

Incremental Accumulation of Linguistic Context in Artificial and Biological Neural Networks

Corresponding Author: Dr Refael Tikochinski

This manuscript has been previously reviewed at another journal. This document only contains information relating to versions considered at Nature Communications.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Thank you for addressing the comments and for providing additional experiments involving the construction of encoding models per subject per voxel, which have resolved some of my concerns. However, some concerns still remain.

- Recent LLMs like Longformer, a Transformer variant, use a modified attention mechanism for handling longer contexts through a combination of local and global attention. Local attention focuses on incremental contexts, where each token attends only to its nearby context within a predefined window (e.g., 512 tokens). As the model processes longer documents, it builds representations incrementally from these segments. Global attention targets key tokens (e.g., special markers or task-specific tokens), allowing them to attend to all other tokens, thereby capturing broader context and integrating critical information. A study by Aw and Toneva (2023) used these deep language models for longer input contexts, showing improved predictions in language regions identified by the Fedorenko lab. Hence, it is crucial to compare these incremental LLM representations with varying context lengths to the incremental-context LLM in this paper.

- The study by Aw and Toneva (2023) explored larger context window sizes, up to 1000 tokens, and found that brain predictions improved only for longer input contexts (20 to 1000 words), not for shorter ones (1 to 5 words). They demonstrated that training language models for deeper narrative understanding enhances predictions of human brain responses, with improvements seen as the number of input words increased up to 500. Aw and Toneva (2023) tested four language models capable of handling longer context lengths, yielding similar insights. How do the authors reconcile these findings with the observed findings in this study?

Reference:

Aw and Toneva (2023), Training language models to summarise narratives improves brain alignment, ICLR-2023.

- The authors responded that a comprehensive layer-by-layer analysis revealed a performance decline in most layers when the context exceeded 32 tokens. However, the better predictions observed when the last layers are used, remain puzzling. Most linguistic brain encoding studies indicate that middle layers generally show better predictions of brain responses. Furthermore, research on longer context models handling both local and global contexts shows that middle layers still achieve better predictions, even with context lengths exceeding 200 tokens. Thus, the authors need to explain why their findings contradict previous studies.

- On Figure 2, why is the prediction performance in the A1 (primary auditory cortex) and STG regions similar to that in high-level language regions, including the IFG (inferior frontal gyrus) and TPJ (temporoparietal junction)? Prior studies have consistently shown that representations from language models better predict brain responses in higher-order language regions than in early sensory regions like A1. However, in this case, the prediction performance is comparable in both A1 and the late-stage language regions. What could be the reason for this unexpected result?

Minor Comments

- Could the authors explain how a single token corresponds to 0.75 words?

- The authors write "the human brain processes the incoming linguistic input serially, word by word, as speech and text unfold over time"

If the authors want to make a statement only for listening then I suggest removing the word "text" in "speech and text unfold over time".

Reviewer #3

(Remarks to the Author)

Thanks for the interesting read. I have been brought on as a replacement for Reviewer #2 due to their unavailability, and have been asked to recommend whether the revisions proposed by the authors have successfully met Reviewer #2's initial concerns. Based on my understanding of Reviewer #2's concerns, I believe that they mostly have. Here is a point-by-point address of the questions posed by Reviewer #2 and the author's responses.

The reviewer's largest concern was due to the averaging of their results across ROIs and participants. This lowered the quality of the signal and reduced the reliability of the results by not acknowledging and accounting for high variation in per-subject responses. The reviewers resolved this by training independent encoding models to predict the BOLD signal at the per-voxel level. I am somewhat surprised by their comment that this took a month, but nevertheless, this concern seems to be largely addressed.

The reviewer was concerned that the permutation test proposed by the authors was unhelpful for comparing the performance of different models. Their move to a Wilcoxon signed rank test should resolve this concern.

The reviewer's concerns about many areas, such as angular gyrus, not being predicted better are largely resolved with the new per-voxel analysis.

One disagreement that I have is with the characterization of Antonello et al., for which I am the first author. In their rebuttal the authors' state:

It is challenging to compare our study with the research conducted by Antonello et al. in 2024, as they vary the contextual window size to search for the optimal window size for each voxel. In their study, they introduced a new concept called "dynamically sized context window" (p. 3, in section 2.3), which means that the size of the context window changes throughout the narrative, ranging from 1 to 256 tokens, depending on the token index. This approach differs from the fixed context window used in our study and other similar works in the literature. You can find more information about this in the equation on page 3. Thus, we are unsure if there is a genuine conflict between the results of the two papers.

While it is true that we use a context window that is not fixed-size, the model is not tuned to search for the optimal window size per voxel. The window size presented to every voxel is the same, growing from 1 token at the start of the story to 512, and then reverting back to 256 tokens at the 513th token, etc. This was mostly a concession to limit the number of forward passes through the LLM, by doing it this way, only one forward pass is needed per 256 tokens, instead of per fixed-length context window. I think it is more likely that the difference between the performance of the models has more to do with the lack of a FIR model to model the hemodynamic lag in the BOLD signal. Most of the BOLD signal is a response to the semantics of what happened several seconds ago and not the present, so not modelling with explicit lags is going to have a negative impact on the model's performance. Another possible reason for the difference is encoding performance is the smaller amount of per-subject data in this work. I do not view the lack of an explicit FIR model as entirely problematic, because it would make interpreting the effect of context harder and that is explicitly what this paper is setting out to do.

The remainder of the comments are minor and largely addressed. I believe the results in this paper are noteworthy enough and encouraging of future research. I would recommend the paper for acceptance to Nature Communications.

Signed,

Richard Antonello

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Incremental Accumulation of Linguistic Context in Artificial and Biological Neural Networks

- Responses to the Reviewers -

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Thank you for addressing the comments and for providing additional experiments involving the construction of encoding models per subject per voxel, which have resolved some of my concerns. However, some concerns still remain.

- Recent LLMs like Longformer, a Transformer variant, use a modified attention mechanism for handling longer contexts through a combination of local and global attention. Local attention focuses on incremental contexts, where each token attends only to its nearby context within a predefined window (e.g., 512 tokens). As the model processes longer documents, it builds representations incrementally from these segments. Global attention targets key tokens (e.g., special markers or task-specific tokens), allowing them to attend to all other tokens, thereby capturing broader context and integrating critical information. A study by Aw and Toneva (2023) used these deep language models for longer input contexts, showing improved predictions in language regions identified by the Fedorenko lab. Hence, it is crucial to compare these incremental LLM representations with varying context lengths to the incremental-context LLM in this paper.

Please note that we did test the Longformer model (as well as other attention-based models for long context, such as LongT5 and Transformer-XL) with various context lengths, and the results are presented in Supplementary Figure 2 (Figure S2; referred to in the main paper on page 8, at the end of the second paragraph). The results indicate that, although these models perform slightly better than vanilla transformers—as shown by Aw & Tenova (2023)—they still fail to encode long-term contextual information in most higher-order (long timescale) language brain areas, compared to our novel incremental model. Moreover, in the discussion section, we thoroughly explore the differences between our novel incremental model and these models, emphasizing the importance of the serial-incremental mechanism over the vanilla attention-based mechanism.

In general, the study by Aw & Tenova (2023) is highly relevant to this work, and the comparisons between their study and ours are discussed in the revised version of our paper (see our response to the next comment). However, a major difference between

our fMRI data and that of Aw & Tenova should be taken into account when comparing the studies: in our dataset (the Narrative dataset; Nastase et al., 2020), participants were passively listening to the narratives, whereas in Aw & Tenova, participants were reading the story aloud. Since it is well established that production and comprehension elicit different neural processes, comparisons between the studies should be made with caution.

- The study by Aw and Toneva (2023) explored larger context window sizes, up to 1000 tokens, and found that brain predictions improved only for longer input contexts (20 to 1000 words), not for shorter ones (1 to 5 words). They demonstrated that training language models for deeper narrative understanding enhances predictions of human brain responses, with improvements seen as the number of input words increased up to 500. Aw and Toneva (2023) tested four language models capable of handling longer context lengths, yielding similar insights. How do the authors reconcile these findings with the observed findings in this study?

Thank you for bringing the important study by Aw & Toneva (2023) to our attention. Their results are highly relevant and fascinating!

In our view, their findings are more aligned with ours than contradictory. In our incremental model, we used prompt engineering to guide the model in integrating long-term information into short summaries and combining it with upcoming (short-term) information. Aw & Toneva presented a different method, fine-tuning the model to perform a summarization task, which ultimately enabled it to process long-term context more efficiently.

Conceptually, our prompt-based model and Aw & Toneva's fine-tuned model are not so different, as both studies converge to the conclusion that summarization is a crucial mechanism in processing long-term context. Moreover, as expected, both methods demonstrated improved brain-alignment exclusively in long contextual windows. The key advantage of our method, however, lies in its use of a single model for both short-term and long-term contextual processing, offering a cognitively plausible framework by avoiding fine-tuning across different contexts. This unified approach not only supports a direct comparison between conditions (with or without the incremental-summary module) but also aligns more closely with how cognitive mechanisms likely operate across varied context lengths. In contrast, Aw & Toneva's method employs distinct fine-tuned models, making it harder to interpret these models as analogous to a unified cognitive process. This distinction is crucial in our study, as our primary aim was to explore hypotheses related to cognitive mechanisms, beyond just optimizing brain-alignment.

Following this comment, as well as the previous one, we added the following paragraph to the discussion of the revised paper:

"...To the best of our knowledge, no study has yet systematically investigated the effect of the LLM input contextual window size on neural encoding performance with language comprehension data. It should be noted here that one recent study by Aw & Toneva²⁹ reported a similar systematic input context-size analysis, but their fMRI data involved participants who were reading the story (i.e., language production) rather than passively listening (language comprehension), as in our study. Their main findings suggest that models trained (fine-tuned) to summarize long narratives were significantly

better in neural encoding compared to baseline models, specifically for large context windows (20 to 1000 words). This method resembles our incremental-model method, as in both studies, summarization plays a crucial role in guiding the model on how to process long contexts. However, unlike their method, which involves two different models (base and fine-tuned), our prompt-based mechanism allows us to use the same model for both short-term and long-term contextual processing (i.e., with or without the incremental-summary module). This single-model approach allows us to make a direct comparison between short- and long-term contexts. Moreover, a single model that operates differently in short vs. long contexts is more cognitively plausible than two distinct (fine-tuned) models for short and long contexts. Nevertheless, future studies are needed to compare both methods, as well as to compare language comprehension and production...”

Reference:

Aw and Toneva (2023), Training language models to summarise narratives improves brain alignment, ICLR-2023.

- The authors responded that a comprehensive layer-by-layer analysis revealed a performance decline in most layers when the context exceeded 32 tokens. However, the better predictions observed when the last layers are used, remain puzzling. Most linguistic brain encoding studies indicate that middle layers generally show better predictions of brain responses. Furthermore, research on longer context models handling both local and global contexts shows that middle layers still achieve better predictions, even with context lengths exceeding 200 tokens. Thus, the authors need to explain why their findings contradict previous studies.

In Figure S4, we show that the *trend* of performance decline over 32 tokens is consistent across layers. However, this figure does not imply that the last layer is *always* necessarily better than the other layers. For example, for a window of 128 tokens, the 8th layer performs better than the last layer (this does not contradict the earlier statement, as the performance decline over 32 tokens is observed in the 8th layer as well). In the main paper, we choose to focus on the results obtained from the last layer, as overall, this layer achieved the best performance (and, of course, the resulting pattern of performance decline in long contexts is replicated in all layers). In general, using the last layer in neural encoding is not new in this paper, but it is a common practice in other studies as well (e.g., Goldstein et al., 2022; Jain et al., 2021) and has also yielded the best results in our previous work with the same dataset (#### et al., 2023). From our impression of the literature, it seems that determining which layer encodes best is delicate and highly sensitive to multiple methodological factors, such as the type of neural encoding model, the evaluation metric, and the input length applied to the model (as suggested in our Figure S4). Moreover, it is possible that the type of dataset is crucial, as, to our knowledge, the studies that demonstrated better performance in the middle layers (e.g., Caucheteux et al., 2022; Aw & Tenova, 2023; Zaid et al., 2023) focused on language production rather than comprehension, making comparison difficult. This issue of last vs. middle layers should be carefully addressed in future work, which is beyond the scope of the current study.

- On Figure 2, why is the prediction performance in the A1 (primary auditory cortex) and STG regions similar to that in high-level language regions, including the IFG (inferior frontal gyrus) and TPJ (temporoparietal junction)? Prior studies have consistently shown that representations from language models better predict brain responses in higher-order language regions than in early sensory regions like A1. However, in this case, the prediction performance is comparable in both A1 and the late-stage language regions. What could be the reason for this unexpected result?

It is unclear which aspect of the figure the reviewer is referring to and in what way the results of A1 are "similar" to the higher areas. Please note that the y-axes in Figure 2b are different for each row, and as you can see, the overall scores in the primary areas are about 0.11, while in the higher-order areas they reach 0.14. Also note that for large context windows (Figure 2a; above 128), only the higher-order areas show significant prediction scores, not the primary areas.

Minor Comments

- Could the authors explain how a single token corresponds to 0.75 words?

This comes from the fact that, on average, a word in English is broken by the tokenizer into approximately 1.3 tokens (OpenAI, 2024; <https://platform.openai.com/tokenizer>), which means that a single token is roughly equivalent to about 0.75 words.

- The authors write "the human brain processes the incoming linguistic input serially, word by word, as speech and text unfold over time"
If the authors want to make a statement only for listening then I suggest removing the word "text" in "speech and text unfold over time".

Thank you!

We omitted the word "text".

Reviewer #3 (Remarks to the Author):

Thanks for the interesting read. I have been brought on as a replacement for Reviewer #2 due to their unavailability, and have been asked to recommend whether the revisions proposed by the authors have successfully met Reviewer #2's initial concerns. Based on my understanding of Reviewer #2's concerns, I believe that they mostly have. Here is a point-by-point address of the questions posed by Reviewer #2 and the author's responses.

Thank you! We deeply appreciate your willingness to review our manuscript.

The reviewer's largest concern was due to the averaging of their results across ROIs and participants. This lowered the quality of the signal and reduced the reliability of the results by not acknowledging and accounting for high variation in per-subject responses. The reviewers resolved this by training independent encoding models to predict the BOLD signal at the per-voxel level. I am somewhat surprised by their

comment that this took a month, but nevertheless, this concern seems to be largely addressed.

Thank you.

The reviewer was concerned that the permutation test proposed by the authors was unhelpful for comparing the performance of different models. Their move to a Wilcoxon signed rank test should resolve this concern.

Thank you.

The reviewer's concerns about many areas, such as angular gyrus, not being predicted better are largely resolved with the new per-voxel analysis.

Thank you.

One disagreement that I have is with the characterization of Antonello et al., for which I am the first author. In their rebuttal the authors' state:

It is challenging to compare our study with the research conducted by Antonello et al. in 2024, as they vary the contextual window size to search for the optimal window size for each voxel. In their study, they introduced a new concept called "dynamically sized context window" (p. 3, in section 2.3), which means that the size of the context window changes throughout the narrative, ranging from 1 to 256 tokens, depending on the token index. This approach differs from the fixed context window used in our study and other similar works in the literature. You can find more information about this in the equation on page 3. Thus, we are unsure if there is a genuine conflict between the results of the two papers.

While it is true that we use a context window that is not fixed-size, the model is not tuned to search for the optimal window size per voxel. The window size presented to every voxel is the same, growing from 1 token at the start of the story to 512, and then reverting back to 256 tokens at the 513th token, etc. This was mostly a concession to limit the number of forward passes through the LLM, by doing it this way, only one forward pass is needed per 256 tokens, instead of per fixed-length context window.

Thank you for the clarification! We understand that your model was not used to search for the optimal window size per voxel. Nevertheless, our main claim—that using non-fixed sizes across the narrative may cause differences in the results—still holds.

I think it is more likely that the difference between the performance of the models has more to do with the lack of a FIR model to model the hemodynamic lag in the BOLD signal. Most of the BOLD signal is a response to the semantics of what happened several seconds ago and not the present, so not modelling with explicit lags is going to have a negative impact on the model's performance. Another possible reason for the difference is encoding performance is the smaller amount of per-subject data in this work. I do not view the lack of an explicit FIR model as entirely problematic, because it would make interpreting the effect of context harder and that is explicitly what this paper is setting out to do.

Very interesting point! It is true that the FIR model is usually not taken into account in the neural encoding field, and it definitely should be. However, we cannot consider the lack of the FIR model as an alternative explanation for the differences between the models, as the lag of the signal (whether it is applied or not) is fixed across conditions (i.e., model context sizes). Therefore, while neglecting the FIR model may influence overall performance, it does not account for the *differences* observed between the models.

The remainder of the comments are minor and largely addressed. I believe the results in this paper are noteworthy enough and encouraging of future research. I would recommend the paper for acceptance to Nature Communications.

Thank you!

Signed,

Richard Antonello