COGNITIVE NEUROSCIENCE

Predicting the next sentence (not word) in large language models: What model-brain alignment tells us about discourse comprehension

Shaoyun Yu[1]*, Chanyuan Gu[1], Kexin Huang[1], Ping Li[1,2]*

Current large language models (LLMs) rely on word prediction as their backbone pretraining task. Although word prediction is an important mechanism underlying language processing, human language comprehension occurs at multiple levels, involving the integration of words and sentences to achieve a full understanding of discourse. This study models language comprehension by using the next sentence prediction (NSP) task to investigate mechanisms of discourse-level comprehension. We show that NSP pretraining enhanced a model's alignment with brain data especially in the right hemisphere and in the multiple demand network, highlighting the contributions of nonclassical language regions to high-level language understanding. Our results also suggest that NSP can enable the model to better capture human comprehension performance and to better encode contextual information. Our study demonstrates that the inclusion of diverse learning objectives in a model leads to more human-like representations, and investigating the neurocognitive plausibility of pretraining tasks in LLMs can shed light on outstanding questions in language neuroscience.

INTRODUCTION

Recent advances in generative artificial intelligence (AI) have put large language models (LLMs) under the spotlight. The impressive performance of LLMs arises from pretraining the models on large-scale text data and representing words and meanings as high-dimensional vectors (or "embeddings"). An increasing number of neurocognitive studies have begun to explore how model embeddings can capture human brain activities during language processing (*1*–*4*), and some argue that the rise of LLMs has enabled us to test the neural mechanisms of language learning and representation in a more principled and explicit way (*5*, *6*). Recently, a number of researchers (*7*) have advocated that connecting the study of computational models and the brain through "representational alignment" will promote knowledge transfer between the AI and the neurocognitive research communities.

To relate computational models to the brain, a link between model embeddings and brain signals must be established through methods such as linear regression or representational geometry analysis (*6*, *8*). In this study, we refer to this general approach as "model-brain alignment." To find out what computational properties of language models are relevant to the processing mechanisms in the brain, researchers compare how well the embeddings from different models align with brain data. Specifically, this means that we can test variations in the model and their relevance to human brain processes. Two computational principles of language comprehension have been proposed in the literature. First, contextual information is represented in the brain during language comprehension as it is in the model. Goldstein *et al.* (*4*) provided the key evidence that contextualized embeddings from GPT-2 (*9*) outperformed static embeddings from GloVe (*10*) in model-brain alignment (*4*). Second, word prediction is a core process of language comprehension as it has been implemented in the model (*4*, *11*). Most state-of-the-art LLMs are trained by either the next word prediction task (i.e., predicting the next word from the previous context) or the masked language modeling (MLM) task (i.e., predicting masked words from both the left and right context, akin to the cloze test). Several large-scale model comparisons have found a strong correlation between a model's word prediction ability and its alignment with brain data (*12*, *13*).

Despite remarkable developments in the literature, several gaps exist in the study of model-brain alignment. To begin with, the word prediction tasks used by LLMs have a very different goal from humans who, instead of just identifying the best candidate based on statistics of words, process and integrate words and sentences to achieve an understanding of discourse (or a spoken conversation involving multiple people) (*14*–*16*). This multilayered nature of human language comprehension is shown in recent findings that the brain predicts multiple ranges and levels of language representations (*17*). Further, the human language system also interacts with other cognitive systems and serves a communicative function (*6*). A second major gap is the lack of communication between the natural language processing (NLP) and the neuroscience of language research communities (*7*). While NLP studies aim to improve model performance on various standardized benchmarks, they generally do not consider insights from neurocognitive findings, with a few exceptions (*2*, *18*). Similarly, few neurocognitive studies have been interested in studying models that vary in pretraining tasks while investigating brain mechanisms, despite that how LLMs learn language representations through pretraining (*19*, *20*) could inform the learning and representation underlying the linguistic brain.

Given the above gaps, the current study asks whether we can leverage LLMs to study discourse comprehension, which is an area where recent investigations in NLP research and the neuroscience of language can be brought together (Fig. 1A). Discourse comprehension is critical to human communication and knowledge acquisition: Whether it be conversations, reading texts, or listening to speeches, we construe the meanings of language at the discourse level (i.e., across multiple sentences) rather than at the individual

[1]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China. [2]Centre for Immersive Learning and Metaverse in Education, The Hong Kong Polytechnic University, Hong Kong SAR, China.
*Corresponding author. Email: shaoyun.yu@polyu.edu.hk (S.Y.); ping2.li@polyu.edu.hk (P.L.)

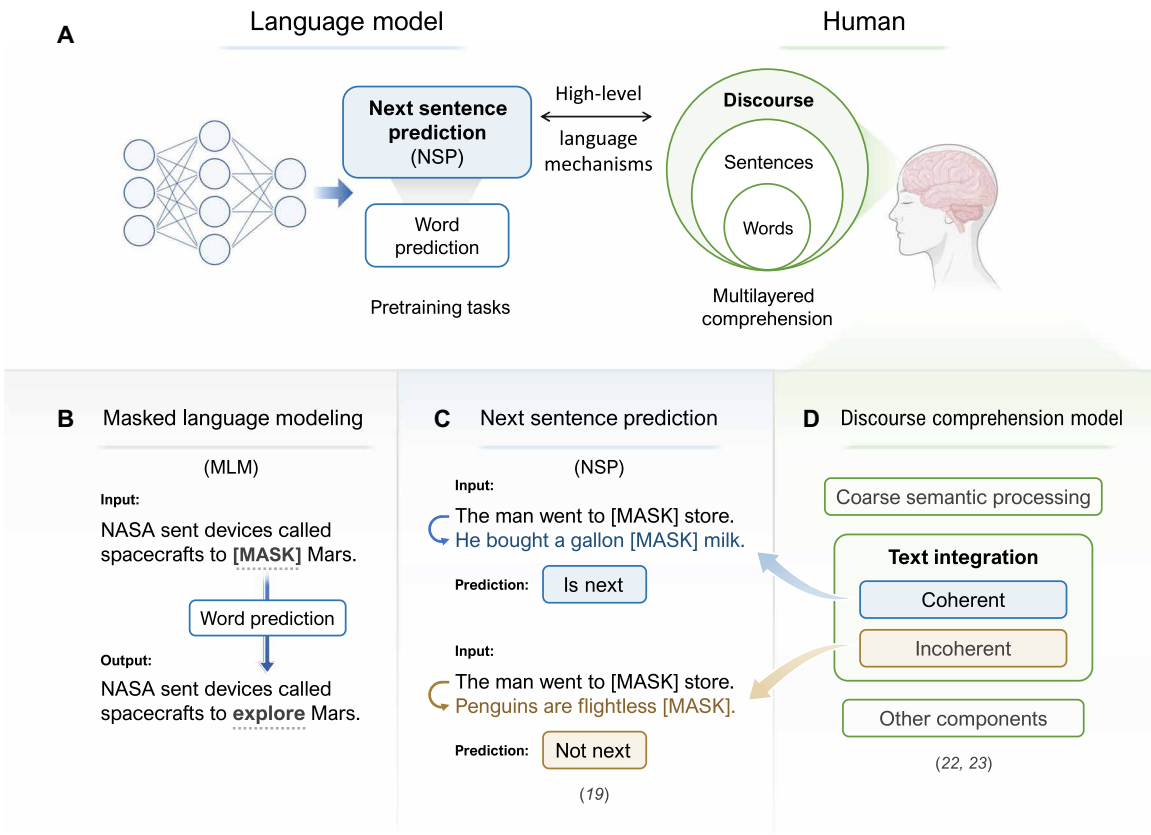


Fig. 1. NSP as a computational account of discourse comprehension. (**A**) Humans integrate words and sentences to achieve a full understanding of discourse. In LLMs, the NSP task proposed by BERT (*19*) can serve as a computational account of human discourse comprehension. (**B**) Illustration of the MLM task. (**C**) Illustration of the NSP task and its relevance to the Mason and Just model (*23*). (**D**) Illustration of Mason and Just's neurocognitive model of discourse processing. Yang *et al.* (*22*) explicitly labeled two key aspects in this model as "coherent text integration" and "incoherent text integration" (originally "text integration" and "coherence monitoring"). The human head illustration in (A) was created with BioRender.com.

word or sentence level (*14*, *21*). A central process of discourse comprehension is text integration, which requires the understanding of the coherence between sentences (*16*, *22*, *23*). To investigate discourse processing mechanisms, neurocognitive studies often contrast the brain responses to coherent sentences and those to incoherent/unconnected sentences (*22*–*26*). As a popular transformer (*27*) model, BERT (*19*) introduced the next sentence prediction (NSP) task along with the well-known MLM task to enhance the model's understanding of sentence relationships (Fig. 1, B and C); the NSP task also uses pairs of coherent or unconnected sentences/texts during pretraining. Notably, this task does not predict the content of the next sentence per se. Instead, NSP predicts whether the second sentence is truly the one next to the first sentence (i.e., one that naturally follows it), demanding the model to distinguish between coherent and unconnected pairs of sentences or texts.

There is recent evidence that using NSP pretraining in BERT models substantially improves the models' performance on discourse-level NLP tasks (*28*, *29*). However, NSP's contribution to general model performance has been in question (*30*); many recent BERT-based models even dropped this task from pretraining (*31*, *32*). From a neurocognitive perspective, the NSP task, in addition to the word prediction task, may serve as a good computational principle for how humans process and understand discourse. Mason and Just (*23*) proposed a model of discourse comprehension in the human brain (Fig. 1D) that includes two types of text integration, one for integrating coherent sentences and the other for integrating unconnected or incoherent sentences. As NSP enables the model to understand what sentence pairs are coherent and what pairs are not, it directly maps onto the two types of integration in the discourse comprehension model.

Despite decades of research, the brain networks and the hemispheric division of labor for discourse comprehension remain less well understood. From the perspective of large-scale brain networks (*33*), the classical left-lateralized language network is known to play an indispensable role in processing words and sentences (*34*); however, recent findings suggest that this network might not be sensitive to the coherence of sentences (*24*). The role of the domain-general multiple demand (MD) network (*35*, *36*) in language and discourse comprehension is also under debate (*24*, *37*, *38*). Traditionally, neurocognitive studies have focused on the left hemisphere (LH) for lexical-semantic processing. In discourse comprehension and language learning, however, the right hemisphere (RH) has been suggested to play an important role (*16*, *23*, *39*–*44*); for example, patients with RH damage often showed difficulties in discourse-level understanding (*45*, *46*). Mason and Just's Parallel Networks of Discourse model (Fig. 1D) proposed five components of discourse comprehension, which also highlighted the contribution of the RH: for example, while coherent text integration was hypothesized to

recruit a left-lateralized brain network, incoherent text integration was considered to engage a bilateral dorsolateral prefrontal network. In coarse semantic processing (*39*, *47*), the RH has been hypothesized to represent semantics on a coarse and more global scale to facilitate higher-level language understanding. However, a recent large-scale meta-analysis (*25*) suggested that text integration only consistently involves the LH, and thus, it remains unclear whether and to what extent the RH is involved in discourse comprehension.

In this study, we focus on leveraging different pretraining tasks to better align LLMs with the human language system, and by using model-brain alignment, we hope to gain insights into how the brain processes discourse. Specifically, we test NSP as a plausible computational mechanism for discourse comprehension and explore brain networks that correspond to this mechanism. To this end, we built two BERT-based deep language models (DLMs) that manipulated the presence of NSP in pretraining and used two functional magnetic resonance imaging (fMRI) datasets that emphasized coherent and unconnected sentence relationships, respectively. Model-brain alignment performance was examined in the language network and the MD network for both hemispheres. Overall, our results showed that LLMs and the brain converged better on high-level language mechanisms beyond word prediction.

RESULTS

Two fMRI datasets about sentence reading were used in this study: the Mars subset of the Reading Brain project (*48*–*50*) in which the sentences are connected to make a coherent story and the dataset from Pereira *et al.* (*1*) in which the stimuli are dominated by unconnected relationships (see Materials and Methods for details). We refer to the two datasets as the "Reading-brain2019 dataset" and the "Pereira2018 dataset." To identify the contributions of NSP to model-brain alignment, we trained two types of models using the BERT architecture: the MLM model that performed only the MLM task and the MLM_NSP model that performed both the MLM (for word prediction) and NSP (for sentence coherence prediction/evaluation) tasks. All other training procedures for the two models were kept identical. To estimate model-brain alignment, we used representational similarity analysis (RSA) (*51*) as the alignment function (*7*), which evaluates the correlation between the model and the brain's representational spaces (see Materials and Methods for details). The brain networks that may reflect NSP's computational mechanism were revealed by examining brain regions that displayed higher alignment with the MLM_NSP model than with the MLM model. An illustration of our overall approach is shown in Fig. 2.

NSP-pretrained model displayed greater model-brain alignment in the language and MD networks

We provided our models with each stimulus sentence from the two datasets and extracted their embeddings. Model-brain alignment was computed for both the MLM_NSP and MLM models. We examined the models' differences in regions of interest (ROIs) from two major brain networks: the language network (*34*) (10 ROIs) and the domain-general MD network (*35*, *52*) (20 ROIs). One-sided Wilcoxon signed-rank tests were performed because of our hypothesis that the MLM_NSP model would better align with the brain and also because of the lack of normal distribution in this type of correlational results. False discovery rate (FDR; $\alpha = 0.05$) correction for multiple comparisons was applied to ROIs from the same brain network. The results from the Reading-brain2019 and the Pereira2018 datasets suggest that a pretraining process that combined NSP and MLM, as compared with pretraining only based on MLM, significantly increased the model's alignment with brain data in the comprehension of both coherent sentences and unconnected sentences.

Reading-brain2019 dataset

The sentences in this dataset consisted of a coherent narrative text about humans going to Mars (see Materials and Methods for details). As illustrated in Fig. 3A, we found that the MLM_NSP model showed significantly higher model-brain alignment than the MLM model in four language network ROIs: the bilateral (left and right) inferior frontal gyri (IFG), the right orbital part of the IFG (IFGorb), and the right anterior temporal gyrus (ATG), $P_{FDR} = 0.038$ for these regions. These ROIs can be considered to have a greater correspondence with NSP's computational mechanism, which involves the understanding of sentence coherence. No significant model differences were found in the MD network (see table S1 for details). Our results suggest that the language network is critically engaged in comprehending coherent sentences, contrasting the null results reported by Jacoby and Fedorenko (*24*). The IFG areas are classic language processing regions (*53*, *54*), whereas the bilateral ATG is a major hub for conceptual and semantic integration (*55*). Further, while the ROIs in the LH of the language network are considered the classical "core" for language processing, the ROIs identified in our results as having higher model-brain alignment reside more in the RH homologs (i.e., the right IFG, IFGorb, and ATG) rather than in the LH regions, suggesting a crucial role of the RH in discourse comprehension.

Pereira2018 dataset

This dataset consisted of 384 sentences dominated by unconnected relationships (see Materials and Methods for details). The sentences were organized into 96 unconnected sentence groups about various topics, with each group consisting of four locally coherent sentences. As illustrated in Fig. 3B, we found that the MLM_NSP model showed significantly higher model-brain alignment than the MLM model in five MD network ROIs. Two ROIs were from the LH: the left middle frontal gyrus (MFG) and the left anterior cingulate cortex/presupplementary motor cortex (ACC/pSMA); three ROIs were from the RH: the right superior frontal gyrus (SFG), right orbital part of the MFG (MFGorb), and right precentral gyrus (PrecG), $P_{FDR} = 0.039$ for the right MFGorb, $P_{FDR} = 0.029$ for the other four ROIs. These ROIs can be considered to have a stronger association with NSP's computational mechanism. However, unlike with the Reading-brain2019 dataset, no significant model differences were found in the language network (see table S2 for details). Our results thus showed that, despite recent debates surrounding the role of the MD network in language comprehension (*24*, *37*, *38*), multiple frontal MD network regions in both hemispheres were implicated in the comprehension of unconnected/incoherent sentences. This finding partially overlaps with Mason and Just's proposal (*23*) that the integration of incoherent sentences recruits a bilateral dorsolateral prefrontal brain network (Fig. 3C).

NSP-pretrained model captured individual differences in reading time

A recent study (*56*) found that model-brain alignment computed with GPT-2 could predict subjects' listening comprehension scores. To test the association between reading performance and model-brain alignment, we performed Pearson correlation analyses for individual

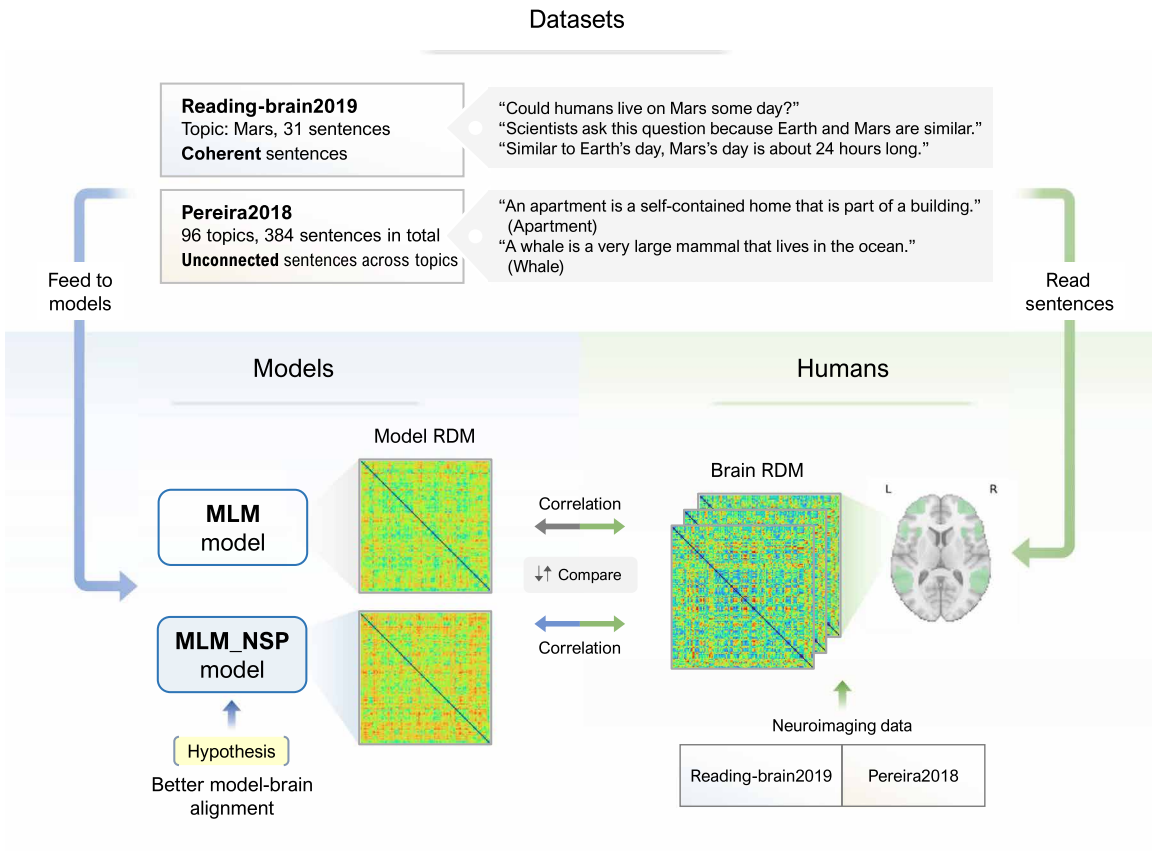


Fig. 2. Overview of the datasets, computational models, and analysis. To investigate model-brain alignment in discourse comprehension, we pretrained two BERT-based models and used two neuroimaging datasets. To extract model-based sentence representations, we fed the two types of models with the same sentences read by human subjects. RSA was used to evaluate the correlation between model embeddings and brain activation.

ROIs from the language and MD networks, respectively, for the model with both MLM and NSP pretraining and the model with MLM pretraining only. We tested the significance of the correlation coefficients for the MLM_NSP versus the MLM model and applied FDR correction ($\alpha = 0.05$) to ROIs from the same brain network. Here, we only report the findings for the Reading-brain2019 dataset as the Perreira2018 dataset did not contain any reading performance data (reading time or accuracy).

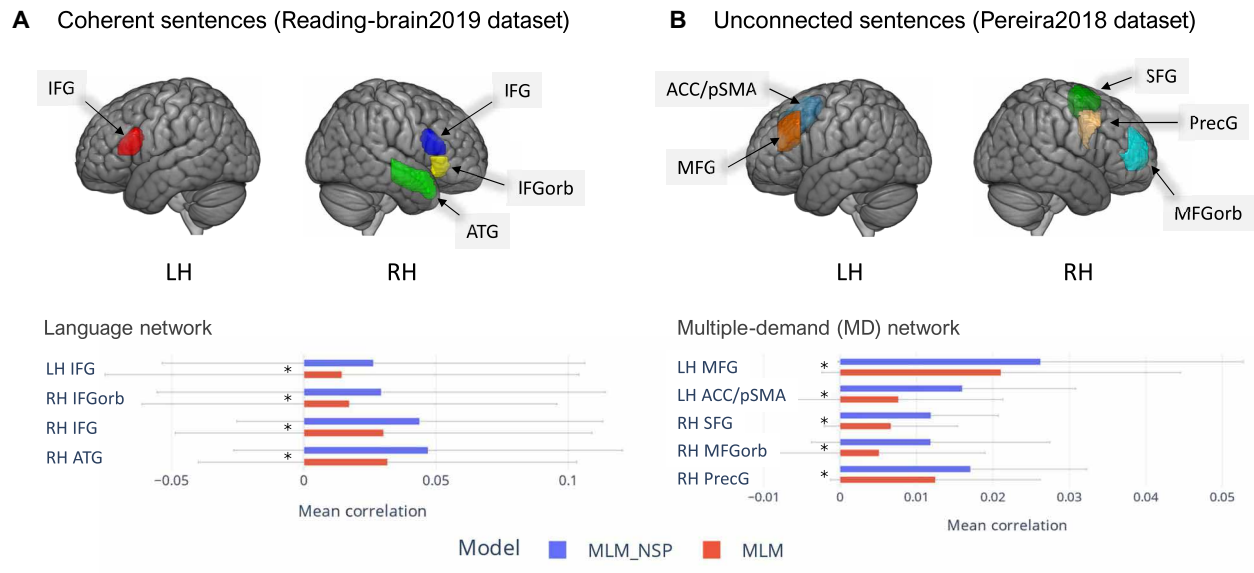
Our findings indicated that reading time was negatively correlated with model-brain alignment in all language ROIs, except the left MFG, for both the MLM_NSP and the MLM models; for the left MFG, only the MLM_NSP pretraining model showed significant correlation (Fig. 4, A and C). Likewise, reading time was negatively correlated with model-brain alignment in the ROIs from the MD network (Fig. 4B) for both types of models, except for the right superior parietal lobule (SPL), where only the MLM_NSP pretraining model showed significant correlation (Fig. 4, B and C). However, in neither type of model did we find significant correlations between reading accuracy and model-brain alignment. These findings demonstrate that model-brain alignment computed with the MLM_NSP and MLM models were both sensitive to reading efforts indexed by reading time rather than reading outcomes indexed by accuracy. Further, model-brain alignment derived from the MLM_NSP model displayed a slight advantage in capturing individual variations in reading time, manifested as the significant correlations in all ROIs from the language and MD networks.

NSP-pretrained model consistently performed better with different context lengths

To investigate how model-brain alignment could be modulated by the contextual window available to LLMs (*2*, *57*, *58*), we presented the two types of models with both the stimulus sentences and their prior context. At each context length, we compared the model-brain alignment performance between the MLM_NSP and MLM models.

Reading-brain2019 dataset

We varied each stimulus sentence's context length from one to seven preceding sentences, which were equal to 10 to 70 words on average. As displayed in Fig. 5A, the impact of context length followed a non-monotonic pattern that peaks at a short-range length. Specifically, for both the MLM_NSP and MLM models, the alignment between the model and brain representational dissimilarity matrices (RDMs) quickly increased within a context length of one to two preceding sentences (an average of 10 to 20 words), and then there was an overall decline as the contextual window expanded further beyond 20 words on average. This advantage of short-range context echoes Toneva and Wehbe's finding (*2*) based on a story-reading dataset, which also demonstrated an effect of short context length for their BERT model. Our study also showed that the MLM_NSP model consistently outperformed the MLM model: The four language network



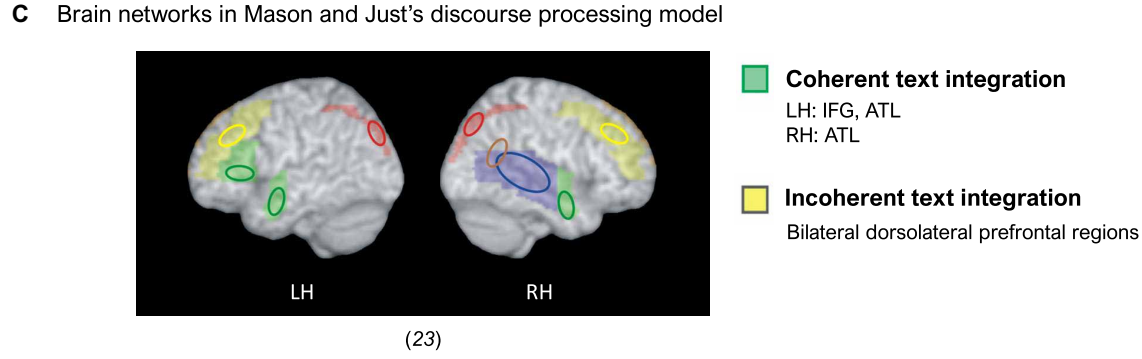


Fig. 3. NSP-pretrained model displayed higher alignment with brain data. NSP significantly improved model-brain alignment for both coherent and unconnected sentence relationships in discourse-level comprehension. (**A**) In the Reading-brain2019 dataset (coherent sentences), four language network ROIs displayed significantly higher alignment with the MLM_NSP model. (**B**) In the Pereira2018 dataset (mainly unconnected sentences), five MD network ROIs displayed significantly higher alignment with the MLM_NSP model. The two models' performances were compared with one-sided Wilcoxon signed-rank tests. Asterisk (*) indicates statistical significance after FDR correction ($\alpha = 0.05$). Error bars indicate SEs. (**C**) Brain networks in Mason and Just's discourse comprehension model, adapted from figure 1 of (*23*).

ROIs identified when no preceding sentences were given (bilateral IFG, right IFGorb, and right ATG) continued to demonstrate a significant advantage of the MLM_NSP model across a range of context lengths, from one to seven sentences. The MLM_NSP model's advantage also extended to most other ROIs in the language and MD networks (see fig. S1 for details), suggesting that the MLM_NSP model benefited more from the prior context compared to the MLM model.

Pereira2018 dataset

Context lengths were limited to a max of three sentences before the stimulus sentence because this dataset was composed of various four-sentence groups (see Materials and Methods for details); contextual windows across different sentence groups were not available due to the lack of trial order information. Therefore, for each stimulus sentence, we only included its preceding sentences from the same group as its context (henceforth, the "group local context"), which varied from one to three sentences, or averagely 11.8 to 35.4 words. As shown in Fig. 5B, model-brain alignment did not substantially vary with the length of group local context for both the MLM_NSP and MLM models; there were only minor fluctuations in alignment, which can also be described as a nonmonotonic pattern. Such a near-flat pattern suggests that increasing group local context did not necessarily improve model-brain alignment when the whole sentence set was dominated by unconnected relationships. The results resonate with Caucheteux and King's (*12*) observation that varying context lengths did not significantly affect model-brain alignment in a dataset where all sentences were unconnected to each other. Despite this lack of context length effect, we observed that the MLM_NSP model had significantly higher model-brain alignment than the MLM model in several MD network regions, including the left SPL, the right SFG, the bilateral MFG, MFGorb, ACC/pSMA, and PrecG (Fig. 5B). Such advantages of the MLM_NSP model were less widespread compared to the results from coherent sentences (Fig. 5A), demonstrating again that the effect of context length was much more limited for unconnected sentences. Further, the MLM_NSP model's advantage was exclusively found in the MD network rather than in the language network (see fig. S2 for details), which is consistent with

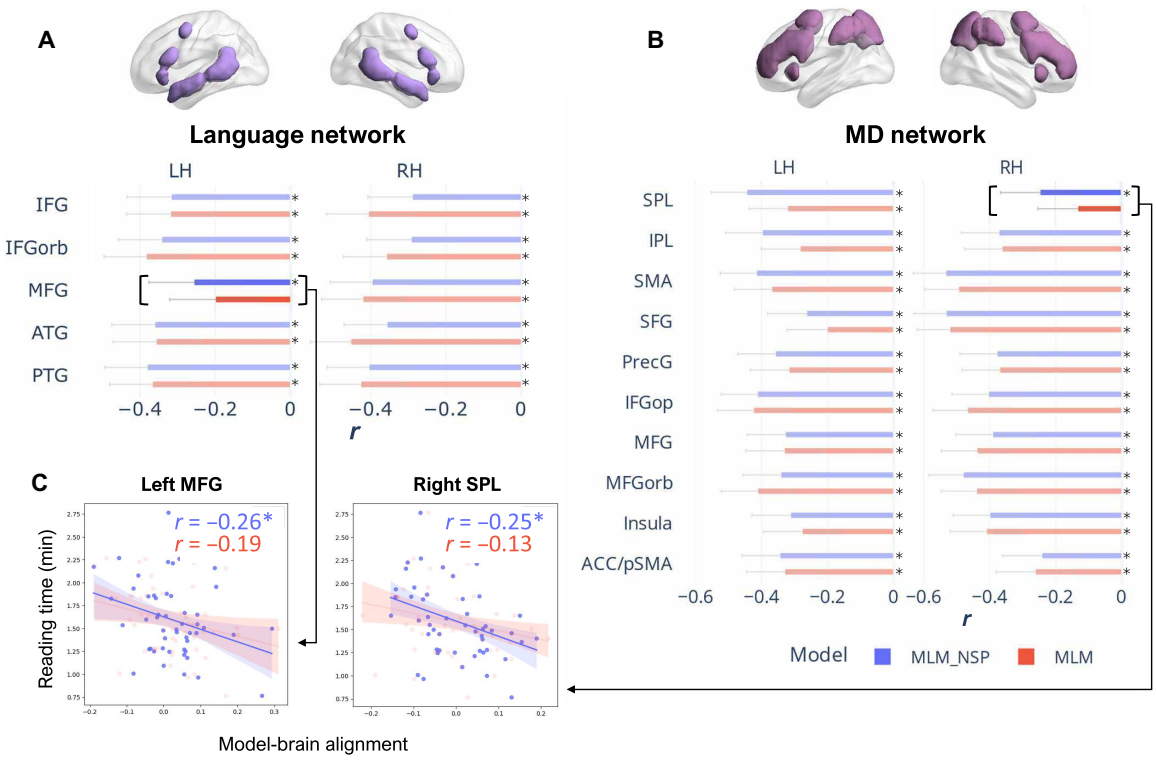


Fig. 4. Correlations between reading time and model-brain alignment. In the Reading-brain2019 dataset, reading time was negatively correlated with model-brain alignment computed from the MLM_NSP and MLM models. (**A**) Estimated correlations for ROIs from the language network. (**B**) Estimated correlations for ROIs from the MD network. The bars represent correlation estimates, and the gray lines indicate SEs. (**C**) Estimated correlations in the left MFG and the right SPL where the correlation between reading time and model-brain alignment was significant for the MLM_NSP model (blue) rather than the MLM (red) model. Asterisk (*) indicates statistical significance after FDR correction ($\alpha = 0.05$).

our suggestion that the MD network plays a critical role in processing unconnected sentences.

DISCUSSION

How can human brain research and AI inform one another in the era of generative AI and LLMs? Here, we advocate the examination of model-brain alignment as an approach to studying brain mechanisms of language processing using models with diverse pretraining tasks, taking advantage of machine learning–based language models and representational alignment analytics to link models and brain data (*7*, *59*).

While word prediction has been validated as a critical computational principle in human language processing (*4*, *12*, *13*), successful discourse comprehension for humans requires multilevel comprehension beyond the word level, and as such, a single computational process (e.g., word prediction) is unlikely to fully account for human language comprehension (*6*, *59*). In this regard, while the word prediction pretraining task has proven to be highly effective for LLMs, it does not match the multilevel processes of human language processing if we aim at computational models that are brain-inspired. In our study, we examined the NSP task in addition to word prediction, an LLM pretraining task that requires the model to classify whether the input sentences or texts are coherent or not. We found that NSP significantly improves a model's alignment with brain data with respect to the correlations between the model and the brain's representations. In neurocognitive models of discourse comprehension, such evaluation of sentence coherence is essential to the understanding of discourse or conversation (*22*, *23*). Thus, we regarded NSP as a viable computational account of discourse-level comprehension. By examining the roles of the brain regions showing greater model-brain alignment with NSP-enhanced versus MLM-only pretraining, our results also enabled us to explore the neural correlates of NSP. Through two datasets that focused on either coherent or unconnected sentences, we found that the language network and the domain-general MD network had differential contributions to discourse comprehension according to the coherence of sentences. Further, our results highlighted the importance of RH brain regions in discourse comprehension.

The exact role of the language network (*34*) in discourse comprehension has been under debate. For example, while previous findings indicate that the processing of coherent texts engages the RH language regions (*46*, *60*), a recent study (*24*) did not observe greater brain activations during the reading of coherent sentences compared with unconnected sentences. Our results provide more support to the view that RH language regions are critically engaged in processing coherent texts, given the model-brain alignment shown in the reading of coherent sentences from the Reading-brain2019 dataset. The MLM_NSP model displayed significantly higher alignment with brain data than the MLM model in three language ROIs from the RH in addition to the left IFG: the right IFG, IFGorb, and ATG. Notably, we found higher model-brain alignment in the right

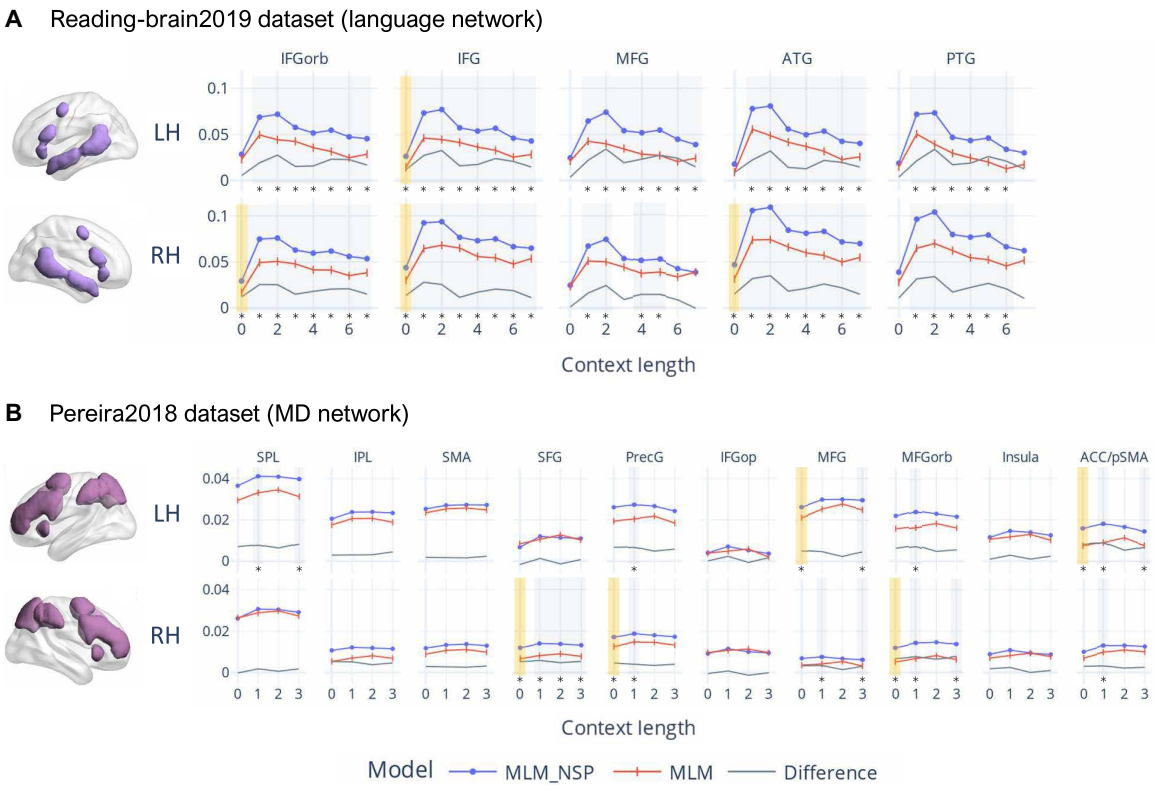


Fig. 5. Model-brain alignment as a function of context length. Increasing the context length available to the models affected model-brain alignment in a nonmonotonic way, and there was evidence that the MLM_NSP model performed better than the MLM model for both coherent and unconnected sentence relationships. (**A**) Effect of context length for the Reading-brain2019 dataset (coherent sentences) in the language network. (**B**) Effect of context length for the Pereira2018 dataset (mainly unconnected sentences) in the MD network. Asterisk (*) indicates statistical significance after FDR correction ($\alpha = 0.05$). Yellow areas indicate ROIs identified when no preceding sentences were given. Gray areas indicate ROIs that showed a significant advantage of the MLM_NSP model at different context lengths.

IFG and IFGorb in the naturalistic reading of coherent sentences, while Jacoby and Fedorenko (*24*) found greater activation in these two regions for their unconnected sentences condition. This discrepancy points to an important difference between the contrast-based experimental fMRI approach and the model-based approach in our study (*61*, *62*). The former approach relies on contrasting and analyzing well-controlled conditions, so it may overlook the involvement of brain regions if the regions are active but do not show statistically stronger levels of activation compared to the contrasted condition. By comparison, the latter approach as used in our work can more directly test whether neural responses in a specific brain region correspond to the engagement of a specific computational mechanism. The significantly higher alignment with the MLM_NSP model in the right IFG and IFGorb suggests that these language network regions are relevant to the understanding of coherent sentences in discourse-level comprehension.

In the past decade, the domain-general MD network has attracted continued attention and also generated debates regarding its role in language comprehension. Some recent studies have argued that the MD network is not essential for language comprehension (*24*, *37*, *38*), while others have suggested that certain MD network regions are reliably involved in discourse comprehension, such as the right SPL (*25*). With the Pereira2018 dataset, our results supported this latter line of argument by showing higher model-brain alignment with the MLM_NSP model in the MD network. It is worth noting that three of the MD network regions found in the current study (i.e., the left MFG, right MFGorb, and right PrecG) overlapped with the MD regions showing stronger activation to unconnected sentences reported by Jacoby and Fedorenko (*24*). However, our results do not lead to an interpretation to attribute the MD network activation to task-induced efforts (*23*, *38*). The MD regions in our results were isolated by comparing the MLM_NSP and MLM models; therefore, activities in these regions should be interpreted as relevant to NSP's computational mechanism, that is, the evaluation of sentence relationships. Consequently, we suggest that the MD network is genuinely engaged in discourse comprehension, particularly for understanding the relationship of unconnected sentences. Future research should examine the impact of text properties such as the sentences' lexical and syntactic structural relationships (*63*) in addition to coherence in modulating model-brain alignment.

Mason and Just's Parallel Networks of Discourse model (*23*) proposed that the integration of coherent sentences recruits a left-lateralized brain network, while the integration of unconnected or incoherent sentences recruits a bilateral dorsolateral prefrontal network. By synthesizing the above findings from the Reading-brain2019 and Pereira2018 datasets, our results indicate consistent recruitment of the RH in integrating both coherent and unconnected sentences. Our findings regarding the integration of unconnected sentences largely agreed with Mason and Just's account, as suggested by higher alignment with the MLM_NSP model in the left

MFG, right MFGorb, and right SFG. However, we also found several frontal regions, including the left ACC/pSMA and the right PrecG, which are not strictly within the dorsolateral prefrontal scope. Our findings regarding the integration of coherent sentences, however, differed from Mason and Just's proposal that focused on the role of the left IFG and ATG; instead, we found a more right-lateralized network, including the RH homologs of the IFG, IFGorb, and ATG. This finding does not mean that the left language network is not engaged in processing coherent sentences but that in discourse comprehension the model-brain alignment might be more sensitive to a more general, higher-level, and perhaps coarse semantic integration process supported by the RH [see (*47*) for a discussion of the division of labor for semantic processing at different levels]. Overall, our results contribute to a growing body of evidence suggesting the RH's crucial role in high-order language functions, including discourse comprehension, first and second language learning, prosody processing, and the understanding of figurative languages (*39*, *41*, *43*, *44*, *64*–*66*).

At the individual difference level, our study showed that model-brain alignment computed with the MLM_NSP and MLM models was negatively correlated with reading time, suggesting that greater alignment between brains and models may be associated with faster reading. Reading time is one of the critical components for assessing reading skills (*67*–*69*), which has been used to differentiate skilled and less skilled readers during discourse comprehension (*50*). Skilled readers, compared with less skilled readers, may be more efficient in selecting and organizing key contents to construct and integrate the mental representation, thus giving rise to quicker reading time (*50*, *70*). Our finding of the significant correlations between model-brain alignment and reading time during discourse comprehension demonstrates that LLMs may be capable of characterizing the neurocognitive map of skilled comprehension. In our modeling, we also observed that the MLM_NSP model had a small advantage in capturing reading speed. Specifically, the model-brain alignment in the left MFG and right SPL, derived from the MLM_NSP model but not the MLM model, exhibited significant correlations with reading time. We speculate that the NSP-enhanced model may be more sensitive to the underlying neural mechanism of discourse comprehension, allowing it to better capture individual reading speed. Sentence coherence is indispensable for the integration and construction of mental representation (*23*, *71*). The purpose of the NSP task is to judge whether sentences are coherent, so it may allow the MLM_NSP model to encode certain high-level information about the upcoming discourse content (*28*).

In transformer models, the attention mechanism allows each word to draw information from other words in computing contextualized embeddings unique to the input context (*27*). Such a mechanism has neurocognitive relevance as information processing in the brain is influenced by memories of the context at various timescales (*72*, *73*). In general, we found evidence that the MLM_NSP model consistently performed better than the MLM model when context was incorporated, suggesting that NSP pretraining allows a model to better use contextual information. Prior studies suggested that model-brain alignment is not a simple linear function of a model's contextual window length. Most findings demonstrated that model-brain alignment peaked or plateaued at a short context length of about 10 words (*2*, *57*, *58*). In contrast, Caucheteux and King (*12*) found that model-brain alignment was not significantly affected by context length; notably, this study used a dataset of unconnected sentences, while most previous work was based on coherent and narrative materials (e.g., podcast stories or book chapters). We leveraged the two different datasets in our study to examine the impact of context length. Our results about coherent sentences (Reading-brain2019 dataset) supported Toneva and Wehbe's study (*2*) based on a narrative text: The model-brain alignment increased within a short-range context (in our case, one to two sentences) and then decreased. Such a pattern could suggest that, when the discourse is continuous and coherent, the recent context consistently contributes to the current sentence's meaning and representation. By contrast, in the Pereira2018 dataset, we did not observe a substantial impact of group local context length (see Results for details) during unconnected sentence reading. This pattern indicated that incorporating the local context had limited contribution when the full sentence set was dominated by unconnected relationships. Combined with previous findings, we suggest that the impact of context length depends on the coherence features of the whole context, which should be further investigated.

Recent LLMs such as GPT-3 and its successors have greatly benefited from exploiting the "scaling laws" by increasing the model size (*74*). In contrast, less progress has been made in pretraining tasks. Mainstream LLMs all base their language pretraining on one task type: word prediction. In this study, we showed that the computational principle of the NSP task (i.e., sentence coherence prediction/evaluation) is neurocognitively plausible and maps onto the theoretical framework of discourse comprehension (*22*, *23*). Our results support the arguments of previous studies that NSP improves a model's discourse-level language competence (*19*, *28*, *29*). The increased model-brain alignment achieved through NSP-pretraining provides evidence that LLMs and the brain can converge on discourse-level language mechanisms rather than only sharing the core principle of word prediction.

One limitation of the current work is that our experiments were based on two research-oriented, comparatively smaller-scale models with the BERT architecture (see Materials and Methods for details). Our choices regarding the model type and size were limited by two practical considerations; first, currently BERT is the only major open-source model that proposed a cognitively plausible pretraining task beyond word prediction, and second, large-scale models such as Meta AI's LLaMA required hundreds or thousands of GPUs for pretraining, which exceeds the capability of single research labs. Nevertheless, we expect our findings to be extendable to larger-sized models because recent LLMs share the same underlying transformer architecture with BERT, and they have mostly been limited to word prediction tasks in pretraining. To evaluate the effectiveness of NSP beyond BERT-like transformer encoder models, future studies may generalize the NSP task and extend it to other transformer decoder models (e.g., GPT models). In addition, future work can also adapt alternative sentence-level training algorithms (*20*) and assess their cognitive plausibility as well as effectiveness compared to NSP. A second limitation of the current study is that our models were evaluated against two neuroimaging datasets collected from different participants. In addition to the difference in the key properties modeled in our study (i.e., coherence versus unconnectedness), the two datasets may differ on other dimensions due to data collection and processing differences beyond our control. Future studies can aim at more controlled fMRI data from a single study, although resource demands in collecting such data may be quite challenging.

Even with the above limitations and constraints, our study of model-brain alignment can shed light on NLP research so that the significance of a model and model components can be evaluated not only on NLP benchmarks but also on its neurocognitive relevance. We highlighted at the beginning the lack of communication as a major gap between NLP research and the neuroscience of language. Our findings demonstrate that pretraining methods as experimented with in NLP studies can inform neuroscientists when testing computational hypotheses about how the human brain processes and represents language. The current study shows how neurocognitive researchers can leverage LLMs to study higher-level language mechanisms by going beyond word prediction. For a true approximation of the human language system, it is important to note that human language involves more diverse and intricate mechanisms than what existing computational models and neurocognitive theories could account for. Our findings align with recent proposals that future LLMs need to embrace more modularity, diverse learning objectives, multimodal information integration of features in the external world, and integrations beyond core linguistic abilities (*6*, *7*, *59*, *75*). By showing the advantage of combining multiple levels of language pretraining, our findings are also in line with the view that hierarchical or multilevel representations might be crucial for AI to approach the efficiency and flexibility of human intelligence, which is a direction explored by recent endeavors such as the Joint Embedding Predictive Architecture (*76*). We conclude that model-brain alignment promotes a close communication of ideas and methods between the AI and the neurocognitive research communities, which will lead to future research in brain-inspired AI and AI-informed brain studies.

MATERIALS AND METHODS

Computational models

We built the MLM_NSP and MLM models with the transformers Python library. Both models were uncased and used a base BERT structure of 12 hidden layers. The MLM model was trained with only the MLM task, while the MLM_NSP model was trained with both the MLM and NSP tasks. We used the entire English Wikipedia (version 20220301, available on Hugging Face) as the pretraining dataset. The masked token ratio of the pretraining data was set to 15%. Both models were trained for 11 epochs with a learning rate of 5×10^{-5}; the amount of epochs was comparable to those used in smaller-scale research-purpose models (*77*). Figure S3 illustrates the training loss curves of the MLM and MLM_NSP models. The pretraining was performed on two NVIDIA Tesla V100S GPUs.

Neuroimaging datasets

Reading-brain2019 dataset

We used the native English speaker dataset from the Reading Brain Project (*48*–*50*), which is a multimodal naturalistic reading database that combined fMRI and eye tracking. Fifty-two right-handed native English speakers participated in reading expository texts in the fMRI scanner. In the current study, two subjects were excluded from the data analysis due to preprocessing errors. We used the subset for the Mars text (31 sentences) as it had the highest Flesch Reading Ease score, indicating that it was the easiest to comprehend (see Supplementary Text for full text). The subjects read the text sentence by sentence on the screen in the natural order. The experiment was self-paced, with an 8-s limit for each sentence. Multiband echo-planar imaging data were acquired with a repetition time of 400 ms. The dataset is available on OpenNeuro (https://openneuro.org/datasets/ds003974/).

Pereira2018 dataset

We used the experiment 2 dataset from the Pereira *et al.* study (*1*) and chose the nine subjects analyzed by Schrimpf *et al.* (*13*). The total 384 sentences consisted of 96 four-sentence groups about 96 different concepts (e.g., elephant and farm). The 96 concepts can be further grouped into 24 broader categories (e.g., animal and place). All sentences were written in an expository style. The subjects read the stimuli on the screen sentence by sentence at a fixed pace, with each trial consisting of a 4-s display followed by a 4-s interval. The 96 groups of sentences were presented randomly for each subject; within each sentence group, the order of the four sentences was fixed. All stimuli were repeated three times across three scanning sessions. Data availability and additional details are given in (*1*) and (*13*).

Neuroimaging data processing

Reading-brain2019 dataset

The fMRI data were preprocessed with fMRIPrep 22.0.0 (*78*). Subjects' structural images (T1-weighted) were corrected for field inhomogeneity, and brain tissue was extracted and segmented. Subjects' functional images were corrected for head motion and slice time. The functional images were then coregistered to their structural reference images. Confound time series were estimated on the basis of the processed blood-oxygen-level-dependent signals. The functional images were resampled into the Montreal Neurological Institute (MNI) space using the MNI ICBM 152 nonlinear sixth asymmetric template (*79*). The brain activation (beta maps) for each sentence was estimated with general linear models (GLMs) in Nilearn, and the least squares single modeling method was applied (*80*). The canonical SPM hemodynamic response function was used to model brain responses. The GLMs also included confound regressors for head motion, white matter, cerebrospinal fluid, and mean global signal. The high-pass filter was set at 100 s.

Pereira2018 dataset

We used the beta values for the stimulus sentences precomputed by Pereira *et al.* The data processing procedures are detailed in (*1*).

Statistical analysis

Representational similarity analysis

We used RSA (*51*) to evaluate the correspondence between model embeddings and brain activation. We computed the brain-based RDMs (henceforth, brain RDMs) using each subject's brain activation elicited by the stimulus sentences (beta values). We generated model embedding-based RDMs (henceforth, model RDMs) by computing pairwise distances between sentences in the embedding spaces of our models.

Model RDMs. We fed each individual sentence from the two datasets to our custom-trained models and averaged the sentence's token embeddings from the 12th hidden layer (the final hidden layer) as the sentence-level representation. The special tokens [CLS] and [SEP] were not included in averaging. We then generated the model RDMs by computing the pairwise cosine distance between the sentence representations for each dataset. In our investigation of context length effects, we provided the models with the stimulus sentence and its preceding sentences, allowing the models to draw information from the prior context in generating embeddings. We obtained sentence-level representations by averaging the token

embeddings from the stimulus sentence, and then we computed the model RDMs using these sentence representations.

Brain RDMs. We computed brain RDMs per subject and per ROI for the two neuroimaging datasets. For each ROI, beta values were extracted to represent brain activation elicited by the stimulus sentences. RDMs were built by computing 1 − Pearson's *r* for all sentence pairs within each dataset. We used 10 fronto-temporal ROIs (5 per hemisphere) defined in the language-selective network (*34*), including the left IFG, IFGorb, MFG, ATG, PTG, and their RH homologs; we chose 20 fronto-parietal ROIs (10 per hemisphere) from the MD network (*52*), including the bilateral SPL, IPL, SMA, SFG, PrecG, IFG (pars opercularis), MFG, MFGorb, insula, and ACC/pSMA. The mask images for these ROIs are available at https://evlab.mit.edu/funcloc/. For the Reading-brain2019 dataset, we extracted the ROI voxels by using the predefined group-level ROI masks for the two brain networks. For the Pereira2018 dataset, the ROIs were determined individually by combining localizer tasks and group-level masks (*13*), and we obtained the ROI voxels from the precomputed result data provided by Pereira *et al.* (*1*).

Model-brain alignment was evaluated with Pearson's correlations between the model RDM and the subjects' brain RDMs (*81*). As we have an a priori hypothesis that the MLM_NSP model would exhibit higher model-brain alignment, we performed one-sided Wilcoxon signed-rank tests (MLM_NSP > MLM). FDR correction was applied to the *P* values for ROIs from the same brain network.

Correlation between reading performance and model-brain alignment

We performed two-sided Pearson correlation analyses to estimate the association between reading performance and model-brain alignment. Our focus was on the Reading-brain2019 dataset due to its inclusion of individual performance data, whereas the Pereira2018 dataset did not provide such information. The Reading Brain project used 10 questions to evaluate the understanding of expository text reading and recorded self-paced reading time and accuracy for each participant (*48*). Model-brain alignment was computed for the ROIs from the language and MD networks. To control for multiple comparisons, we applied FDR correction ($\alpha = 0.05$) for ROIs from the same brain network and reported the corrected results.

Terminology for language models and pretraining tasks

We refer to BERT as an instance of LLM in a broad sense. When viewed more specifically, BERT can be defined as a neural language model (NLM). From a technical standpoint, LLMs are on a continuum of NLMs because current LLMs share the same fundamental transformer architecture (*27*) and pretraining method (i.e., word prediction) with NLMs. In our reading of the literature and in the LLM field's rapid development, we found that the usage of NLM and LLM has become blurred. For example, Wikipedia has explicitly referred to BERT as a type of LLM (https://en.wikipedia.org/wiki/Large_language_model, accessed on 22 February 2024). The situation becomes even more complicated when some researchers use DLM to refer to BERT and GPT series models (*4*). We think that the terms NLM, DLM, and LLM tend to highlight different aspects of these language models (neural net-based, deep-layered, or large-sized, respectively), and researchers are using them interchangeably (sometimes not technically correct).

Following the convention in the literature (*12*, *13*, *19*), we refer to the next word prediction and the MLM tasks as word prediction tasks, although technically they predict subword tokens instead of complete words. Language models internally work with subword tokens instead of words for efficiency and generalizability considerations (e.g., to handle rare words by combining subword tokens known to the model); the segmentation of a word into tokens is determined by the tokenizer algorithm, and tokens do not necessarily correspond to linguistic units such as letters or syllables.

The next word prediction task is more formally known as the language modeling (LM) task in NLP studies (*9*). The LM (i.e., next word prediction) task predicts the upcoming word based on the preceding context; the task is unidirectional in that it always uses the one-sided context (e.g., left context in left-to-right languages) to predict the next word. The MLM task is closely related to the LM task (*19*), and it can predict masked words (hence "masked" in the name) in any position instead of just the last/next word. MLM is bidirectional in that the task uses both the left and right context to predict the masked word. Despite the differences, both LM (next word prediction) and MLM are word prediction tasks (*12*).

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S3
Tables S1 and S2

REFERENCES AND NOTES

1. F. Pereira, B. Lou, B. Pritchett, S. Ritter, S. J. Gershman, N. Kanwisher, M. Botvinick, E. Fedorenko, Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* **9**, 963 (2018).
2. M. Toneva, L. Wehbe, *"Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)" in Proceedings of the 33rd International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2019).
3. A. J. Anderson, D. Kiela, J. R. Binder, L. Fernandino, C. J. Humphries, L. L. Conant, R. D. S. Raizada, S. Grimm, E. C. Lalor, Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *J. Neurosci.* **41**, 4100–4119 (2021).
4. A. Goldstein, Z. Zada, E. Buchnik, M. Schain, A. Price, B. Aubrey, S. A. Nastase, A. Feder, D. Emanuel, A. Cohen, A. Jansen, H. Gazula, G. Choe, A. Rao, C. Kim, C. Casto, L. Fanda, W. Doyle, D. Friedman, P. Dugan, L. Melloni, R. Reichart, S. Devore, A. Flinker, L. Hasenfratz, O. Levy, A. Hassidim, M. Brenner, Y. Matias, K. A. Norman, O. Devinsky, U. Hasson, Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).
5. P. Contreras Kallens, R. D. Kristensen-McLachlan, M. H. Christiansen, Large language models demonstrate the potential of statistical learning in language. *Cognit. Sci.* **47**, e13256 (2023).
6. S. Arana, J. Pesnot Lerousseau, P. Hagoort, Deep learning models to study sentence comprehension in the human brain. *Lang. Cogn. Neurosci.* **1**, 1–19 (2023).
7. I. Sucholutsky, L. Muttenthaler, A. Weller, A. Peng, A. Bobu, B. Kim, B. C. Love, E. Grant, J. Achterberg, J. B. Tenenbaum, K. M. Collins, K. L. Hermann, K. Oktar, K. Greff, M. N. Hebart, N. Jacoby, Qiuyi, Zhang, R. Marjieh, R. Geirhos, S. Chen, S. Kornblith, S. Rane, T. Konkle, T. P. O'Connell, T. Unterthiner, A. K. Lampinen, K.-R. Müller, M. Toneva, T. L. Griffiths, Getting aligned on representational alignment. arXiv arXiv:2310.13018 [Preprint] (2023). http://arxiv.org/abs/2310.13018.
8. T. M. Mitchell, S. V. Shinkareva, A. Carlson, K. M. Chang, V. L. Malave, R. A. Mason, M. A. Just, Predicting human brain activity associated with the meanings of nouns. *Science* **320**; 1191–1195 (2008).
9. A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, "Improving language understanding by generative pre-training" (2018). https://openai.com/research/language-unsupervised.
10. J. Pennington, R. Socher, C. Manning, "GloVe: Global vectors for word representation" in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Doha, Qatar, 2014; http://aclweb.org/anthology/D14-1162), pp. 1532–1543.
11. R. Ryskin, M. S. Nieuwland, Prediction during language comprehension: What is next? *Trends Cogn. Sci.* **27**, 1032–1052 (2023).
12. C. Caucheteux, J.-R. King, Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5**, 134 (2022).

13. M. Schrimpf, I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. B. Tenenbaum, E. Fedorenko, The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2105646118 (2021).
14. U. Hasson, G. Egidi, M. Marelli, R. M. Willems, Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition* **180**, 135–157 (2018).
15. E. C. Ferstl, J. Neumann, C. Bogler, D. Y. Von Cramon, The extended language network: A meta-analysis of neuroimaging studies on text comprehension. *Hum. Brain Mapp.* **29**, 581–593 (2008).
16. P. Li, R. B. Clariana, Reading comprehension in L1 and L2: An integrative approach. *J. Neurolinguistics* **50**, 94–105 (2019).
17. C. Caucheteux, A. Gramfort, J.-R. King, Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat. Hum. Behav.* **7**, 430–441 (2023).
18. E. Chersoni, E. Santus, C.-R. Huang, A. Lenci, Decoding word embeddings with brain-based semantic features. *Comput. Linguist.* **47**, 663–698 (2021).
19. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "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)* (Association for Computational Linguistics, Minneapolis, Minnesota, 2019; https://aclanthology.org/N19-1423), pp. 4171–4186.
20. D. Iter, K. Guu, L. Lansing, D. Jurafsky, "Pretraining with contrastive sentence objectives improves discourse performance of language models" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Online, 2020; https://aclweb.org/anthology/2020.acl-main.439), pp. 4859–4870.
21. M. L. van Moort, D. D. Jolles, A. Koornneef, P. van den Broek, What you read versus what you know: Neural correlates of accessing context information and background knowledge in constructing a mental representation during reading. *J. Exp. Psychol. Gen.* **149**, 2084–2101 (2020).
22. X. Yang, H. Li, N. Lin, X. Zhang, Y. Wang, Y. Zhang, Q. Zhang, X. Zuo, Y. Yang, Uncovering cortical activations of discourse comprehension and their overlaps with common large-scale neural networks. *Neuroimage* **203**, 116200 (2019).
23. R. A. Mason, M. A. Just, "Neuroimaging contributions to the understanding of discourse processes" in *Handbook of Psycholinguistics* (Elsevier, 2006; https://linkinghub.elsevier.com/retrieve/pii/B9780123693747500201), pp. 765–799.
24. N. Jacoby, E. Fedorenko, Discourse-level comprehension engages medial frontal Theory of Mind brain regions even for expository texts. *Lang. Cogn. Neurosci.* **35**, 780–796 (2020).
25. X. Yang, N. Lin, L. Wang, Situation updating during discourse comprehension recruits right posterior portion of the multiple-demand network. *Hum. Brain Mapp.* **44**, 2129–2141 (2023).
26. E. C. Ferstl, D. Y. von Cramon, What does the frontomedian cortex contribute to language processing: Coherence or theory of mind? *Neuroimage* **17**, 1599–1612 (2002).
27. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need" in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2017) *NIPS'17*, pp. 6000–6010.
28. W. Shi, V. Demberg, "Next sentence prediction helps implicit discourse relation classification within and across domains" 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)* (Association for Computational Linguistics, Hong Kong, China, 2019; https://aclweb.org/anthology/D19-1586), pp. 5789–5795.
29. F. Koto, J. H. Lau, T. Baldwin, Discourse probing of pretrained language models. arXiv arXiv:2104.05882 [Preprint] (2021). http://arxiv.org/abs/2104.05882.
30. A. Rogers, O. Kovaleva, A. Rumshisky, A primer in BERTology: What we know about how BERT works. *Trans. Assoc.Comput. Linguist.* **8**, 842–866 (2020).
31. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach. arXiv arXiv:1907.11692 [Preprint] (2019). http://arxiv.org/abs/1907.11692.
32. K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, MPNET: Masked and permuted pre-training for language understanding. arXiv arXiv:2004.09297 [Preprint] (2020). http://arxiv.org/abs/2004.09297.
33. B. T. Thomas Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, B. Fischl, H. Liu, R. L. Buckner, The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).
34. E. Fedorenko, P.-J. Hsieh, A. Nieto-Castañón, S. Whitfield-Gabrieli, N. Kanwisher, New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *J. Neurophysiol.* **104**, 1177–1194 (2010).
35. J. Duncan, The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends Cogn. Sci.* **14**, 172–179 (2010).
36. E. Fedorenko, The role of domain-general cognitive control in language comprehension. *Front. Psychol.* **5**, (2014).
37. L. Wehbe, I. A. Blank, C. Shain, R. Futrell, R. Levy, T. Von Der Malsburg, N. Smith, E. Gibson, E. Fedorenko, Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. *Cereb. Cortex* **31**, 4006–4023 (2021).
38. E. Diachek, I. Blank, M. Siegelman, J. Affourtit, E. Fedorenko, The domain-general multiple demand (MD) network does not support core aspects of language comprehension: A large-scale fMRI investigation. *J. Neurosci.* **40**, 4536–4550 (2020).
39. M. Jung-Beeman, Bilateral brain processes for comprehending natural language. *Trends Cogn. Sci.* **9**, 512–518 (2005).
40. J. M. Zacks, E. C. Ferstl, "Discourse comprehension" in *Neurobiology of Language* (Elsevier, 2016), pp. 661–673.
41. P. Hagoort, The neurobiology of language beyond single-word processing. *Science* **366**, 55–58 (2019).
42. C. S. Prat, D. L. Long, K. Baynes, The representation of discourse in the two hemispheres: An individual differences investigation. *Brain Lang.* **100**, 283–294 (2007).
43. C. S. Prat, J. Gallée, B. L. Yamasaki, Getting language right: Relating individual differences in right hemisphere contributions to language learning and relearning. *Brain Lang.* **239**, 105242 (2023).
44. Z. Qi, J. Legault, "Neural hemispheric organization in successful adult language learning: Is left always right?" in *Psychology of Learning and Motivation* (Elsevier, 2020) vol. 72, pp. 119–163.
45. H. Brownell, G. Martino, Deficits in inference and social cognition: The effects of right hemisphere brain damage in *Right Hemisphere Language Comprehension: Perspectives from Cognitive Neuroscience*. (Lawrence Erlbaum, 1998), pp. 309–328.
46. M. Beeman, E. M. Bowden, M. A. Gernsbacher, Right and left hemisphere cooperation for drawing predictive and coherence inferences during normal story comprehension. *Brain Lang.* **71**, 310–336 (2000).
47. B. Schloss, P. Li, Disentangling narrow and coarse semantic networks in the brain: The role of computational models of word meaning. *Behav. Res.* **49**, 1582–1596 (2017).
48. P. Li, C.-T. Hsu, B. Schloss, A. Yu, L. Ma, M. Scotto, F. Seyfried, C. Gu, The Reading Brain project L1 adults, OpenNeuro (2022). https://doi.org/doi:10.18112/openneuro.ds003974.
49. C.-T. Hsu, R. Clariana, B. Schloss, P. Li, Neurocognitive signatures of naturalistic reading of scientific texts: A fixation-related fMRI study. *Sci. Rep.* **9**, 10678 (2019).
50. X. Ma, Y. Liu, R. Clariana, C. Gu, P. Li, From eye movements to scanpath networks: A method for studying individual differences in expository text reading. *Behav. Res.* **55**, 730–750 (2023).
51. N. Kriegeskorte, Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Sys. Neurosci.* **2**, 4 (2008).
52. E. Fedorenko, J. Duncan, N. Kanwisher, Broad domain generality in focal regions of frontal and parietal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 16616–16621 (2013).
53. P. Hagoort, L. Hald, M. Bastiaansen, K. M. Petersson, Integration of word meaning and world knowledge in language comprehension. *Science* **304**, 438–441 (2004).
54. A. D. Friederici, The role of left inferior frontal and superior temporal cortex in sentence comprehension: Localizing syntactic and semantic processes. *Cereb. Cortex* **13**, 170–177 (2003).
55. M. A. Lambon Ralph, E. Jefferies, K. Patterson, T. T. Rogers, The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.* **18**, 42–55 (2017).
56. C. Caucheteux, A. Gramfort, J.-R. King, Deep language algorithms predict semantic comprehension from brain activity. *Sci. Rep.* **12**, 16327 (2022).
57. S. Jain, A. G. Huth, "Incorporating context into language encoding models for fMRI" in *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2018) *NIPS'18*, pp. 6629–6638.
58. S. Abnar, L. Beinborn, R. Choenni, W. Zuidema, "Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains" in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (Association for Computational Linguistics, Florence, Italy, 2019; https://aclweb.org/anthology/W19-4820), pp. 191–203.
59. K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, E. Fedorenko, Dissociating language and thought in large language models *Trends Cogn. Sci.* (2024); https://doi.org/10.1016/j.tics.2024.01.011.
60. M. A. Gernsbacher, M. P. Kaschak, Neuroimaging studies of language production and comprehension. *Annu. Rev. Psychol.* **54**, 91–114 (2003).
61. S. Jain, V. A. Vo, L. Wehbe, A. G. Huth, Computational language modeling and the promise of in silico experimentation. *Neurobiol. Lang.* **5**, 1–27 (2024).
62. U. Hasson, S. A. Nastase, A. Goldstein, Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron* **105**, 416–434 (2020).
63. M. J. Pickering, V. S. Ferreira, Structural priming: A critical review. *Psychol. Bull.* **134**, 427–459 (2008).
64. P. Li, H. Jeong, The social brain of language: Grounding second language learning in social interaction. *npj Sci. Learn.* **5**, 8 (2020).

65. R. L. C. Mitchell, Right hemisphere language functions and schizophrenia: The forgotten hemisphere? *Brain* **128**, 963–978 (2005).
66. S. Kita, O. De Condappa, C. Mohr, Metaphor explanation attenuates the right-hand preference for depictive co-speech gestures that imitate actions. *Brain Lang.* **101**, 185–197 (2007).
67. M. Krieber, K. D. Bartl-Pokorny, F. B. Pokorny, C. Einspieler, A. Langmann, C. Körner, T. Falck-Ytter, P. B. Marschik, The relation between reading skills and eye movement patterns in adolescent readers: Evidence from a regular orthography. *PLOS ONE* **11**, e0145934 (2016).
68. R. Cohen-Mimran, R. Yifat, K. Banai, Size matters? Rapid automatized naming of shape sizes, reading accuracy and reading speed. *J. Res. Read.* **44**, 882–896 (2021).
69. S. P. McGeown, L. G. Duncan, Y. M. Griffiths, S. E. Stothard, Exploring the relationship between adolescent's reading skills, reading motivation and reading habits. *Read. Writ.* **28**, 545–569 (2015).
70. K. Krstić, A. Šoškić, V. Ković, K. Holmqvist, All good readers are the same, but every low-skilled reader is different: An eye-tracking study using PISA data. *Eur. J. Psychol. Educ.* **33**, 521–541 (2018).
71. W. Kintsch, T. A. van Dijk, Toward a model of text comprehension and production. *Psychol. Rev.* **85**, 363–394 (1978).
72. J. L. Elman, Finding structure in time. *Cognit. Sci.* **14**, 179–211 (1990).
73. U. Hasson, J. Chen, C. J. Honey, Hierarchical process memory: Memory as an integral component of information processing. *Trends Cogn. Sci.* **19**, 304–313 (2015).
74. J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models. arXiv arXiv:2001.08361 [Preprint] (2020). http://arxiv.org/abs/2001.08361.
75. M. Mitchell, AI's challenge of understanding the world. *Science* **382**, eadm8175 (2023).
76. Y. LeCun, A Path Towards Autonomous Machine Intelligence Version 0.9.2, 2022-06-27. Open Review (2022).
77. A. Alajrami, N. Aletras, "How does the pre-training objective affect what large language models learn about linguistic properties?" in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Association for Computational Linguistics, Dublin, Ireland, 2022; https://aclanthology.org/2022.acl-short.16), pp. 131–147.
78. O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, H. Oya, S. S. Ghosh, J. Wright, J. Durnez, R. A. Poldrack, K. J. Gorgolewski, fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
79. A. C. Evans, A. L. Janke, D. L. Collins, S. Baillet, Brain templates and atlases. *Neuroimage* **62**, 911–922 (2012).
80. J. A. Mumford, B. O. Turner, F. G. Ashby, R. A. Poldrack, Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage* **59**, 2636–2643 (2012).
81. D. J. Acunzo, D. M. Low, S. L. Fairhall, Deep neural networks reveal topic-level representations of sentences in medial prefrontal cortex, lateral anterior temporal lobe, precuneus, and angular gyrus. *Neuroimage* **251**, 119005 (2022).

Acknowledgments: We thank F. Pereira for sharing of the Pereira2018 database for the modeling and analysis reported in this study. We are also grateful to the Sin Wai Kin Foundation Endowed Professorship in Humanities and Technology to P.L. for supporting this research. S.Y. has been supported by a Research Postdoctoral Fellowship, and C.G. and K.H. have been supported by Research Postgraduate Scholarships from the Hong Kong Polytechnic University. Figure 3 contains an adapted brain image from (*22*), which the original authors publicly shared at http://ccbi.cmu.edu/reprints/Mason_Figure1.tif. **Funding:** The Reading Brain Project was supported by the NSF (#NCS-FO-1533625; principal investigator: P.L.), and the work reported here is supported by the Hong Kong Research Grants Council (Project #PolyU15601520; principal investigator: P.L.). **Author contributions:** S.Y., C.G., and P.L. designed the research and wrote the manuscript. S.Y. performed model building with input from K.H., C.G., P.L., and S.Y. performed data analysis. K.H. contributed to data analysis and the revision of the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The code for pretraining the models, model weights, precomputed beta values, and sentence embeddings is available at https://osf.io/94y7h/. The Reading-brain2019 dataset is available on OpenNeuro at https://openneuro.org/datasets/ds003974/. The availability of the Pereira2018 dataset is stated in (*1*).

Submitted 1 January 2024
Accepted 18 April 2024
Published 23 May 2024
10.1126/sciadv.adn7744