can more effectively integrate over words than recurrent neural networks (Vaswani et al., 2017; Devlin et al., 2019; Raffel et al., 2020; Touvron et al., 2023). However, it remains unclear how LLMs integrate multi-scale information despite having theoretical access to all input tokens (Clark et al., 2019; Tenney et al., 2019). One possibility is that, like humans, LLMs may dynamically adjust their information integration according to language structures. Here, we examine whether the information integration windows of LLMs are modulated by a key structure of language, i.e., sentence boundary.

The information integration window is a well-established concept for studying human cognition, including human language comprehension (Poeppel, 2003; Hasson et al., 2008; Ding et al., 2016; Norman-Haignere et al., 2022), and is recently introduced to characterize information integration behavior of LLMs (Keshishian et al., 2021; Skrill and Norman-Haignere, 2023). For example, Skrill and Norman-Haignere (2023) examine the information integration window by analyzing how a perturbation influences the internal representations within an LLM and reveals a dynamically changing integration window across different layers. Here, we propose a method to characterize the information integration window purely based on model behavior, so that (1) the method can be easily applied to both humans and LLMs, and facilitate comparisons between LLMs and between LLM and human; (2) the method avoids analyzing a large number of internal nodes within an LLM, which may or may not directly contribute to model behavior.

In human studies, the information integration window is shown to be gated by structural boundaries in language. One example is the sentence wrap-up effect, in which the reading time is much longer for the final word of a sentence compared with non-sentence-final words (Rayner et al., 1989; Hirotani et al., 2006; Stowe et al., 2018). No-

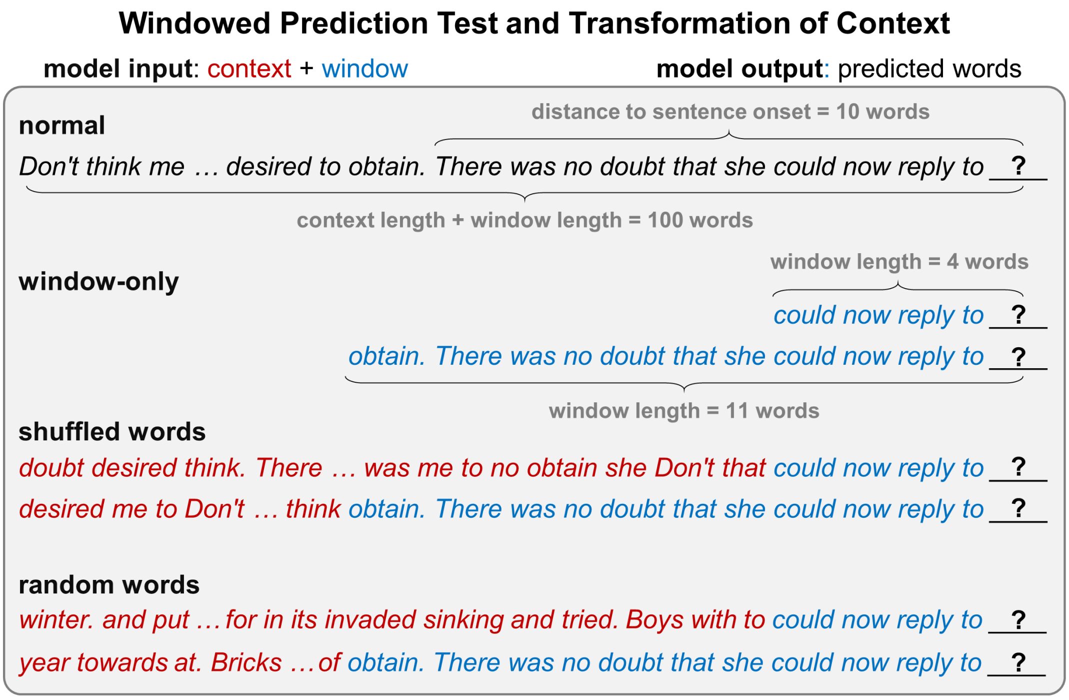


Figure 1: Demonstration of the windowed prediction test. Models are required to predict the next word based on either a local window or the full context. Predictions based on the local window are compared with predictions when the full context is available (**normal** condition).

tably, this effect diminishes when the sentence-final period is removed (Warren et al., 2009). Similarly, in the brain, a closure positive shift (CPS) EEG response typically occurs at the end of an intonation phrase in speech, and can be elicited by a comma during text reading (Steinhauer and Friederici, 2001). It has been hypothesized that punctuation is a structural marker that guide information integration across words (Rayner et al., 2000; Steinhauer, 2003; Moore, 2016). A main goal of the current study is to investigate whether structural boundaries modulate the information integration windows of LLMs, using a novel windowed prediction test to characterize the information integration window.

The windowed prediction test requires LLMs to predict the next word based on either a local window (local prediction) or the full context (global prediction). By systematically varying the window length, we characterize the integration windows of LLMs using the JS divergence between the predicted distributions under local and global conditions. Based on the windowed prediction test, we conduct a series of experiments based on GPT-2 (Radford et al., 2019) and Qwen2.5 (Qwen et al., 2025), and compare the results with human participants. It is found that the integration windows of both humans and LLMs are gated by sentence boundaries. Furthermore, the boundary-gating effect is primarily driven by overt structural markers, i.e., punctuation, rather than syntactic or semantic cues. The contributions of our study include: (1) introducing the windowed prediction test to characterize the information integration windows of both humans and LLMs, and (2) demonstrating that the windows are gated by linguistic structural markers.

2 Data construction

2.1 Tasks

In a windowed prediction test, LLMs are required to predict the next word based on the model input, which is divided into two parts: the local window and the broader context (Fig. 1). The total length of the local window and the broader context is always 100 words (see Appendix B for an extended length setting), while the window length is systematically varied. Words in the window remain unchanged across conditions, whereas the context is either intact (the **normal** condition) or transformed into one of three manipulated conditions:

1. **window-only**: The broader context is removed and the model input only consists of the window.

2. **shuffled words**: The order of words in the broader context is randomly shuffled.

3. **random words**: Each word in the broader context is replaced by a random word.

These conditions are designed to test the model's ability to utilize partial or degraded context, ranging from relying solely on local input (**window-only**), to integrating shuffled distal context (**shuffled words**), to remaining undistracted by irrelevant distal context (**random words**). Model predictions under each manipulated condition are compared with predictions based on the full context (**normal**).

2.2 Test Materials

For both Chinese and English, the test materials are articles sourced from three distinct domains: Wikipedia, news, and books (Koupaee and Wang, 2018; Cui et al., 2019; Kryściński et al., 2021). All articles are publicly available and distributed under the CC-BY-SA 3.0 license. We exclude articles that contain characters from other languages (i.e., non-Chinese or non-English), as well as those shorter than 300 characters (for Chinese) or 300 words (for English). Finally, for each language, we retain a total of 7,500 articles, with 2,500 articles from each domain.

2.3 Window Length and Distance to Sentence Onset

We define two parameters, the window length and the distance to sentence onset, to examine the information integration window at different positions within a sentence. The distance to sentence onset refers to the number of words between the target word (i.e., the word to be predicted) and the first word of the same sentence. The window length refers to the number of words included in the window (Fig. 1). When the distance to sentence onset exceeds the window length, the window contains a sentence fragment. In contrast, when the distance to sentence onset is less than or equal to the window length, the window contains a complete sentence.

3 Experiment 1: Modulation by Sentence Boundary

3.1 Experimental setup

In Experiment 1, we examined whether sentence boundaries modulate the information integration windows of both humans and LLMs. For LLMs, we tested the base version of GPT-2 and Qwen2.5-1.5B on Chinese and English articles. For GPT-2, we used separate Chinese and English model variants for testing. In contrast, since Qwen2.5-1.5B was a multilingual model (Qwen et al., 2025), we used the same model variant for both languages. Both models were only pretrained without any task-specific fine-tuning, and were required to predict the next word based on the input. We varied the distance to sentence onset from 1 to 20 words. For each distance to sentence onset, we sampled 1,000 articles and truncated the articles to meet the criteria. For each article, the window length was increased from 1 to 20 words, starting from the final word in the article. No linguistic structural markers (e.g., dots and commas) occurred within the window. The context outside the window was transformed into one of the four different conditions described previously. In total, we constructed 20 × 1,000 × 20 × 4 tests for each model and language. All experiments were repeated across 10 different random seeds.

We used Jensen-Shannon (JS) divergence to measure the information loss from relying solely on a local window instead of the full context, referred to as the local-prediction deficit:

$$Deficit(w, d) = JS(N_{w,d}, M_{w,d})$$

where w denotes the window length and d denotes the distance to sentence onset. $N_{w,d}$ and $M_{w,d}$ represent the output probability distributions under the normal and manipulated conditions, respectively, for an input constructed based on a given w and d. We utilized the local-prediction deficits to construct a two-dimensional deficit matrix (Fig. 2a), where each element in the matrix represented the average local-prediction deficit for a specific window length and distance to sentence onset. We hypothesized that words outside sentence boundaries would have less impact on model predictions than words within the boundary. Therefore, the diagonal of the deficit matrix was expected to be salient since the window exceeded the sentence boundary on the diagonal. To quantify this boundary effect, we first performed a regression analysis to control the confounding effects of the window length and distance to sentence onset (see Appendix A for more details). The strength of boundary effect was then quantified as the difference in residual deficits between adjacent positions on either side of the diagonal, averaged across all distances to sentence onset.

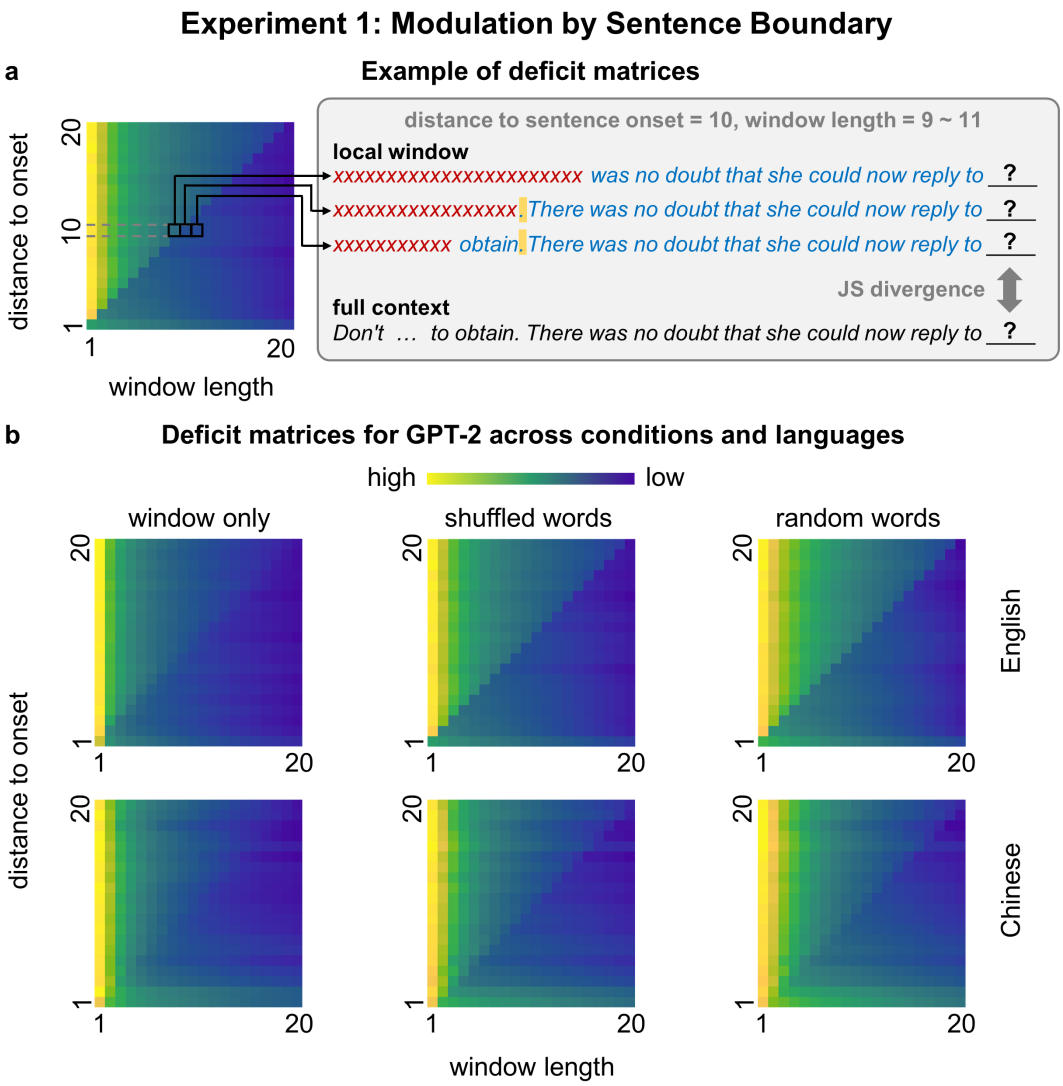


Figure 2: Divergence between predictions based on the full context and predictions based on a local window in Experiment 1. **a.** Example of the deficit matrices. In tests above the diagonal, the local window does not cover the current sentence. In tests below the diagonal, the local window exceeds the current sentence. The sentence boundary is highlighted. **b.** Deficit matrices for GPT-2 across conditions and languages. See the results of Qwen2.5 in Appendix Fig. 1.

We conducted the human experiment using the Chinese version of Experiment 1. To control the experiment time, we fixed the distance to sentence onset at 10 words, and varied the window length from 8 to 12 words. Fifty articles that met the criteria were sampled. The boarder context of each article was either unchanged (**normal**) or replaced by randomly selected words (**random words**). A total of 100 participants were recruited, with each participant receiving 50 tests. In each test, the participant was shown an article and instructed to continue the article by writing 1 to 6 Chinese character(s). Test assignments were counter-balanced, with each participant receiving 10 tests per window length and 25 tests per condition. All participants provided written consent and were paid. Human responses were pooled to compute the output distribution of the first continued character. JS divergence was then calculated between the output distributions under the **normal** and **random words** conditions.

3.2 Result

The results of GPT-2 are shown in Fig. 2b, with the full set of results presented in Appendix Fig. 1. For both Chinese and English, the local-prediction deficits decreased as the window length increased, showing a sharp drop when the window crossed the sentence boundary and then stabilized. This pattern resulted in a salient diagonal in the deficit matrices for GPT-2, indicating that the model predictions

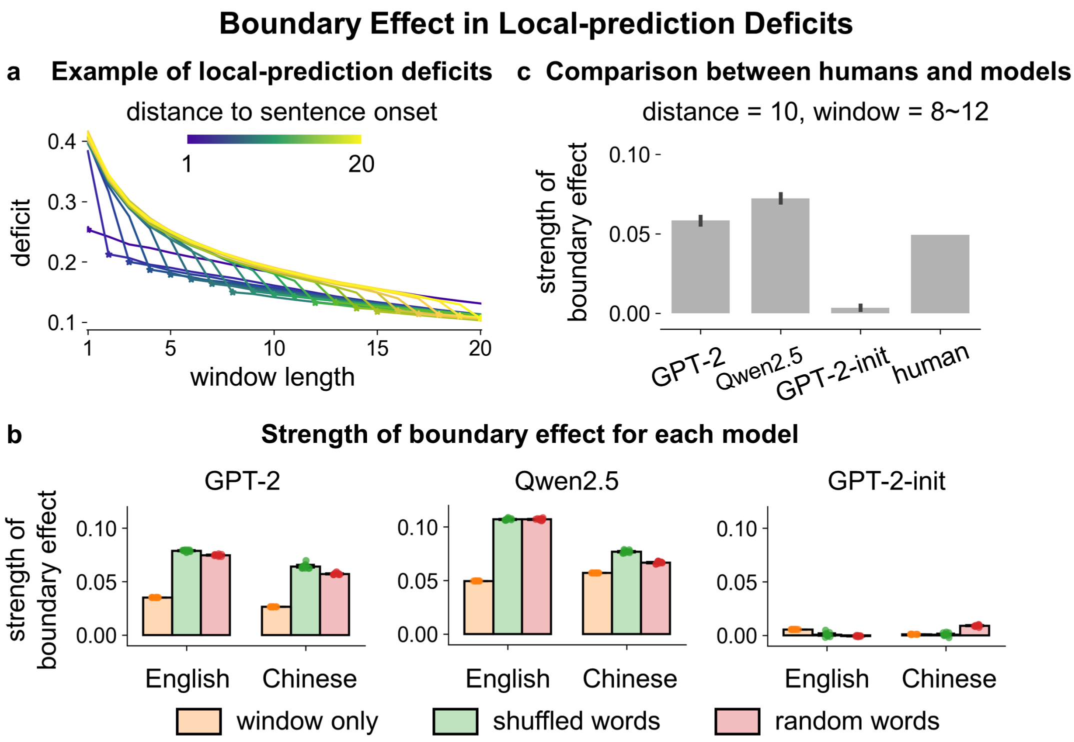


Figure 3: The boundary effect in the local-prediction deficits. **a.** Local-prediction deficits in the English version of Experiment 1, under the **shuffled words** condition. The sentence boundary is marked with a star. **b.** The strength of boundary effect for each model, i.e., GPT-2, Qwen2.5, and GPT-2 with randomly initialized weights. Each dot represents data from a single model run under a different random seed. Error bars represent 95% confidence intervals (CIs) of the mean across runs, estimated using bootstrap. **c.** Comparison between humans and models in the Chinese version of Experiment 1, under the **random words** condition.

relied more on words within the sentence boundaries than on words outside the boundaries across all conditions. Additionally, the local-prediction deficits exhibited a non-linear decay as the window length increased (Fig. 3a). We fitted multiple linear and nonlinear functions to the deficit matrices for each model, and found that a power-law function provided the best fit (see Appendix Fig. 2). Based on the residuals obtained after fitting, we quantified the strength of boundary effect to assess how sentence boundaries modulated the windows. As shown in Fig. 3b, GPT-2 and Qwen2.5 exhibited a significant boundary effect in both languages, whereas no boundary effect was observed in the model without language training (i.e., GPT-2 with randomly initialized weights). For both GPT-2 and Qwen2.5, the **shuffled words** and **random words** conditions consistently elicited stronger boundary effects compared to the **window-only** condition. The results indicated that sentence boundaries significantly gated the contribution of distal context beyond the current sentence, and this boundary-gating effect strengthened when degraded context was provided.

The results of the human experiment are shown in Fig. 3c. A boundary effect was also observed in human responses, though the strength was weaker than that in GPT-2 and Qwen2.5. This discrepancy might reflect that humans could implicitly infer sentence boundaries from the context – The local-prediction deficits of humans decreased sharply before the sentence boundary (i.e., at a window length of 9 words; see Appendix Fig. 3). In contrast, language models might rely more heavily on explicit cues (e.g., punctuation) to identify the boundary. Altogether, these results demonstrated that the information integration windows of both humans and LLMs were gated by sentence boundaries, and such boundary-gating effect might arise from language training. Experiment 1 was also conducted on a larger language model (i.e., Qwen2.5-72B) and with a longer context (context length + window length = 1000 words) to examine the generalizability of our results. The results remained consistent (see Appendix B and Appendix Fig. 4).

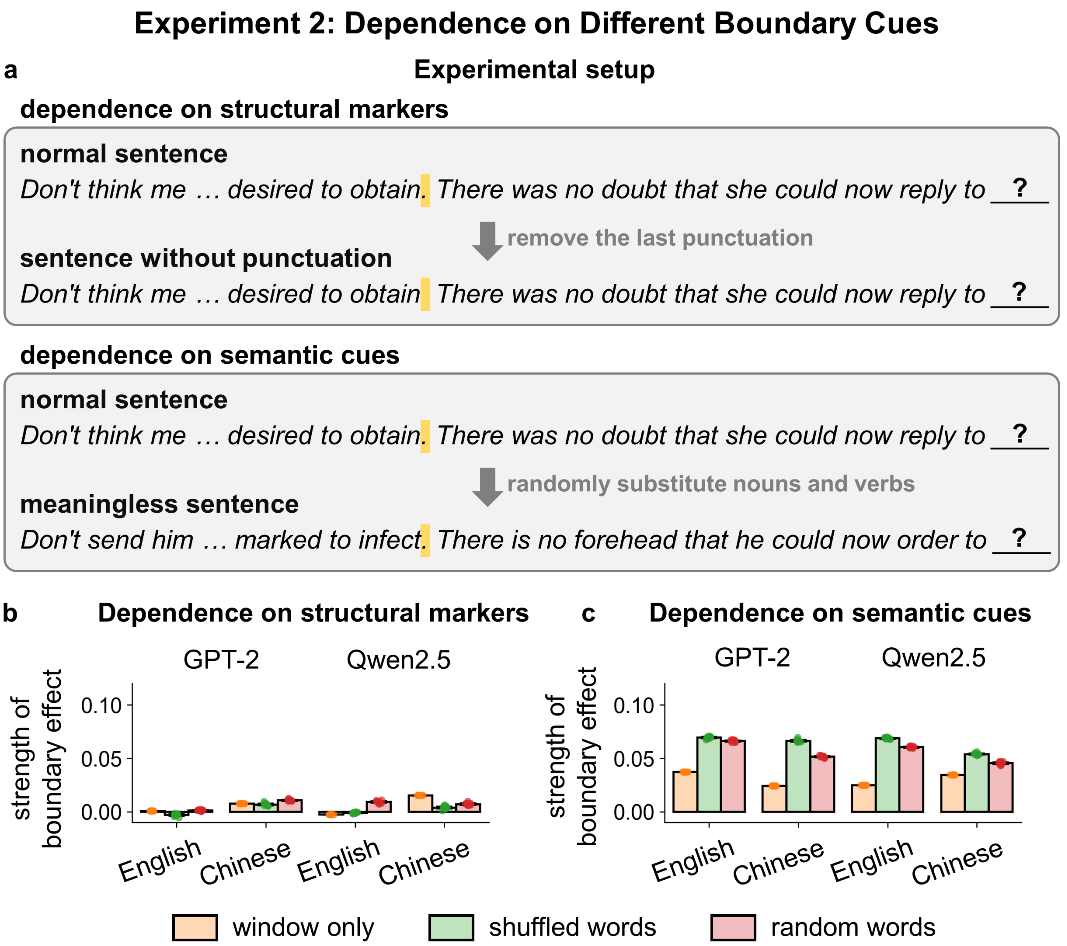


Figure 4: Results of models in Experiment 2. **a.** Experimental setup of Experiment 2. We separately construct the sentences without punctuation and meaningless sentences for testing. **b.** The strength of boundary effect for each model when structural markers are removed. **c.** The strength of boundary effect for each model when semantic cues are disrupted.

4 Experiment 2: Dependence on Different Boundary Cues

4.1 Experimental setup

As suggested in Experiment 1, LLMs used sentence boundaries to modulate the integration windows. However, sentence boundaries can manifest based on various cues, including implicit syntactic boundaries, semantic coherence, and linguistic structural markers such as punctuation. Experiment 2 aimed to disentangle the contributions of different boundary cues by selectively removing structural markers and semantic cues from the model input. We tested GPT-2 and Qwen2.5-1.5B on inputs where either structural markers or semantic cues were removed. To remove structural markers, we eliminated the last punctuation from the model input (Fig. 4a). To disrupt semantic cues, we constructed meaningless sentences by randomly substituting nouns and verbs with other words of the same part of speech. All other experimental setups were consistent with those of Experiment 1.

4.2 Result

The strength of boundary effect in Experiment 2 is shown in Figs. 4b and 4c. When structural markers were removed, the boundary effect nearly disappeared (Fig. 4b), indicating that GPT-2 and Qwen2.5 failed to utilize implicit syntactic cues to modulate the integration window. For the meaningless sentences, where semantic cues were disrupted, the boundary effect diminished compared to Experiment 1 but was still retained (Fig. 4c). These results suggested that both GPT-2 and Qwen2.5 primarily relied on linguistic structural markers, rather than implicit syntactic and semantic cues, to gate the information integration.

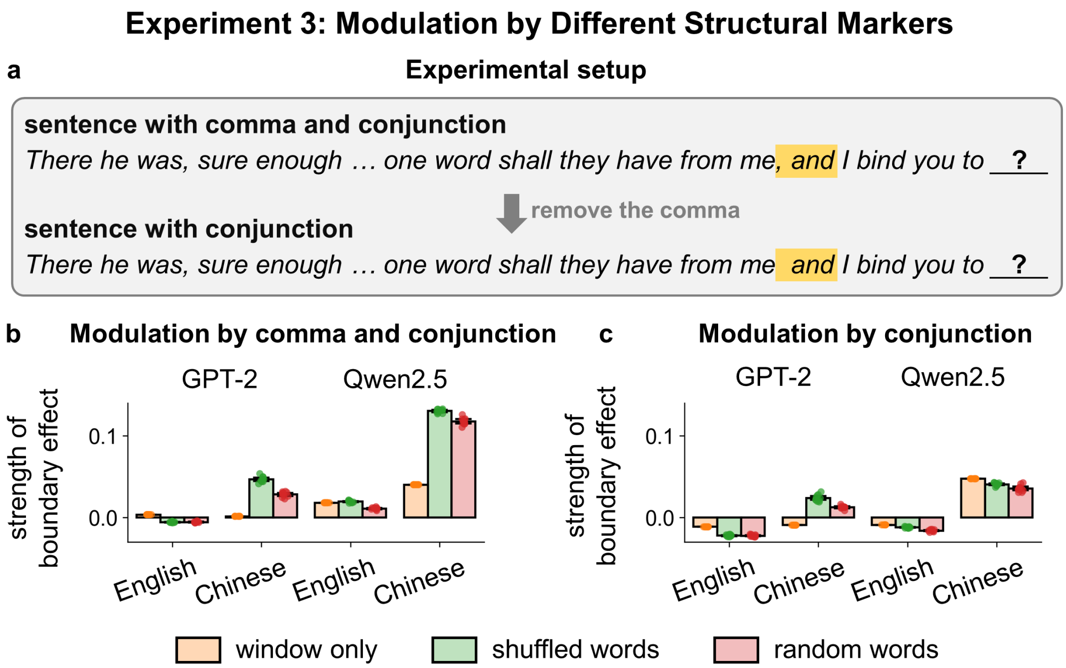


Figure 5: Results of models in Experiment 3. **a.** Experimental setup of Experiment 3. We focus on comma-conjunction pairs as structural markers. **b.** The strength of boundary effect for each model when both commas and conjunctions are retained. **c.** The strength of boundary effect for each model when only conjunctions are retained.

5 Experiment 3: Modulation by Different Structural Makers

5.1 Experimental setup

Experiments 1 and 2 demonstrated that the integration windows of LLMs were primarily modulated by linguistic structural markers. In Experiment 3, we investigated how different types of markers modulated the integration windows. Specifically, we focused on comma-conjunction pairs (e.g., ", and", ", or", ", but") as structural markers (Fig. 5a), and calculated a revised distance to sentence onset based on these markers (i.e., the number of words between the target word and the comma). The revised distance was used to select the articles for testing. The comma was either retained or removed to isolate the effect of commas and conjunctions. We tested GPT-2 and Qwen2.5 in Experiment 3, and all other experimental setups were consistent with those in Experiment 1.

5.2 Result

The strength of boundary effect in Experiment 3 is shown in Figs. 5b and 5c. When both commas and conjunctions were retained, a significant boundary effect was observed in GPT-2 and Qwen2.5 in Chinese (Fig. 5b). However, in English, the boundary effect was relatively weak for Qwen2.5 and absent for GPT-2. One possible explanation for this cross-linguistic pattern was that Chinese generally contained fewer complex relative clauses than English (Li and Thompson, 1989; Lin, 2011). In Chinese, a comma was typically followed by a complete sentence rather than a dependent clause, which might lead to stronger sentence boundary cues being associated with the comma. Language models might capture the cross-linguistic difference, and therefore rely more heavily on commas to modulate the integration window in Chinese than in English.

When commas were removed (Fig. 5c), the strength of boundary effect declined across all models and languages. However, a residual effect remained for Qwen2.5 in Chinese. The results suggested that more extensive language training might allow the model to utilize more structural markers to modulate the integration window, and therefore Qwen2.5 appeared to rely not only on commas, but also on conjunctions to guide its information integration.

6 Related work

Recent advancements in LLMs have increasingly focused on enabling language comprehension over extremely long context. While it is crucial for

LLMs to extract relevant information from such extended sequences, there is growing evidence that LLMs prioritize information within a limited span of preceding context (Keshishian et al., 2021; Skrill and Norman-Haignere, 2023). This phenomenon parallels findings from cognitive science, which suggest that humans integrate information within constrained temporal windows during language comprehension (Poeppel, 2003; Hasson et al., 2008; Norman-Haignere et al., 2022). Inspired by these findings, recent studies have attempted to characterize information integration windows of LLMs by analyzing internal representations such as activations of hidden states. For instance, Keshishian et al. (2021) have explored the integration windows of deep speech models using the temporal context invariance paradigm, while Skrill and Norman-Haignere (2023) have developed a word-swap procedure that reveals a dynamically changing integration window across different layers in LLMs. However, prior work has predominantly focused on a large number of internal nodes within LLMs, which cannot intuitively inform how these integration windows may contribute to model behavior. Our study aims to directly analyze information integration in terms of model behavior and compare it with that of humans under the same experimental paradigm. Furthermore, we focus on whether the integration windows are gated by sentence boundaries, examining the effects of different boundary cues in a multilingual setting.

The structure of language can manifest based on various cues, including implicit syntactic boundaries and semantic coherence. Researchers have explored the encoding of structured sentence representations (e.g., dependency and constituency) in LLMs. Such representations can be reconstructed from internal activations (Hewitt and Manning, 2019; Arps et al., 2022) or model behavior (Cao et al., 2020; Liu et al., 2024), and can influence the processing dynamics of LLMs (Kovaleva et al., 2019; Wu et al., 2020). Our study contributes to this body of literature, and further demonstrates that explicit linguistic structural markers can also gate the information integration in LLMs. One of the interesting findings of our study is that the boundary-gating effect disappears when the linguistic structural markers are removed, which echoes the sentence wrap-up effect observed in human reading. The sentence wrap-up effect refers to increased reading times at sentence-final words, and this effect diminishes when the sentence-final markers are removed (Warren et al., 2009; Stowe et al., 2018). It has been argued that the wrap-up effect reflects the low-level reaction to visual cues (Hill and Murray, 2000). Our results show that a similar effect of markers arises in LLMs, even though these models lack any visual modality. This suggests that the wrap-up effect may not merely reflect a hesitation response to visual stimuli, but instead emerges as a general information integration strategy—one that facilitates structural integration near sentence boundaries across both biological and artificial systems.

In addition, processing long context imposes significant computational and memory costs due to the quadratic complexity of attention in transformer-based architectures (Vaswani et al., 2017; Duman Keles et al., 2023). To address this, some researchers have proposed hybrid architectures that combine sliding window mechanisms with retrieval modules (Beltagy et al., 2020; Xiao et al., 2024; Yuan et al., 2025). Our findings suggest that LLMs may already implicitly adopt a sliding-window-like mechanism during prediction, independent of explicit architectural designs. We provide behavioral evidence that LLMs prioritize information within sentence boundaries, informing the development of more efficient architectures, such as by dynamically adjusting sliding windows based on language structures. Overall, our study not only offers insights into the internal information integration strategies of current LLMs, but also suggests pathways for improving long-context processing in a more linguistically grounded manner.

7 Conclusion

In summary, our study examines whether information integration in LLMs is gated by sentence boundaries. Using the windowed prediction test, we show that, for both humans and LLMs, next word prediction relies more on words within the same sentence or clause than on words beyond the sentence or clause boundary. This boundary-gating phenomenon is not observed in a randomly initialized model. Furthermore, the effect of sentence/clause boundaries is primarily attributed to linguistic structural markers, rather than syntactic and semantic cues. These results indicate LLMs rely on structural markers to guide their information integration strategies.

Limitations

Although our study systematically examined the information integration windows of LLMs, we did not investigate how such windows emerged. The differences between the initialized and pretrained models suggested that the structured integration window might arise from language training, but the specific linguistic features responsible for these windows remained unclear. Future work could explore the integration windows across different model sizes and amounts of training data, or analyze how the windows evolve over the course of pretraining.

Our study focused on sentence boundaries as a key structure of language, since sentence boundaries represented a relatively well-defined language structure. However, natural language is hierarchically structured at many scales. Future research could explore whether information integration in LLMs exhibits hierarchical organization across linguistic scales, from phrases to discourse.

References

David Arps, Younes Samih, Laura Kallmeyer, and Hassan Sajjad. 2022. Probing for Constituency Structure in Neural Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6738–6757, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150*.

Robert C. Berwick, Angela D. Friederici, Noam Chomsky, and Johan J. Bolhuis. 2013. Evolution, brain, and the nature of language. *Trends in Cognitive Sciences*, 17(2):89–98.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Unsupervised Parsing via Constituency Tests. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4798–4808, Online. Association for Computational Linguistics.

Noam Chomsky. 1957. *Syntactic Structures*. De Gruyter Mouton, Berlin, Boston.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT‘s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A Span-Extraction Dataset for Chinese Machine Reading Comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5883–5889, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nai Ding, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel. 2016. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, 19(1):158–164.

Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. 2023. On the computational complexity of self-attention. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pages 597–619. PMLR.

Angela D. Friederici, Noam Chomsky, Robert C. Berwick, Andrea Moro, and Johan J. Bolhuis. 2017. Language, mind and brain. *Nature human behaviour*, 1(10):713–722.

Uri Hasson, Eunice Yang, Ignacio Vallines, David J. Heeger, and Nava Rubin. 2008. A Hierarchy of Temporal Receptive Windows in Human Cortex. *The Journal of Neuroscience*, 28(10):2539.

John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Gregory Hickok and David Poeppel. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402.

Robin L. Hill and Wayne S. Murray. 2000. Chapter 22 - Commas and Spaces: Effects of Punctuation on Eye Movements and Sentence Parsing. In Alan Kennedy, Ralph Radach, Dieter Heller, and Joël Pynte, editors, *Reading as a Perceptual Process*, pages 565–589. North-Holland, Oxford.

Masako Hirotani, Lyn Frazier, and Keith Rayner. 2006. Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language*, 54(3):425–443.

Menoua Keshishian, Sam V. Norman-Haignere, and Nima Mesgarani. 2021. Understanding adaptive, multiscale temporal integration in deep speech recognition systems. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Mahnaz Koupaee and William Yang Wang. 2018. WikiHow: A Large Scale Text Summarization Dataset. _eprint: 1810.09305.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the Dark Secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. BookSum: A Collection of Datasets for Long-form Narrative Summarization. _eprint: 2105.08209.

Yulia Lerner, Christopher J. Honey, Lauren J. Silbert, and Uri Hasson. 2011. Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *Journal of Neuroscience*, 31(8):2906–2915. Publisher: Society for Neuroscience _eprint: https://www.jneurosci.org/content/31/8/2906.full.pdf.

Charles N Li and Sandra A Thompson. 1989. *Mandarin Chinese: A functional reference grammar*. Univ of California Press.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. LooGLE: Can Long-Context Language Models Understand Long Contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333, Bangkok, Thailand. Association for Computational Linguistics.

Chien-Jer Charles Lin. 2011. Chinese and English relative clauses: Processing constraints and typological consequences. In *Proceedings of the 23rd North American Conference on Chinese Lingusitics (NACCL-23). University of Oregon, Eugene*, volume 1, pages 191–199.

Wei Liu, Ming Xiang, and Nai Ding. 2024. Active Use of Latent Constituency Representation in both Humans and Large Language Models. _eprint: 2405.18241.

Nick Moore. 2016. What's the point? The role of punctuation in realising information structure in written English. *Functional Linguistics*, 3(1):6.

Sam V. Norman-Haignere, Jenelle Feather, Dana Boebinger, Peter Brunner, Anthony Ritaccio, Josh H. McDermott, Gerwin Schalk, and Nancy Kanwisher. 2022. A neural population selective for song in human auditory cortex. *Current Biology*, 32(6):1454–1455. Publisher: Elsevier.

Colin Phillips. 2003. Linear order and constituency. *Linguistic inquiry*, 34(1):37–90.

David Poeppel. 2003. The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time.'. *Speech Communication*, 41(1):245–255. Place: Netherlands Publisher: Elsevier Science.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 Technical Report. _eprint: 2412.15115.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1). Publisher: JMLR.org.

Keith Rayner, Gretchen Kambe, and Susan A. Duffy and. 2000. The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology Section A*, 53(4):1061–1080. Publisher: Routledge _eprint: https://doi.org/10.1080/713755934.

Keith Rayner, Sara C. Sereno, Robin K. Morris, A. Réne Schmauder, and Charles Clifton. 1989. Eye movements and on-line language comprehension processes. *Language and Cognitive Processes*, 4(3-4):SI21–SI49. Place: United Kingdom Publisher: Taylor & Francis.

Tamar I. Regev, Colton Casto, Eghbal A. Hosseini, Markus Adamek, Anthony L. Ritaccio, Jon T. Willie, Peter Brunner, and Evelina Fedorenko. 2024. Neural populations in the language network differ in the size of their temporal receptive windows. *Nature Human Behaviour*, 8(10):1924–1942.

David Skrill and Sam V. Norman-Haignere. 2023. Large language models transition from integrating across position-yoked, exponential windows to structure-yoked, power-law windows. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc. Event-place: New Orleans, LA, USA.

Karsten Steinhauer. 2003. Electrophysiological correlates of prosody and punctuation. *Brain and Language*, 86(1):142–164.

Karsten Steinhauer and Angela D. Friederici. 2001. Prosodic boundaries, comma rules, and brain responses: The closure positive shift in ERPs as a universal marker for prosodic phrasing in listeners and readers. *Journal of Psycholinguistic Research*, 30(3):267–295. Place: Germany Publisher: Springer.

Laurie A. Stowe, Edith Kaan, Laura Sabourin, and Ryan C. Taylor. 2018. The sentence wrap-up dogma. *Cognition*, 176:232–247.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, and Faisal Azhar. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

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. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tessa Warren, Sarah J. White, and Erik D. Reichle. 2009. Investigating the causes of wrap-up effects: Evidence from eye movements and E—Z Reader. *Cognition*, 111(1):132–137. Place: Netherlands Publisher: Elsevier Science.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient Streaming Language Models with Attention Sinks. In *The Twelfth International Conference on Learning Representations*.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Y. X. Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. 2025. Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention. _eprint: 2502.11089.

A Regression Analysis

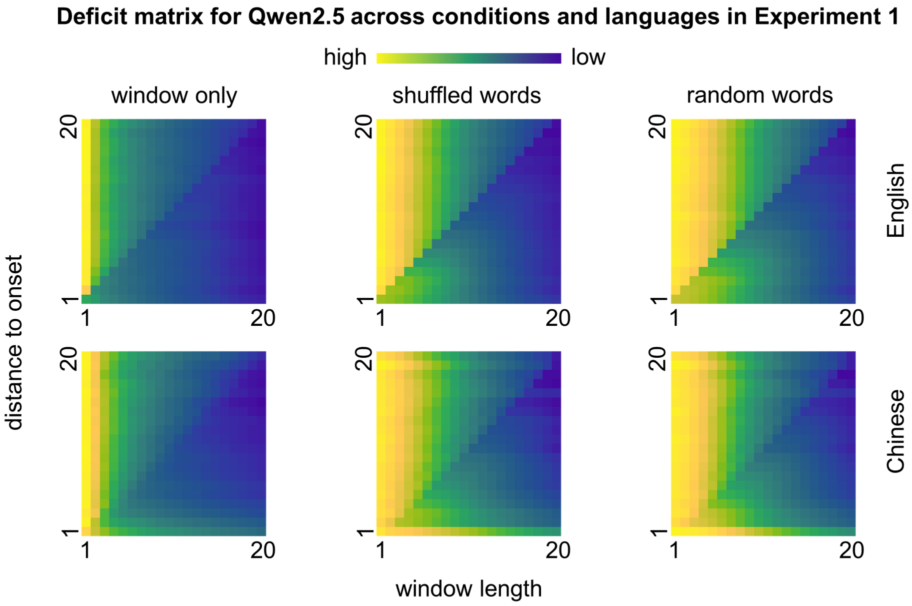
We fit the deficit matrices from humans and language models using the window length and distance to sentence onset, and then calculated the strength of boundary effect based on the residuals after fitting. All matrices were normalized by the maximum value before regression. We used three functions to fit the deficit matrices:

1. linear: $D(w, d) = -x_1 \cdot w - x_2 \cdot d + x_3$
2. exponential: $D(w, d) = e^{-x_1 \cdot w} + e^{-x_2 \cdot d} + x_3$
3. power-law: $D(w, d) = w^{-x_1} + d^{-x_2} + x_3$

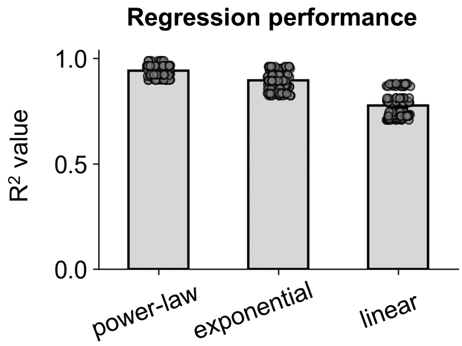
where w denotes the window length, and d denotes the distance to sentence onset. x_1, x_2, and x_3 are fitting parameters. Since the power-law function yielded the best fit in most cases (Appendix Fig. 2), it was selected for subsequent analyses. The strength of boundary effect was calculated based on the residuals of the fitted power-law function.

B Generalizability to Longer Context and Larger Model

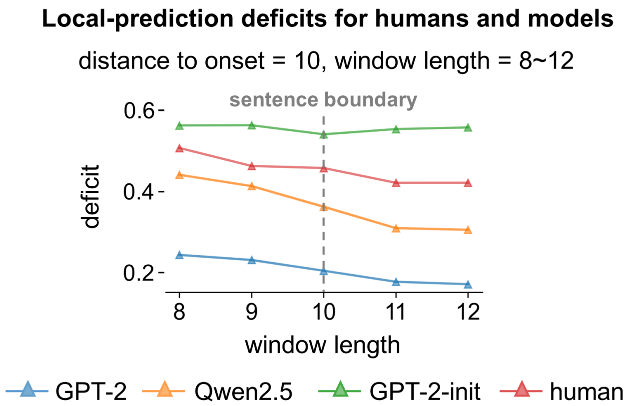
We extended Experiment 1 with longer context and larger language models to assess the generalizability of our results. Long-context articles were obtained from Loogle (Li et al., 2024), with only those exceeding 10,000 words retained. For the context length extension, we replicated the English version of Experiment 1 using Qwen2.5-1.5B, with the total length of the context and window set to 1,000 words. For model size extension, we conducted the English version of Experiment 1 using Qwen2.5-72B. For the experiment with Qwen2.5-72B, we sampled only 100 articles for each distance to sentence onset to reduce computational cost. The results indicated that neither context length nor model size significantly affected the strength of boundary effect (see Appendix Figure 4).



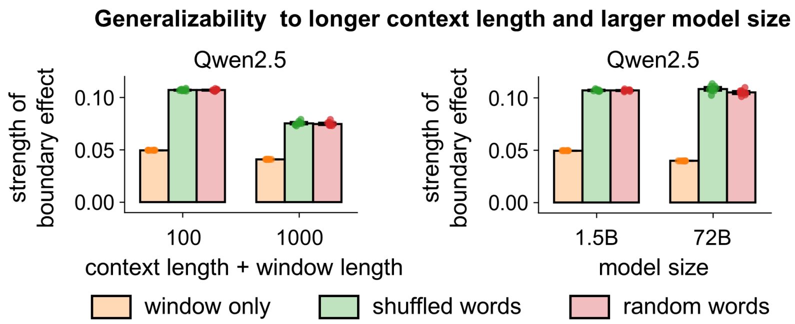
Appendix Figure 1. Deficit matrices for Qwen2.5 across conditions and languages in Experiment 1.



Appendix Figure 2. The regression performance when fitting the deficit matrices. Each dot represents a deficit matrix under a random seed for GPT-2 or Qwen2.5 in Experiment 1.



Appendix Figure 3. Local-prediction deficits for humans and LLMs in the Chinese version of Experiment 1, under the **random words** condition.



Appendix Figure 4. The strength of boundary effect for Qwen2.5 in the English version of Experiment 1, tested across two combined context and window lengths (100 vs. 1000 words) and two model sizes (1.5B vs. 72B). The boundary effect remains generally consistent across different context lengths and model sizes.