npj | science of learning

Article

Published in partnership with The University of Queensland

https://doi.org/10.1038/s41539-025-00337-y

Reading comprehension in L1 and L2 readers: neurocomputational mechanisms revealed through large language models

Chanyuan Gu[1], Samuel A. Nastase[2], Zaid Zada[2] & Ping Li[1,3] ✉

While evidence has accumulated to support the argument of shared computational mechanisms underlying language comprehension between humans and large language models (LLMs), few studies have examined this argument beyond native-speaker populations. This study examines whether and how alignment between LLMs and human brains captures the homogeneity and heterogeneity in both first-language (L1) and second-language (L2) readers. We recorded brain responses of L1 and L2 English readers of texts and assessed reading performance against individual difference factors. At the group level, the two groups displayed comparable model-brain alignment in widespread regions, with similar unique contributions from contextual embeddings. At the individual level, multiple regression models revealed the effects of linguistic abilities on alignment for both groups, but effects of attentional ability and language dominance status for L2 readers only. These findings provide evidence that LLMs serve as cognitively plausible models in characterizing homogeneity and heterogeneity in reading across human populations.

Language comprehension is one of the main means through which humans acquire world knowledge. Successful comprehension requires the understanding of the meanings of words within a given context[1]. However, it is challenging to study word comprehension in the context of naturalistic language comprehension, due to the absence of explicit linguistic models. To address these challenges, in this study we leverage advances in generative artificial intelligence and large language models (LLMs) through the lens of 'shared computational principles' for natural language processing in humans and machines[2–7].

LLMs use multi-layer neural network architectures and self-supervised learning algorithms to learn the statistical structure of natural language from large-scale text or visual corpora. These models project acquired structure onto high-dimensional vectors, often called "contextual embeddings", which encapsulate linguistic nuances of individual words relevant to the context. For example, the word 'bank' will be assigned different sets of contextual embeddings to reflect its varied linguistic nuances in different contexts (riverbank or financial institution). The principle of context-specific embedding, a key principle shared by human brains and LLMs, posits that preceding context is actively engaged in word comprehension during language processes[4,5]. Support for this principle has documented the superior performance of contextual embeddings in predicting brain activity in higher-order (e.g., prefrontal and temporal cortices) and lower-order auditory/visual regions[2,4,5,8]. Researchers argue that LLMs can serve as a unified theoretical framework to shed light on the neurocomputational mechanisms of human language comprehension.

So far, most existing neuroimaging work has applied LLMs to only L1 speakers, that is, individuals whose first/dominant language aligns with the target language used in the task[2–6]. Given that more than half of the world's population is bilingual[9], there is an urgent need to extend the development of LLMs to the study of L2 speakers. In light of this, our study addresses two significant gaps in the literature: (1) whether LLMs can capture the neurobiological mechanisms of language comprehension across L1 and L2 populations, and (2) whether LLM-derived measures can reflect individual differences in human language processing.

In the bilingualism literature, there has been a long-standing debate about the mechanisms underlying L1 and L2 language processes. Some suggest that L2 speakers can never achieve native-like comprehension due to the fundamental differences in language acquisition[10,11], while others argue

[1]Department of Language Science and Technology, The Hong Kong Polytechnic University, Hong Kong SAR, China. [2]Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ, USA. [3]Centre for Immersive Learning and Metaverse in Education and PolyU-Hangzhou Technology and Innovation Research Institute, The Hong Kong Polytechnic University, Hong Kong SAR, China. ✉e-mail: ping2.li@polyu.edu.hk

that L2 speakers can show native-like behaviors or brain responses if provided with the right learning environments[12]. Neuroimaging studies have shown largely overlapping brain activation during L1 and L2 reading, depending on the readers' L2 proficiency, linguistic competency, and cognitive abilities[13–15].

Nonetheless, current neuroimaging work has several limitations. First, traditional studies, averaging brain responses across entire reading sessions, do not reflect the brain responses associated with the encoding of word meanings, due to the absence of explicit linguistic models. Second, many studies do not have a baseline group of L1 speakers against which the comparison of L1 and L2 speakers can be evaluated. These limitations suggest that evaluating the alignment between LLMs and the brain may provide novel insights into the debate about language comprehension among L1 and L2 readers. Furthermore, such cross-group comparisons can inform the development of LLMs by testing their applicability as cognitively plausible models across populations (Fig. 1a).

In addition, it is so far unclear to what extent individual differences modulate the neurobiological mechanisms of language comprehension within the framework of model-brain alignment. While the proposal of 'shared computational principles' hinges on the mechanisms shared by LLMs and humans, it is agnostic to substantial variations in human language processing. Explaining heterogeneity in human reading demands a theoretical framework that incorporates individual differences in linguistic and cognitive abilities[1,16]. Behavioral studies have documented the effects of linguistic and cognitive abilities, such as vocabulary size and attentional ability, on reading comprehension across L1 and L2 readers[17–20]. Furthermore, bilingual language processes may recruit additional cognitive resources due to reduced automaticity and L1 interference[15,18,21]. Despite ample behavioral evidence, the impacts of individual differences on the neural mechanisms underlying L2 reading remain under-investigated.

Language experience for every L2 reader is unique and dynamic. Neurocognitive studies suggest that language experience, such as language

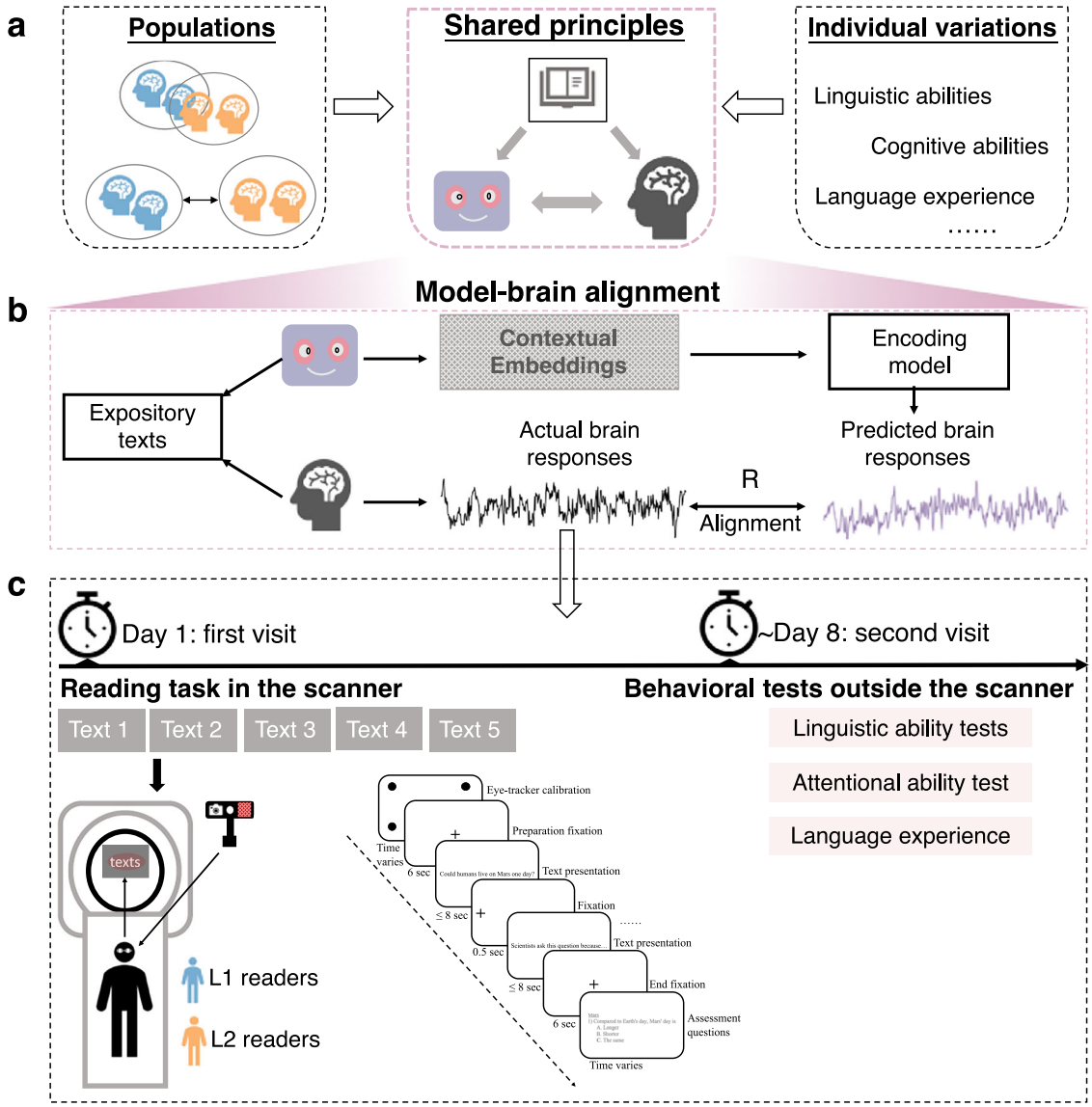

Fig. 1 | Model-brain alignment and experimental procedure. a Contributions of diverse populations and individual differences to the proposal of 'shared computational principles' between LLMs and human language processing. **b** Model-brain alignment approach. We constructed participant-specific encoding models using LLM-based features and computed alignment between LLMs and human brains by correlating predicted brain responses with actual brain responses (i.e., *R* scores). **c** Experimental procedure. During the first visit, we acquired simultaneous fMRI and eye-tracking while participants underwent a self-paced reading task, consisting of five expository scientific texts. Following each text, participants answered a battery of comprehension questions. During the second visit, participants completed a battery of behavioral tests that evaluated their linguistic competency, attentional ability, and language experience.

dominance, may interact with expertise in linguistic and cognitive abilities, jointly impacting brain responses[22,23]. Prior neuroimaging work has documented the effects of L2 readers' language experience during reading[24]. However, how individual language experience and expertise (i.e., linguistic and cognitive abilities) jointly affect reading comprehension remains unclear.

While extant literature has documented individual variations in human language processing[15,25], it remains largely unknown whether the "model-brain alignment" approach, leveraging the power of LLMs, can capture this heterogeneity. To address this gap, this work aims to probe into how model-brain alignment is shaped by individual differences by collecting data on readers' linguistic abilities, cognitive abilities, and language experience (Fig. 1a).

In this study, we extend the application of LLMs to study L1 and L2 readers by testing the principle of context-specific embedding, along with the effects of individual differences. To quantify alignment between LLMs and human brains, we constructed LLM-based encoding models (see Fig. 1b and "Methods") and tested this principle by evaluating the performance of contextual embeddings in predicting brain responses across populations. As an extension, we estimated the association between reading performance and model-brain alignment, as well as the impacts of individual differences. Toward this goal, we simultaneously recorded eye movements and brain activity using fMRI during a naturalistic reading paradigm (see Fig. 1c). Finally, we assessed expertise (i.e., linguistic and cognitive abilities) and L2 dominance.

Based on the proposal of 'shared computational principles'[2,4–6], we made the following predictions: (1) humans and LLMs would show significant alignment in distributed brain regions across L1 and L2 readers, with contextual embeddings outperforming other linguistic features in predicting brain responses; (2) model-brain alignment would correlate with reading outcomes across both groups: if L1 readers exhibit better reading outcomes, greater alignment in L1 readers compared with L2 readers may be observed; if the two groups had comparable reading outcomes, similar alignments would be expected; (3) linguistic abilities would impact both groups, while attentional ability would impact more strongly L2 readers than L1 readers; and finally (4) language experience would affect model-brain alignment in L2 readers. Together, the model-brain alignment approach enables us to examine the similarities and differences between L1 and L2 processing at a finer granularity and to identify the origins of homogeneity and heterogeneity within a naturalistic reading framework.

Results

L2 and L1 readers displayed comparable reading outcomes

We used multiple-choice questions during the scanning session (first visit) to evaluate reading comprehension performance (see "Methods"). We also recorded the total reading time that participants spent completing the five texts. First, we found that (a) the two groups showed comparable reading accuracy ($t(102) = 1.68$, $p = 0.10$, $d = 0.33$), (b) L2 readers showed longer total reading time ($t(104) = -4.81$, $p < 0.001$, $d = -0.94$), greater number of fixations ($t(104) = -2.92$, $p = 0.004$, $d = -0.57$), and longer mean fixation duration ($t(104) = -4.90$, $p < 0.001$, $d = -0.96$) than L1 readers. Next, we evaluated group-level differences in behavioral assessments: (c) L2 readers displayed lower vocabulary size ($t(94) = 11.64$, $p < 0.001$, $d = 2.58$) and general reading ability ($t(104) = 3.41$, $p < 0.001$, $d = 0.67$), and (d) there was no significant group difference in attentional abilities, including alerting ($t(104) = -1.40$, $p = 0.17$, $d = -0.27$), executive control ($t(104) = 1.33$, $p = 0.19$, $d = 0.26$), and orienting networks ($t(104) = 0.27$, $p = 0.79$, $d = 0.05$). Details for behavioral measures were provided in Supplementary Table 2. These findings suggest that despite longer reading times, L2 readers achieved reading comprehension accuracy comparable to that of L1 readers. Moreover, the linguistic abilities of L2 readers lagged behind those of L1 readers, while the two groups had comparable attentional abilities.

Contextual embeddings captured brain responses similarly in L1 and L2 readers

To test Prediction (1) that LLM-based embeddings would align with brain responses across groups, we constructed participant-specific encoding models using the Generative Pre-trained Transformer 2 (GPT-2) language model[26]. Specifically, we quantified the model-brain alignment by extracting contextual embeddings from the eighth layer of GPT-2, in line with previous work[7,27]. Subsequently, we iteratively constructed the participant-specific encoding model using a leave-one-run-out cross-validation approach. We used four runs as the training sample and the remaining run as the test sample, ensuring that each run served as the test sample once (see "Methods"). We adopted a 1000-parcel brain atlas to partition the brain into 1000 parcels[28] and constructed the encoding model on the averaged brain responses for individual parcels. Next, we selected 12 language regions of interest (ROIs) and 6 visual ROIs (see "Methods") to estimate ROI-based alignment.

We observed significant model-brain alignment in widespread bilateral brain areas spanning visual and language in both L1 and L2 readers (Fig. 2a). For both groups, the strongest alignment was observed in bilateral lower-order visual regions and higher-order comprehension-related brain regions, such as bilateral precuneus/posterior cingulate cortex (PCC; see abbreviations in Supplementary Table 1), bilateral lateral prefrontal cortex (LPFC), left temporal gyrus (TG), right inferior parietal lobe (IPL), and right superior parietal lobe (SPL). In addition, the language and visual ROIs displayed significant alignments (Fig. 2d). There was no significant group difference in the whole-brain and ROI-based alignment (Fig. 2b–d), suggesting that brain responses were equally captured by contextual embeddings between the two groups.

These findings support Prediction (1) and provide neural evidence that LLM-based embeddings capture brain responses similarly across L1 and L2 readers during naturalistic reading. The comparable alignment across L1 and L2 readers, particularly in higher-order comprehension-related and lower-order visual regions, demonstrates that LLMs can be effectively extended to the study of L2 readers.

Models embedded with context-specific meaning showed the best model-brain alignment across L1 and L2 readers

To further test prediction (1), we examined the superiority of contextual embeddings in predicting brain responses over other linguistic features, such as part-of-speech (POS) labels that indicate a word's grammatical role vs static embeddings that represent words with fixed high-dimensional vectors independent of context. Specifically, we compared a model constructed using contextual embeddings (i.e., $R_{contextual}$; as reported in Fig. 2a) with (a) a model constructed using both contextual embeddings and POS labels (i.e., $R_{contextual\_POS}$) and (b) a model constructed using both contextual and static embeddings (i.e., $R_{contextual\_static}$). We extracted POS labels by assigning a one-hot vector of 11 features to individual words (see "Methods"). These labels do not convey the conceptual meaning but indicate the grammatical role. We extracted static embeddings from the word token embedding matrix of GPT-2 to represent the context-independent state of word representations. Unlike contextual embeddings, static embeddings capture the 'average' meaning of words independent of specific contexts.

Our analyses indicated comparable model-brain alignment between the model including contextual embeddings and the model consisting of contextual and POS labels for both groups ($ps > 0.001$; see the non-thresholded whole-brain differences in Fig. 3a). Similarly, there was no significant difference in model-brain alignment between the model including contextual embeddings and the model consisting of contextual and static embeddings for both groups, except for a significant parcel labeled PCC in L1 readers ($ps > 0.001$; see the non-thresholded whole-brain differences in Fig. 3b).

To reflect the reliance on contextual information, we further estimated the model-brain alignment uniquely contributed by contextual embeddings (R_{unique}; see "Methods"). We identified that both L1 and L2 readers showed model-brain alignment (i.e., $R_{unique} > 0$) uniquely contributed by contextual embeddings in frontoparietal and frontotemporal brain regions (i.e., $ps < 0.001$; Fig. 3c), such as bilateral precuneus/PCC, IPS, PFCl, SPL, IFG, and visual regions. In addition, we identified comparable R_{unique} between L1 and L2 readers (i.e., $ps > 0.001$; see the non-thresholded whole-brain differences in Fig. 3d).

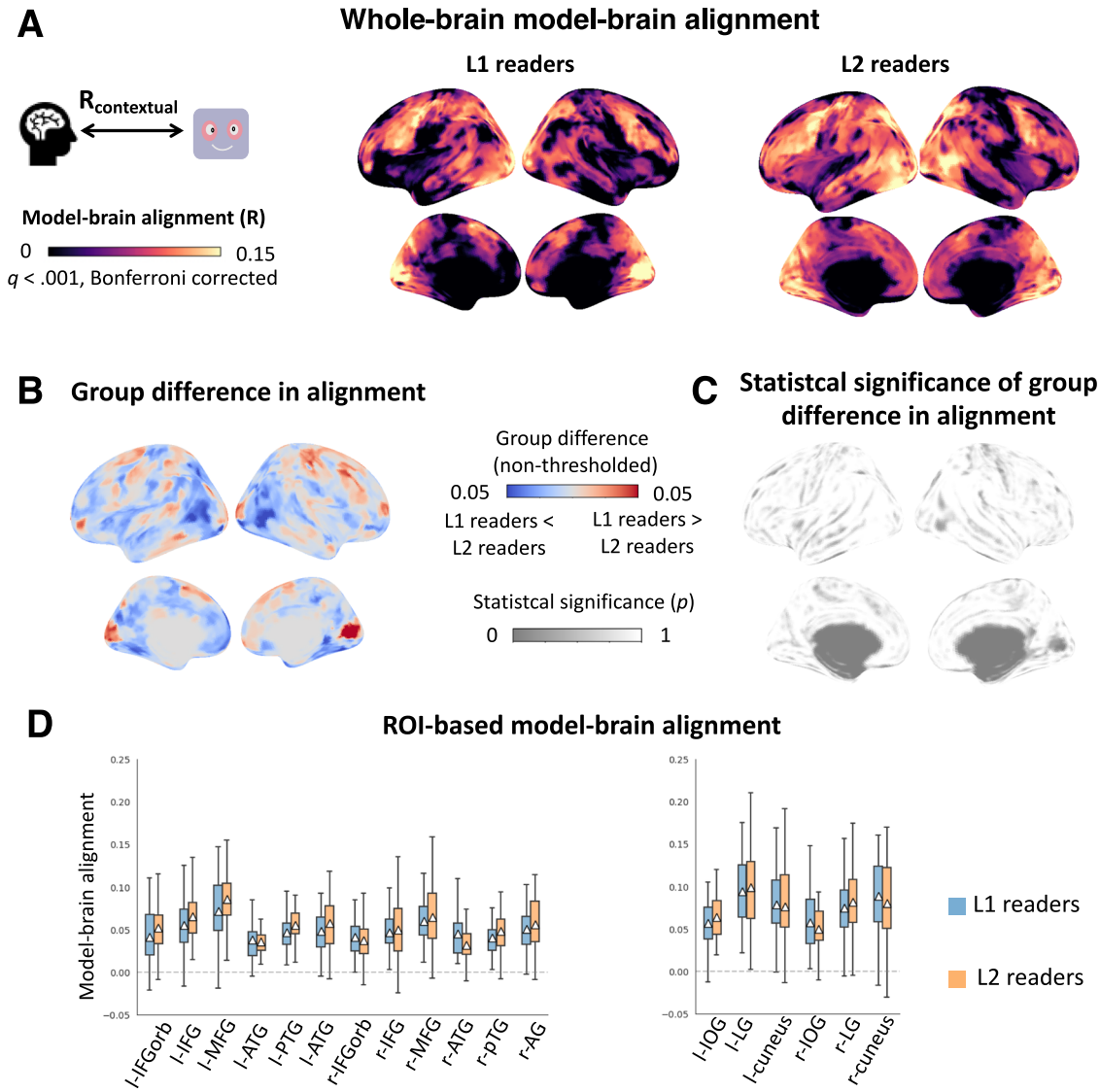

Fig. 2 | Model-brain alignment. a Whole-brain model-brain alignment. L1 and L2 readers both showed significant alignment in widespread brain regions. **b** Group differences in model-brain alignment. Parcels in red indicate greater alignment in L1 readers, but parcels in blue indicate greater alignment in L2 readers. **c** Statistical significance of group difference in alignment. **d** ROI-based model-brain alignment. The left and right panels show the alignment in language and visual regions, respectively. Note: the left and right brain regions were marked with 'l-' and 'r-' respectively; see a list of abbreviations in Supplementary Table 1; error bars indicate 95% CI.

These findings convergently reinforce the superiority of contextual embeddings in predicting brain responses across populations and highlight their unique contribution beyond other features, demonstrating the homogeneity of LLMs in modeling brains across populations.

Model-brain alignment predicts reading accuracy across readers

To test prediction (2) that greater model-brain alignment would be associated with better comprehension, we correlated reading accuracy with alignment in ROIs. This analysis hinges on the premises that (a) LLM representations correctly capture the meanings of the text and (b) humans comprehend the meanings of the text. Specifically, we performed correlation analyses in language and visual ROIs. We concatenated model-brain alignment and reading accuracy of the two groups, due to no group differences in both measures (Fig. 2d).

We found positive correlations between reading accuracy and alignment in bilateral language ROIs (Fig. 4a), including the left inferior frontal gyrus (IFG; $r = 0.23$, $p = 0.03$), left angular gyrus (AG; $r = 0.33$, $p = 0.003$), right anterior temporal gyrus (ATG; $r = 0.22$, $p = 0.03$), bilateral middle frontal gyrus (MFG; left: $r = 0.24$, $p = 0.03$; right: $r = 0.26$, $p = 0.03$), and bilateral posterior temporal gyrus (PTG; left: $r = 0.36$, $p = 0.002$; right: $r = 0.21$, $p = 0.04$). In contrast, none of the visual ROIs exhibited significant correlations between model-brain alignment and reading accuracy (Fig. 4a; $ps > 0.1$). These findings suggest that the better the reading outcome, the stronger the model-brain alignment in the language areas.

Furthermore, we built linear regression models using the leave-one-run-out cross-validation approach to predict reading accuracy from model-brain alignment, and then correlated the *predicted* with *actual* reading accuracy scores. This model predicted reading score in left-out subjects in the left PTG ($r = 0.30$, $p = 0.02$) and AG ($r = 0.27$, $p = 0.03$; see Fig. 4B). Together, these findings suggest (a) there is a tight association between reading outcomes and model-brain alignment, (b) this association only occurs for the language ROIs, and (c) alignment in the key left-hemisphere language regions predicts reading outcomes.

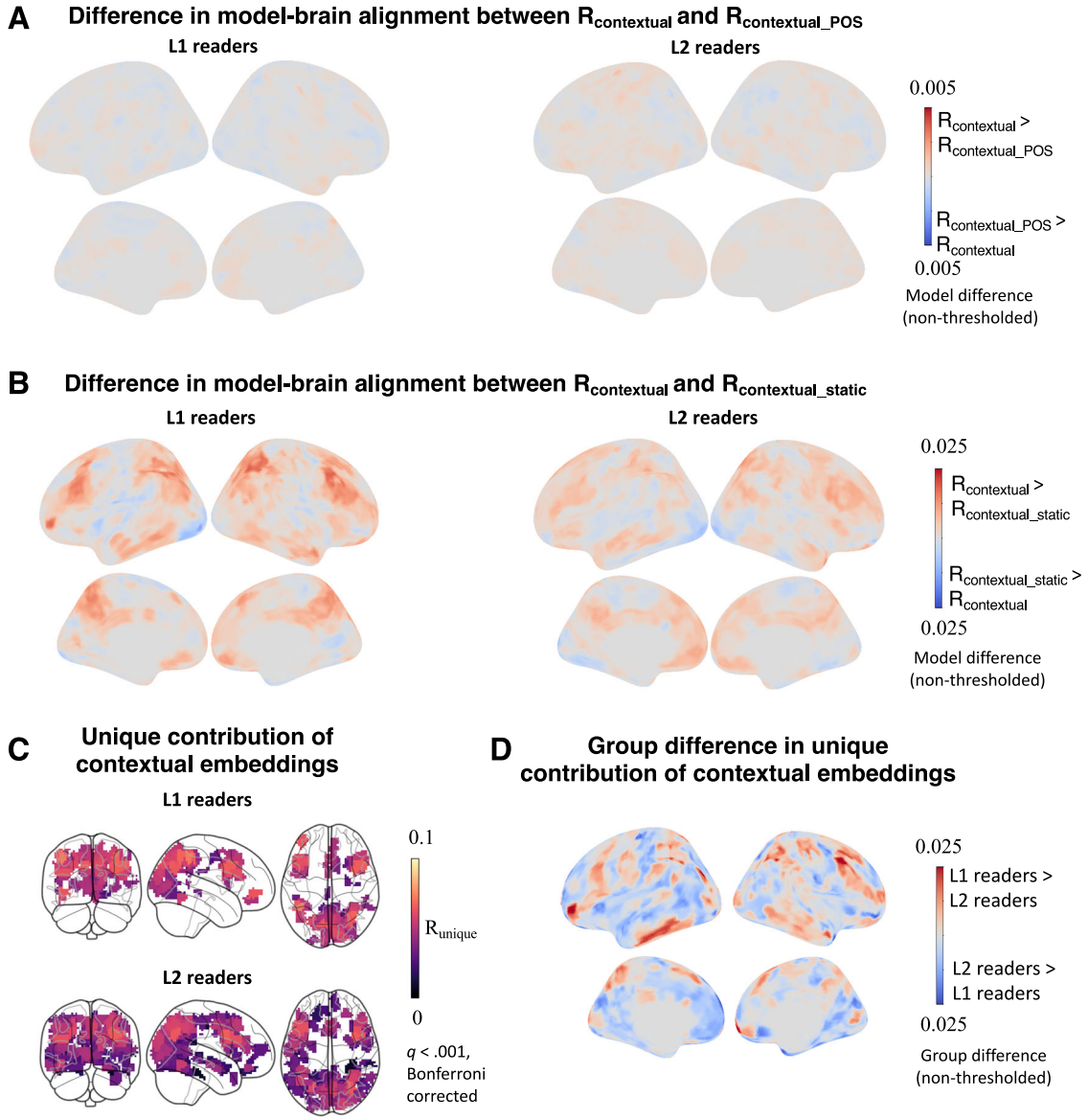

Fig. 3 | Advantages of contextual embeddings in predicting model-brain alignment. a Differences in model-brain alignment between $R_{contextual}$ and $R_{contextual\_POS}$. Parcels in red indicate greater $R_{contextual}$, while parcels in blue indicate greater $R_{contextual\_POS}$. No significant differences were observed between $R_{contextual}$ and $R_{contextual\_POS}$. **b** Differences in model-brain alignment between $R_{contextual}$ and $R_{contextual\_static}$. Parcels in red indicate greater $R_{contextual}$, while parcels in blue indicate greater $R_{contextual\_static}$. No widespread differences were observed between $R_{contextual}$ and $R_{contextual\_static}$. **c** Unique contribution of contextual embeddings. The unique contribution of contextual embeddings was significant across L1 and L2 readers. **d** Group differences in the unique contribution of contextual embeddings. The unique contributions of contextual embeddings were comparable between L1 and L2 readers.

Impact of expertise on model-brain alignment differed between L1 and L2 readers

To assess the impact of individual differences on model-brain alignment as outlined in Prediction (3), we constructed regression-based models to predict alignment in language ROIs with 'expertise' as the predictor for L1 and L2 readers, separately. We call this predictor 'expertise' as it includes linguistic and attentional abilities of the learner[22]. We included vocabulary size and general reading ability in the first regression model (referred to as the 'Ling model'), given the effects of linguistic abilities addressed by earlier work[1,16].

This model was predictive of alignment between LLMs and L1 readers in the left language regions comprising the orbital part of IFG (IFGorb), IFG, MFG, and PTG ($ps \leq 0.05$; Fig. 5a and Supplementary Table 4). Specifically, vocabulary size positively predicted model-brain alignment in the left IFG (Fig. 5b), IFGorb, and MFG. In addition, as general reading ability improved, the effect of vocabulary size on alignment improved, including the IFGorb, IFG, MFG, and PTG. By contrast, the Ling model did not significantly predict model-brain alignment for L2 readers.

The second regression model included linguistic and attentional abilities as predictors (referred to as the 'LingANT model'). Interestingly, the LingANT model was not predictive of the alignment between LLMs and L1 readers ($ps > 0.10$). By contrast, the LingANT model, including the linguistic and alerting abilities, significantly predicted the alignment between LLMs and L2 readers in the left IFG, MFG, anterior TG (ATG), and PTG ($ps \leq 0.055$; Fig. 5c and Supplementary Table 5). In particular, better alerting and general reading abilities positively predicted greater alignment in those regions (Fig. 5d). In addition, as vocabulary size increased, the prediction of general reading ability on the alignment decreased in the left IFG, ATG, and PTG (Supplementary Table 5 and Supplementary Fig. 1a, b). Furthermore, the LingANT model, consisting of linguistic and executive control abilities, predicted alignment between LLMs and L2 readers in the left MFG ($p < 0.05$; Fig. 5c and Supplementary Table 5). Better executive control and general

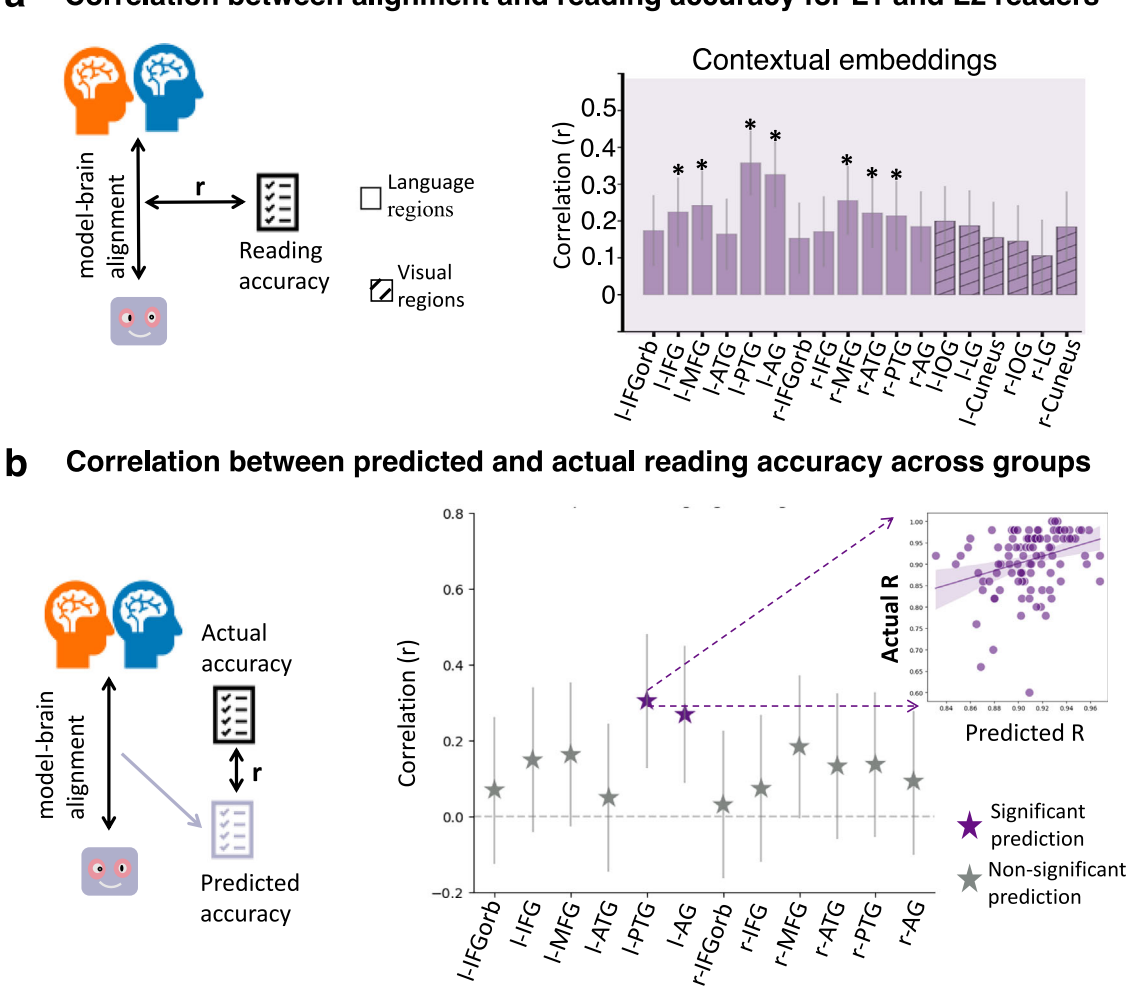

Fig. 4 | Association between reading accuracy and model-brain alignment. **a** Correlation between reading accuracy and model-brain alignment. Reading accuracy positively correlated with model-brain alignment in language ROIs using contextual embeddings. **b** Correlation between predicted and actual reading accuracy concatenated across L1 and L2 readers. Predictions of reading accuracy in the left PTG and AG positively correlated with actual reading accuracy. Note: left and right brain regions are labeled with 'l-' and 'r-', respectively; see a list of abbreviations in Supplementary Table 1; $^{*}p < 0.05$.

reading abilities positively predicted greater alignment, while vocabulary size negatively predicted alignment. In addition, the effect of executive control ability on alignment increased as vocabulary size increased (Supplementary Table 5 and Supplementary Fig. 1c).

Our findings demonstrate a population-general role of linguistic abilities in shaping model-brain alignment during naturalistic reading, but a population-specific contribution of attentional abilities among L2 readers.

Interplay between language dominance and expertise in L2 readers

Following existing neurocognitive theories of bilingual language processing[22,23,29], we extended our analysis by testing the effect of language experience in L2 readers. Among these experience factors, language dominance—which quantifies the dominance of language use in real-life activities—exerts significant effects on the structure and function of brain regions[23]. To examine the impact of multifaceted individual differences on the model-brain alignment, we developed a regression model (referred to as the 'full model') that incorporated the same expertise factors as predictors (i.e., vocabulary size, general reading ability, and attentional ability), plus the factor 'language dominance score' (the degree to which English is the dominant language in real-life reading) derived from the Language history questionnaire (LHQ 2.0)[30].

Our analyses indicated that this model was predictive of alignment between LLMs and L2 readers in the left IFG, MFG, and PTG ($ps \leq 0.05$; Fig. 6a and Supplementary Table 6). Specifically, we observed significant interaction effects between vocabulary size and language dominance and between alerting ability and language dominance (Fig. 6b): the prediction of language dominance on alignment in the left IFG decreased as vocabulary size increased, and on alignment in the left MFG and PTG also decreased as alerting ability increased. In line with the LingANT model for L2 readers, we identified the effects of general reading and alerting abilities on the alignment in the left IFG and MFG. Together, these findings demonstrate the modulation of language dominance during L2 reading, highlighting the heterogeneity in brain responses among L2 readers driven by language experience.

Discussion

How does the alignment between LLMs and human brains serve to study reading across diverse populations? This work provides neural evidence for both the homogeneity and heterogeneity in model-brain alignment during naturalistic reading. LLM-based contextual embeddings showed comparable performance in predicting brain activity between L1 and L2 readers, suggesting homogeneity in alignment across populations. Concurrently, model-brain alignment was modulated by individual differences in expertise and language experience, reflecting the heterogeneity in alignment

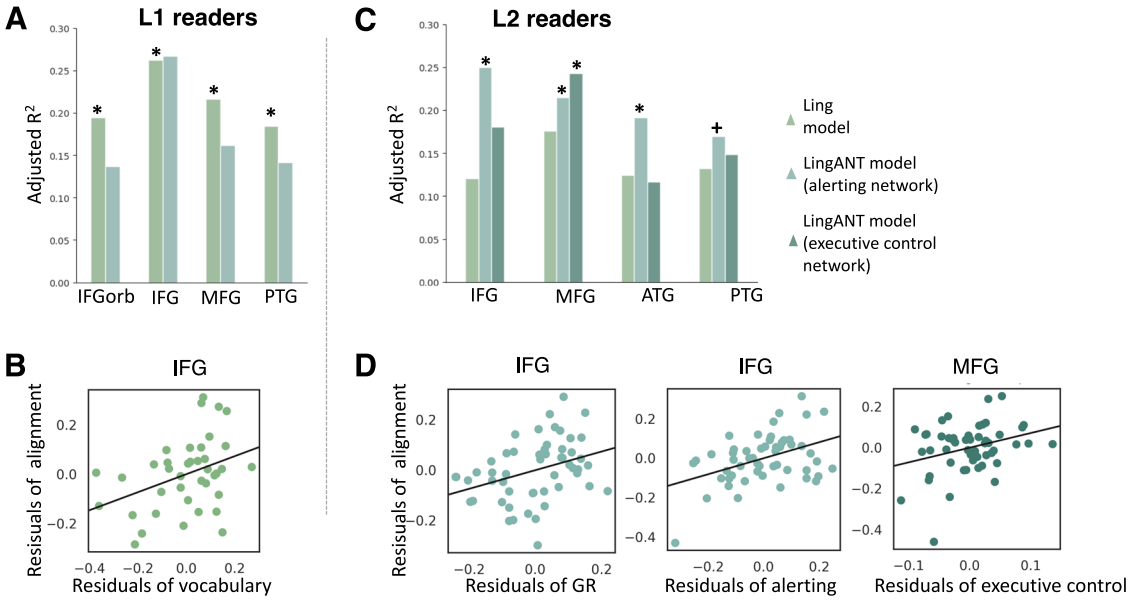

Fig. 5 | Ling and LingANT models for predicting model-brain alignment in L1 and L2 readers. a Prediction of model-brain alignment in L1 readers. The adjusted R^2 was significant or marginally significant ($ps \leq 0.050$) in the left IFGorb, IFG, MFG, and PTG. **b** Effect of vocabulary size. **c** Prediction of model-brain alignment in L2 readers. The adjusted R^2 was significant or marginally significant ($ps \leq 0.055$) in the left IFG, MFG, ATG, and PTG. Bars in muted teal represent the model including linguistic and alerting abilities, while the bar in deep teal represents the model including linguistic and executive control abilities. **d** Effects of general reading and attentional abilities. We inverted the sign of alerting and executive control scores for visualization. Note. GR = General reading ability; $^{*}p \leq 0.050$; $0.050 < {}^{+}p \leq 0.055$.

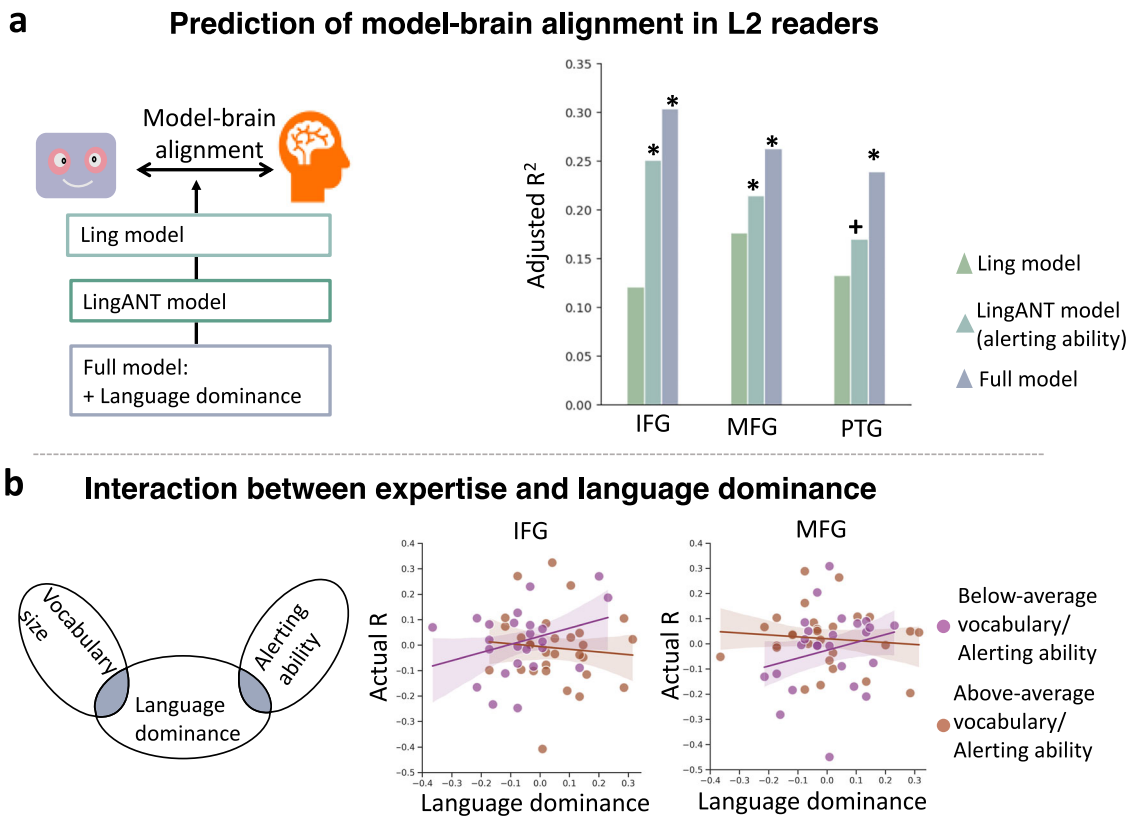

Fig. 6 | Full model for predicting model-brain alignment in L2 readers. **a** Prediction of model-brain alignment in L2 readers. Full model, compared to the Ling or LingANT models, displayed the greatest adjusted R^2 in the left IFG, MFG, and PTG. **b** Interaction effects. The effect of language dominance on the model-brain alignment in the left IFG and MFG decreased as vocabulary size (left) or alerting ability (right) increased. We inverted the sign of alerting scores for visualization. Note. $^{*}p \leq 0.050$; $0.050 < {}^{+}p \leq 0.055$.

between LLMs and human reading. Together, our work provides compelling evidence for the applicability of LLMs in studying the homogeneity and heterogeneity in human language processes.

While the encoding of contextual information is crucial for language comprehension, it is challenging for traditional neurocognitive work to identify related brain activity due to the absence of an explicit linguistic model. In this regard, recent neurocomputational work has advocated that the principle of context-specific embeddings based on LLMs can be applied to study human language processing using the model-brain alignment approach[2,4,7,27]. Aligning with these studies, we found that contextual embeddings predicted brain responses in widespread brain regions during reading. This is not unexpected given that LLMs are trained on a massive amount of text corpora to learn rich linguistic structures of human language[31], enabling contextual embeddings to encapsulate contextual information and word meanings. However, our work further contributes to the understanding of this principle by extending the potential of LLMs to the comparison of L1 and L2 readers within the model-brain alignment approach and by incorporating individual variations.

Our study found comparable alignment between models, including contextual embeddings alone and those combining them with additional features, indicating the unique contribution of contextual embeddings in predicting human brains during reading. Unlike static embeddings and POS labels, contextual embeddings encode context-level conceptual meanings[32]. This high-level process is fundamental for discourse comprehension, as it reflects the interplay between preceding context and current words[4]. Notably, contextual embeddings predict brain responses across diverse brain networks, such as default mode, language, frontoparietal, and visual networks. Prior reviews demonstrate that these networks are involved in various linguistic processes during discourse comprehension[14,33,34]. Our findings, using the model-brain alignment approach, underscore that these brain networks are tuned to context-level conceptual information during naturalistic reading across groups.

Furthermore, our study identified comparable contributions of contextual embeddings between L1 and L2 readers, reflecting that both groups similarly relied on contextual information. Coincidentally, using an LLM-derived surprisal measure, a recent study reported the emergence of a surprisal effect across L1 and L2 readers during naturalistic reading, reflecting similar predictive/integrative processes across the two groups[35]. Notably, our work used pre-trained English data (GPT-2), and the native language of L2 readers in this study is Mandarin. Yet, when L2 readers read in English, the LLM-based model using English data could capture their brain responses equally well as it did for L1 readers. Given comparable reading achievement across groups, these findings suggest that the association between models and humans may depend more on comprehension success than on language background. Supporting this argument, we found that model-brain alignment did not predict whether a reader was L1 or L2 (see Supplementary Note 2 and Supplementary Fig. 2). Collectively, these findings highlight the potential of LLMs as cognitively plausible models for studying naturalistic language processing and offer methodological insight for cross-linguistic work without bias toward language background.

However, in contrast to comparable model-brain alignment between L1 and L2 readers, the univariate brain activation analysis identified differential brain activations between the two groups (see Supplementary Note 1 and Supplementary Fig. 3). Specifically, we found greater activations in bilateral cuneus and calcarine for L1 readers but greater activations in the right IOG and left precuneus for L2 readers. This discrepancy may arise from the differences in approaches. Traditional neuroimaging work averaged brain responses throughout the course of reading, potentially collapsing word-by-word linguistic nuances. By contrast, the model-brain alignment approach pinpoints brain responses associated with explicit linguistic models. Consequently, it prompts us to ask how we can use the LLM-based approach to study human language as a complement to traditional approaches[36–38].

Furthermore, comprehension achievement was associated with model-brain alignment in higher-order language regions but not in lower-order visual regions. These language regions are implicated in diverse linguistic processes crucial for discourse comprehension, such as inference processing or coherence making[33,39,40]. Specifically, the left IFG plays a key role in the unification system of language comprehension[40], which involves integrating linguistic elements. The selected visual regions (e.g., IOG and LG) are known to be engaged in letter identification and pre-lexical processing[41,42], and they may interact with the left IFG to function as part of the ventral lexical-semantic pathway during naturalistic reading[43,44]. In addition, the left IFG is engaged in syntactic and semantic unification by interacting with other higher-order language regions such as PTG and AG[40]. Together, our findings provide direct evidence of the engagement of these language regions in the encoding of contextual information. In light of the interactive nature of the neurocognitive system underlying language processing, future research should examine how various brain regions interact to contribute to language comprehension within the model-brain alignment framework.

A key contribution of our work is our finding that considerable variations in the correspondence between LLMs and human brains indicate that 'shared computational principles' may depend on the expertise and experience of individual language users. In line with recent work reporting the mediation of L2 proficiency on the surprisal effect[35], we found that linguistic abilities influence model-brain alignment in language regions among L2 readers. We also identified similar effects among L1 readers, aligning with prior behavioral studies[1,15,16]. Complementing previous neuroimaging work, our study provides the first neural evidence, within the model-brain alignment framework, for a population-general role of linguistic abilities in reading comprehension across both L1 and L2 readers. These findings further demonstrate the heterogeneity in model-brain alignment and its association with linguistic processes, contributing to the refinement of the proposal of 'shared computational principles'.

Our study also discovered an additional role of attentional abilities in predicting model-brain alignment for L2 readers but not L1 readers. These attentional networks, involving the allocation of attentional resources for sustained vigilance and conflict resolution, are associated with discourse comprehension[45–47]. The differential roles of the attentional networks may reflect distinct attentional demands on reading between the two groups. The Flesch-Kincaid Grade Level of the texts used in this work ranged from sixth to eighth grade, likely requiring less demanding cognitive resources from L1 readers. Previous work found the effects of cognitive skills on structural neuroplasticity under the cognitively demanding learning condition, but not in the less demanding condition[48]. Similarly, less challenging attentional requirements may diminish the effects of attentional networks on L1 readers.

In contrast, L2 readers have heightened demands on cognitive processes due to less automatic linguistic processes[15,18,19]. The compensatory mechanism proposes that increased cognitive demands can enhance L2 comprehension by devoting additional cognitive resources[49,50]. Similarly, neurocognitive meta-analyses have shown greater activations in brain regions (e.g., IPL, IFG, and MFG) implicated in language control in L2 reading than in L1 reading[13,14]. Therefore, the additional role of attentional ability prompts us to conclude that, relative to L1 readers, it is more challenging for L2 readers to achieve successful reading. This conclusion is substantiated by the findings of longer reading time, greater number of fixations, and longer mean fixation durations among L2 readers. Thus, the population-specific effect of attentional abilities for L2 readers highlights the importance of individual differences when using LLMs to study naturalistic language comprehension.

To what extent do language experiences modulate the neurocognitive patterns of bilingual speakers? In our study, we identified the impact of L2 dominance on alignment. In particular, language dominance impacted the alignment by interacting with linguistic and attentional abilities. Unlike prior work that shows the effect of language experience alone on brain functions[24,51], this study focuses on the joint contribution of language experience and expertise. The interaction effect suggests that L2 readers in our study who scored low on expertise could compensate for their low abilities when having greater language dominance, aligning with the

compensatory mechanism[49,50]. This interpretation is also consistent with the proposal that language experience shapes brain architectures and increases the efficiencies of linguistic and cognitive processes[23,29]. These findings highlight the significance of language experience for L2 readers using the model-brain alignment approach.

This work has several limitations. First, we only used GPT-2, a widely used transformer-based model, but variations in model architecture may lead to distinct patterns of model-brain associations. A recent study[35] reported that the surprisal measure, derived from a model capturing the hierarchical structures, was the best predictor of N400 for L2 readers. By contrast, for L1 readers, brain responses were best predicted by the surprisal measure derived from the transformer model encapsulating semantic associations between words. While the comparison of different LLM models could be further pursued, such work is out of this study's scope, requiring more resources and time to complete in future studies. Second, in line with previous model-brain alignment work using fMRI data[2,27], the overall model-brain alignment is not high. Notably, naturalistic language comprehension involves complex linguistic and cognitive processes, but LLM-based contextual embeddings pinpoint brain responses associated with the encoding of contextual information. Mapping brain signals, acquired over time and space, to high temporal resolution LLM-based metrics is a novel approach, but is still at its early stage of development. In addition, variation in model architecture may lead to different model-brain alignment patterns. To address these challenges, more sophisticated LLMs and enhanced methods for data acquisition are warranted to reflect the nuanced processes involved in human language processing. Finally, other domain-general executive functions may differentially influence language processing between L1 and L2 readers. For example, working memory updating plays a critical role in discourse comprehension[52] and in the development of the bilingual language system[53]. Future investigations are needed to fully understand the impact of domain-general executive functions on the alignment between LLMs and L2 processing.

To conclude, our work is the first systematic application of the neurocomputational approach to investigate the neurocomputational mechanism underlying naturalistic reading in both L1 and L2 readers, along with the examination of the impacts of individual differences. This work is consistent with recent calls to integrate cognitive, computational, and neuroscience perspectives to study cognition and language[8,54]. It is important to note that despite novel insights into human language processing, LLMs as a unified model have limitations in fully accounting for how human brains process language. One important direction for future research is to harness interdisciplinary approaches across cognitive neuroscience and natural language processing to enhance the cognitive plausibility of LLMs in capturing human language processing, including its use in real-world contexts.

Methods

Participants

Fifty-two native English speakers (L1 readers; 24 males; mean age ± SD = 22.85 ± 4.66) and fifty-six Chinese-speaking learners of English (L2 readers; 26 males; mean age ± SD = 25.14 ± 4.74) were recruited. All L2 readers were required to pass the Chinese English Test 6 (CET6). Participants in both groups were right-handed with normal or corrected normal vision and had no history of mental or neurological disorders. This research was approved by the Pennsylvania State University Institutional Review Board (IRB; Study ID: STUDY00002823) and followed ethical standards. Written informed consent was obtained from all participants before the experiments. Half of the L2 readers were recruited from Pennsylvania State University, and the other half were recruited from Beijing Normal University and Peking University. We excluded one L1 reader and one L2 reader due to missing eye movement data. Eventually, this work included 51 L1 and 55 L2 readers.

Stimuli

Five short expository texts about STEM content, adopted from a previous study[55], were presented to participants in the MRI scanner. The topics of the texts include Mars, Supertanker, Math, Global Positioning System (GPS), and Electric Circuit. Text characteristics (e.g., the length of texts and mean word count per sentence; see Supplementary Table 3) and psychological variables of lexical properties (e.g., age of acquisition, familiarity for content words, word frequency, and Coh-Metrix measurements) in the five texts were both controlled.

Behavioral measurements

We assessed the general reading ability of each participant using the Gray Silent Reading Test (GSRT)[56], comprising 13 narrative tests. Each narrative was presented to participants, along with five multiple-choice questions, to assess reading comprehension. The test started with the 8th narrative (i.e., middle-level difficulty), and was conducted downward until the basal (i.e., all questions were answered correctly) was reached, and upward until the ceiling (i.e., three out of five answers were wrong) was reached. The total number of correct questions was counted as the raw score, ranging from 0 to 65. We converted the raw score into standardized quotient scores and percentile ranks following the formulas. This test has been normed on a sample of 1,400 individuals, with reliability coefficients alpha at or above 0.97. Following previous work[57], we adopted the percentile rank to reflect general reading ability. Two L1 readers were excluded when the analysis involving general reading ability due to missing data.

We assessed English receptive vocabulary size using the Peabody Picture Vocabulary Test (PPVT IV)[58], consisting of 228 items distributed across 19 item sets. For each item, a word was aurally presented, accompanied by four pictures on the screen. Participants were required to select a picture matching the meaning of the word. This test was administered downward until the basal set (i.e., until only one or zero errors within the item set) was reached, and upward until the ceiling set (i.e., eight or more errors within the item set) was reached. Participants obtained one score for each correct item, so the total score ranges from 0 to 228. Twelve L1 readers were excluded when the analysis involving vocabulary size due to missing data.

We used the attentional network test (ANT)[59,60], to assess domain-general cognitive abilities in three attention networks, including the alerting network, the orienting network, and the executive control network. For each trial, a central arrow was presented with congruent or incongruent flanking arrows accompanied by attentional and/or spatial cues. Participants were required to indicate the direction of central arrows as fast and accurately as possible. Following previous work[59,61], the alerting network score was calculated by subtracting the mean reaction time (RT) of the double-cue condition from that of the no-cue condition. The orienting network score was derived by subtracting the mean RT of the spatial-cue condition from that of the central-cue condition. The executive control network score was obtained by subtracting the mean RT of the congruent flanker condition from that of the incongruent flanker condition. We only included correct trials for both conditions. A smaller value indicates greater attentional network scores. Two L1 readers were excluded from the analyses involving attentional ability due to missing data.

We used LHQ 2.0 to measure language dominance[30], one of the important experience measurements[23,29]. This measure quantifies how dominant a language is in daily life. Since this work investigates reading comprehension, we computed the language dominance score relevant to reading activities, following the formula proposed by a recent work[62]. A greater value indicates that the target language is more dominant in real-life reading. Two L2 readers were excluded from the analysis using language dominance because their data were either missing or inaccurate.

MRI task procedure and image acquisition

We used the fixation-related fMRI paradigm[63] to simultaneously record eye movements and BOLD signals while participants performed the self-paced reading task in the MRI scanner (Fig. 1c). Each text was presented on the screen sentence by sentence, and the maximum duration of each sentence was 8 s. Participants pressed the button to proceed to the next sentence once they finished the current sentence. Ten multiple-choice comprehension

questions were used to assess reading performance at the end of the text. In total, each participant completed five runs containing one text for each run, and the order of texts was randomized across participants.

3-T Siemens scanner with a 64-channel phased array coil was used to acquire T_1-weighted, T_2* weighted, and Diffusion tensor images. T_1 weighted images employed MPRAGE sequence (TR = 1540 ms; TE = 2.34 ms; flip angle = 9°; GRAPPA in-plane acceleration factor = 2; voxel size = 1 mm × 1 mm × 1 mm; acquisition time = 216 s), covering cerebrum, cerebellum, and brain stem. Functional runs of T_2* weighted images utilized the echo-planar sequence (TR = 400 ms; TE = 30 ms; flip angle = 35°; voxel size = 3 mm × 3 mm × 4 mm; acquisition time varied based on reading speed). To correct for distortions caused by the multiband acquisition, a pair of echo-planar spin-echo sequence images (A/P and P/A phase encoding direction) were acquired (TR = 3000 ms; TE = 51.2 ms; flip angle = 90°; voxel size = 3 mm × 3 mm × 4 mm)[64].

Eye-tracking data acquisition

The Eye-Link 1000 Plus long-range mount MRI-compatible eye tracker (SR-Research) was used to record eye movements monocularly (from the right eye). Detailed parameters are as follows: sampling rate = 1000 HZ; distance between the camera and the participants' eyes = 120 cm; mean word length on the screen = 3.08 cm; average distance between words = 0.95 cm; On average, a reader's visual angle when fixating on a word is 1°14¢. Before the experiment and each run, a 13-point calibration routine was employed.

Reading performance

We assessed reading comprehension achievement using reading accuracy (ACC). ACC was quantified by the proportion of questions that participants correctly answered after reading each text. In total, there were 50 questions for all five texts. Four L2 readers were excluded from the analysis involving ACC because their responses to several questions were missing. We measured the average total reading time for each participant across five texts. In addition, we quantified the number of fixations and mean fixation duration per participant. We performed two-sample t-tests to examine group differences in ACC and eye-movement measures.

fMRI data preprocessing

SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/, Wellcome Department of Cognitive Neurology, United Kingdom) and fsl-5.0.11 (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FslInstallation#Installing_FSL) were adopted for preprocessing, including slice-time correction, voxel displacement map calculation, realignment & unwrapping (motion and distortion correction), co-registration, segmentation, normalization, and smoothing. Head motion during realignment was measured by three translational and three rotational parameters. The average framewise displacement (FD) for five runs was estimated[65]. No participant exceeded the exclusion criterion (FD > 2.5 mm). We used unsmoothed functional images to implement confound regression with the Nilearn Python library (http://nilearn.github.io). We included six head movement parameters and three physiological regressors (i.e., WM, CSF, and ventricular CSF) as confound regressors, and applied a high-pass filter with a cut-off of 100 s to remove the low-frequency signals. These preprocessed images were used for the subsequent model-brain alignment analysis.

ROI-based analyses

For the ROI-based analyses, we selected language regions as ROIs from a predefined language mask[66], including bilateral IFGorb, IFG, MFG, ATG, PTG, and AG (Supplementary Fig. 4a). Additionally, we used Neurosynth[67] to generate a reading mask with the search term 'reading' through an automated meta-analysis of 521 "reading" related studies. Next, we chose the peak coordinates of clusters in the occipital cortex with more than 50 contiguous voxels to generate 10-mm sphere visual ROIs. Lastly, we flipped these visual ROIs to obtain bilateral regions akin to language ROIs, including bilateral IOG, LG, and cuneus (Supplementary Fig. 4b).

Embeddings extracted from LLMs

We extracted contextual embeddings from the GPT-2 model[26], a recognized transformer-based model applied in recent work[2–4]. GPT-2 uses a multi-layer self-supervised mechanism to acquire linguistic structure during natural language processing. Each successive layer processes the output from the preceding layer, dynamically generating embeddings of the input. Previous work[1,2] has shown the best performance of GPT-2 with the middle layers, like the eighth layer, so this work derived contextual embeddings from the eighth layer. In addition, we extracted static embeddings from GPT-2, capturing the "average" meaning of that word independent of contexts. These embeddings are produced when the input layer converts tokenized text into vectors, which are then processed by subsequent layers of self-attention. The following procedure was used to extract the embeddings: (1) all five texts were individually tokenized using the tokenizer provided by HuggingFace; (2) the tokens of each text were used as the language input to run the GPT-2 model; (3) the embeddings of all tokens were extracted from the model, resulting in a 768-dimensional vector for each token; and (4) the embeddings of all tokens from a word were averaged. Finally, the GPT-2 model generated a 768-dimensional vector for each word.

Model-brain alignment

Following prior work[2,6], we constructed encoding models to estimate the alignment between brain responses and LLM-based metrics (e.g., contextual embeddings). To do so, we adopted a leave-one-run-out cross-validation approach, using four of the five runs as the training sample and the remaining run as the test sample. This procedure was iteratively performed five times, ensuring imaging data of each run served as the test sample once. Next, we built a linear ridge regression model to predict brain responses on the training sample using LLM-based metrics. Given the trained encoding model, we then predicted brain responses for the test sample (i.e., the left-out run) and computed the Pearson correlation between predicted and actual brain responses. Lastly, we averaged the correlation coefficients derived from each cross-validation fold to determine the model-brain alignment for each participant.

In line with previous work[2,6], we adopted a finite impulse response (FIR) model with five delays to construct the linear ridge regression model, consisting of 0, 5, 10, 15, and 20 time points corresponding to 0 s, 2 s, 4 s, 6 s, and 8 s. To align language stimuli with continuous imaging data, we adopted a downsampling approach as proposed by prior work[2,3,6]. Specifically, individual eye fixations were registered to each repetition time (TR) of the imaging data. This registration resulted in each TR containing a varying number of words, ranging from 0 to 4 for both L1 and L2 readers. To match the sampling frequency between brain responses and embeddings, we averaged the embeddings of all words within the same TR. However, no attempt was made to match the actual number of words within each TR, given the variability that may exist across readers. We implemented scikit-learn to z-score brain responses and embeddings before running the regression model, and adopted L2-penalized linear regression using RidgeCV from the Himalaya Python library[68] to build the regression model. The regularization coefficients (i.e., L2 penalty terms) of RidgeCV were selected with nested leave-one-run-out cross-validation from 11 log-spaced values ranging from 10^{-1} to 10^{9} for each training fold.

We used a 1000-parcel brain atlas to partition the brain into 1000 parcels[28] and estimated the model-brain alignment for each parcel by implementing the parcel-wise encoding model. We also quantified model-brain alignment in language and visual ROIs by constructing the voxel-wise encoding model. In line with recent work[2], we performed a two-sided Wilcoxon rank-sum test to evaluate the significance of model-brain alignment within individual groups. In addition, we conducted a Mann–Whitney U-test to compare model-brain alignment between groups. To correct for the multiple comparisons, we applied a strict threshold ($q < 0.001$, Bonferroni corrected).

Model-brain alignment specific to contextual information

To demonstrate the advantage of contextual embeddings in predicting human brain responses over other features, we compared the model

including contextual embeddings ($R_{contextual}$) with (a) the model incorporating contextual embeddings and POS labels ($R_{contextual\_POS}$) and (b) the model incorporating contextual and static embeddings ($R_{contextual\_static}$). To do so, we used the NLTK Python package to assign individual words a one-hot vector of 11 categories, representing POS features including noun, verb, proper noun, past tense verb, present participle, base verb, adjective, adverb, determiner, coordinating conjunction, and preposition/subordinating conjunction. In addition, we derived static embeddings from the non-contextual word token embedding matrix generated by the GPT-2 model.

To estimate the model-brain alignment uniquely contributed by contextual embeddings, following recent work[32], we established the full encoding model by concatenating all features, including contextual embeddings, static embeddings, and POS labels. Next, we constructed the partial encoding model by excluding contextual embeddings. Finally, we subtracted the model-brain alignment derived from the partial model (R_{static_pos}) from that derived from the full model (R_{full}) to obtain the unique model-brain alignment predicted by contextual embeddings (R_{unique}).

We conducted two-sided Wilcoxon rank-sum tests to compare model-brain alignment between models and the significance of R_{unique} for each group, respectively. Furthermore, we implemented Mann–Whitney *U*-tests to compare whether R_{unique} would differ between L1 and L2 readers. To correct for multiple comparisons, we applied a strict threshold ($q < .001$, Bonferroni corrected).

Association between reading performance and model-brain alignment

To reveal the association between reading achievement and model-brain alignment in language and visual ROIs, we conducted the correlation analyses for L1 and L2 readers, separately. Next, we concatenated the alignment and reading accuracy of the two groups, due to no significant group differences in accuracy or model-brain alignment. Four L2 readers were excluded from this analysis because their responses to several questions were missing. To correct for multiple comparisons, we applied FDR correction ($q < 0.05$) for ROIs from the same brain network.

Further, we established linear regression models using the Sklearn Python Library to predict reading accuracy using model-brain alignment when significant correlations were shown in the above analyses. To do so, we applied the leave-one-run-out cross-validation approach and Pearson correlation to evaluate model performance. We applied FDR correction ($q < 0.05$), and reported the corrected significance. To test whether model-brain alignment predicts which population the reader comes from (i.e., L1 vs L2 readers), we built a classifier model using the leave-one-run-out cross-validation approach. We evaluated the model performance using the bootstrap approach (for details, see Supplementary Note 2).

Impacts of individual differences on model-brain alignment

To examine the impacts of individual differences, we constructed regression-based models to predict model-brain alignment in language ROIs by including linguistic and attentional abilities as predictors. In light of the importance of linguistic abilities in language comprehension[1,69], the first model included linguistic abilities (i.e., vocabulary size and general reading ability) as predictors (referred to as the 'Ling model'). The subsequent model consisted of linguistic and attentional abilities as predictors (referred to as the 'LingANT model'), aligning with recent work that highlights the role of cognitive abilities[15,16]. We established the LingANT model for each of the three attentional networks for both L1 and L2 readers.

In the field of neurocognitive work, bilingual language experience also impacts the structure and function of brain regions engaged in language processes[22,23,29]. Consequently, we constructed a regression-based model that included linguistic abilities, cognitive abilities, and language dominance as predictors (referred to as the 'Full model') for L2 readers. Two L2 readers and twelve L1 readers were excluded from the regression models due to missing behavioral assessments.

To discern the interaction effects among individual differences proposed by recent studies[22,23,29], we added the interaction terms to the regression-based models. We limited the regression model to the two-way interaction effect because it is complicated to explain the three-way effects or above. All predictors included in the regression model were mean-centered and normalized. The variance inflation factor (VIF) across regression models was below 10 (ranging from 1.0 to 8.4), indicating an acceptable level of multicollinearity[70]. The regression model performance was corrected using the FDR ($q < 0.05$) approach.

Data availability

All data needed to evaluate the conclusions of this paper are present in the paper and/or the Supplementary Information. Data that support the findings of this study are available at https://openneuro.org/datasets/ds003988 and https://openneuro.org/datasets/ds003974.

Code availability

All scripts used in this study can be accessed upon request by contacting the authors.

Received: 31 May 2025; Accepted: 24 June 2025;
Published online: 10 July 2025

References

1. Perfetti, C. & Stafura, J. Word knowledge in a theory of reading comprehension. *Sci. Stud. Read.* **18**, 22–37 (2014).
2. Caucheteux, C., Gramfort, A. & King, J.-R. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat. Hum. Behav.* **7**, 430–441 (2023).
3. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5**, 134 (2022).
4. Goldstein, A. et al. Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).
5. Goldstein, A. et al. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nat. Commun.* **15**, 2768 (2024).
6. Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
7. Zada, Z. et al. A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. *Neuron* **112**, 3211–3222 (2024).
8. Yu, S., Gu, C., Huang, K. & Li, P. Predicting the next sentence (not word) in large language model pretraining: what model-brain alignment tells us about discourse comprehension. *Sci. Adv.* **10**, eadn7744 (2024).
9. Grosjean, F. & Li, P. *The Psycholinguistics of Bilingualism* (John Wiley & Sons, Inc., New York, 2013).
10. Bley-Vroman, R. The evolving context of the fundamental difference hypothesis. *Stud. Second Lang. Acquis.* **31**, 175–198 (2009).
11. Clahsen, H. & Felser, C. How native-like is non-native language processing?. *Trends Cogn. Sci.* **10**, 564–570 (2006).
12. Jeong, H. & Li, P. Neurocognition of social learning of second language: How can second language be learned as first language? in *The Routledge Handbook of Second Language Acquisition and Neurolinguistics* 217–229 (Routledge, 2023).
13. Li, H., Zhang, J. & Ding, G. Reading across writing systems: a meta-analysis of the neural correlates for first and second language reading. *Biling. Lang. Cogn.* **24**, 537–548 (2021).
14. Sulpizio, S., Del Maschio, N., Fedeli, D. & Abutalebi, J. Bilingual language processing: a meta-analysis of functional neuroimaging studies. *Neurosci. Biobehav. Rev.* **108**, 834–853 (2020).
15. Li, P. & Clariana, R. B. Reading comprehension in L1 and L2: an integrative approach. *J. Neurolinguist.* **50**, 94–105 (2019).

16. Duke, N. K. & Cartwright, K. B. The science of reading progresses: communicating advances beyond the simple view of reading. *Read. Res. Q.* **56**, S25–S44 (2021).
17. Raudszus, H., Segers, E. & Verhoeven, L. Situation model building ability uniquely predicts first and second language reading comprehension. *J. Neurolinguist.* **50**, 106–119 (2019).
18. Raudszus, H., Segers, E. & Verhoeven, L. Lexical quality and executive control predict children's first and second language reading comprehension. *Read. Writ.* **31**, 405–424 (2018).
19. Taboada Barber, A., Cartwright, K. B., Hancock, G. R. & Klauda, S. L. Beyond the simple view of reading: the role of executive functions in emergent bilinguals' and English monolinguals' reading comprehension. *Read. Res. Q.* **56**, S45–S64 (2021).
20. Jung, J., Zhang, W. & Lee, M. The role of working memory and attention control in incidental L2 vocabulary learning from reading-while-listening. *ITL Int. J. Appl. Linguist.* **176**, 44–75 (2025).
21. Grant, A. M., Fang, S.-Y. & Li, P. Second language lexical development and cognitive control: a longitudinal fMRI study. *Brain Lang.* **144**, 35–47 (2015).
22. Claussenius-Kalman, H., Hernandez, A. E. & Li, P. Expertise, ecosystem, and emergentism: dynamic developmental bilingualism. *Brain Lang.* **222**, 105013 (2021).
23. DeLuca, V., Segaert, K., Mazaheri, A. & Krott, A. Understanding bilingual brain function and structure changes? U bet! a unified bilingual experience trajectory model. *J. Neurolinguist.* **56**, 100930 (2020).
24. Liu, H. & Cao, F. L1 and L2 processing in the bilingual brain: a meta-analysis of neuroimaging studies. *Brain Lang.* **159**, 60–73 (2016).
25. Melby-Lervåg, M. & Lervåg, A. Reading comprehension and its underlying components in second-language learners: a meta-analysis of studies comparing first-and second-language learners. *Psychol. Bull.* **140**, 409–433 (2014).
26. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
27. Caucheteux, C., Gramfort, A. & King, J.-R. Deep language algorithms predict semantic comprehension from brain activity. *Sci. Rep.* **12**, 16327 (2022).
28. Kong, R. et al. Individual-specific areal-level parcellations improve functional connectivity prediction of behavior. *Cereb. Cortex* **31**, 4477–4500 (2021).
29. Li, P., Legault, J. & Litcofsky, K. A. Neuroplasticity as a function of second language learning: anatomical changes in the human brain. *Cortex* **58**, 301–324 (2014).
30. Li, P., Zhang, F., Tsai, E. & Puls, B. Language history questionnaire (LHQ 2.0): a new dynamic web-based research tool. *Biling. Lang. Cogn.* **17**, 673–680 (2014).
31. Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U. & Levy, O. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl. Acad. Sci. USA* **117**, 30046–30054 (2020).
32. LeBel, A., Jain, S. & Huth, A. G. Voxelwise encoding models show that cerebellar language representations are highly conceptual. *J. Neurosci.* **41**, 10341–10355 (2021).
33. Ferstl, E. C., Neumann, J., Bogler, C. & Von Cramon, D. Y. The extended language network: a meta-analysis of neuroimaging studies on text comprehension. *Hum. Brain Mapp.* **29**, 581–593 (2008).
34. Yang, X., Lin, N. & Wang, L. Situation updating during discourse comprehension recruits right posterior portion of the multiple-demand network. *Hum. Brain Mapp.* **44**, 2129–2141 (2023).
35. Oralova, G. et al. Surprisal in reading: language models predict the N400 for L2 readers. *Lang. Cogn. Neurosci*. (in press).
36. Hasson, U., Nastase, S. A. & Goldstein, A. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron* **105**, 416–434 (2020).
37. Nastase, S. A., Goldstein, A. & Hasson, U. Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *Neuroimage* **222**, 117254 (2020).
38. Arana, S., Pesnot Lerousseau, J. & Hagoort, P. Deep learning models to study sentence comprehension in the human brain. *Lang. Cogn. Neurosci.* **1**, 19 (2023).
39. Yang, X. et al. Uncovering cortical activations of discourse comprehension and their overlaps with common large-scale neural networks. *Neuroimage* **203**, 116200 (2019).
40. Hagoort, P. MUC (memory, unification, control) and beyond. *Front. Psychol.* **4**, 416 (2013).
41. Borowsky, R. et al. FMRI of ventral and dorsal processing streams in basic reading processes: insular sensitivity to phonology. *Brain Topogr.* **18**, 233–239 (2006).
42. Olulade, O. A., Flowers, D. L., Napoliello, E. M. & Eden, G. F. Dyslexic children lack word selectivity gradients in occipito-temporal and inferior frontal cortex. *NeuroImage Clin.* **7**, 742–754 (2015).
43. Zhou, W. et al. Neural mechanisms of dorsal and ventral visual regions during text reading. *Front. Psychol.* **7**, 1399 (2016).
44. Zhou, W. & Shu, H. A meta-analysis of functional magnetic resonance imaging studies of eye movements and visual word reading. *Brain Behav.* **7**, e00683 (2017).
45. Arrington, C. N., Kulesz, P. A., Francis, D. J., Fletcher, J. M. & Barnes, M. A. The contribution of attentional control and working memory to reading comprehension and decoding. *Sci. Stud. Read.* **18**, 325–346 (2014).
46. Conners, F. A. Attentional control and the simple view of reading. *Read. Writ.* **22**, 591–613 (2009).
47. Kocaarslan, M. The relationships between oral reading fluency, sustained attention, working memory, and text comprehension in the third-grade students. *Psychol. Sch.* **59**, 744–764 (2022).
48. Legault, J. et al. Immersive virtual reality as an effective tool for second language vocabulary learning. *Languages* **4**, 13 (2019).
49. Serafini, E. J. & Sanz, C. Evidence for the decreasing impact of cognitive ability on second language development as proficiency increases. *Stud. Second Lang. Acquis.* **38**, 607–646 (2016).
50. Stevenson, M., Schoonen, R. & de Glopper, K. Inhibition or compensation? A multidimensional comparison of reading processes in Dutch and English. *Lang. Learn.* **57**, 115–154 (2007).
51. DeLuca, V., Rothman, J., Bialystok, E. & Pliatsikas, C. Redefining bilingualism as a spectrum of experiences that differentially affects brain structure and function. *Proc. Natl. Acad. Sci. USA* **116**, 7565–7574 (2019).
52. Linares, R. & Pelegrina, S. The relationship between working memory updating components and reading comprehension. *Cogn. Process.* **24**, 253–265 (2023).
53. Lukasik, K. M. et al. Bilingualism and working memory performance: Evidence from a large-scale online study. *PLoS One* **13**, e0205916 (2018).
54. Kriegeskorte, N. & Douglas, P. K. Cognitive computational neuroscience. *Nat. Neurosci.* **21**, 1148–1160 (2018).
55. Follmer, D. J., Fang, S.-Y., Clariana, R. B., Meyer, B. J. F. & Li, P. What predicts adult readers' understanding of STEM texts?. *Read. Writ.* **31**, 185–214 (2018).
56. Wiederholt, J. L. & Blalock, G. *GSRT: Gray Silent Reading Tests* (Pro-Ed., 2000).
57. Rief, S. F. & Stern, J. *The Dyslexia Checklist: A Practical Reference for Parents and Teachers*, Vol. 3 (John Wiley & Sons, 2010).
58. Dunn, L. M. & Dunn, D. *The Peabody Picture Vocabulary Test* (NCS Pearson, Inc, Minnesota, 2007).
59. Fan, J. et al. Testing the behavioral interaction and integration of attentional networks. *Brain Cogn.* **70**, 209–220 (2009).
60. Huang, F., Lin, G., Meng, Y., Lin, Y. & Zheng, S. The role of alerting in the attentional boost effect. *Front. Psychol.* **14**, 1075979 (2023).

61. Antón, E. et al. Is there a bilingual advantage in the ANT task? Evidence from children. *Front. Psychol.* **5**, 398 (2014).
62. Li, P., Zhang, F., Yu, A. & Zhao, X. Language history questionnaire (LHQ3): an enhanced tool for assessing multilingual experience. *Biling. Lang. Cogn.* **23**, 938–944 (2020).
63. Henderson, J. M., Choi, W., Lowder, M. W. & Ferreira, F. Language structure in the brain: a fixation-related fMRI study of syntactic surprisal in reading. *Neuroimage* **132**, 293–300 (2016).
64. Todd, N. et al. Evaluation of 2D multiband EPI imaging for high-resolution, whole-brain, task-based fMRI studies at 3T: sensitivity and slice leakage artifacts. *Neuroimage* **124**, 32–42 (2016).
65. Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* **59**, 2142–2154 (2012).
66. Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S. & Kanwisher, N. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *J. Neurophysiol.* **104**, 1177–1194 (2010).
67. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670 (2011).
68. la Tour, T. D., Eickenberg, M., Nunez-Elizalde, A. O. & Gallant, J. L. Feature-space selection with banded ridge regression. *Neuroimage* **264**, 119728 (2022).
69. Hoover, W. A. & Gough, P. B. The simple view of reading. *Read. Writ.* **2**, 127–160 (1990).
70. O'brien, R. M. A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.* **41**, 673–690 (2007).

Acknowledgements

The Reading Brain Project was supported by the NSF (#NCS-FO-1533625) and the work reported here is supported by the Hong Kong Research Grants Council (Project #PolyU15607623), the Hong Kong Polytechnic University and the Sin Wai Kin Foundation. C.G. was supported by a postgraduate scholarship from the Hong Kong Polytechnic University. We would also like to acknowledge the computational support from PolyU University Research Facility in Big Data Analytics (UBDA) and the University Research Facility in Behavioral and Systems Neuroscience (UBSN).

Author contributions

C.G.: conceptualization, investigation, formal analysis, writing—original draft, review and editing. S.A.N.: methodology, supervision and writing—review and editing. Z.Z.: methodology and writing—review. P.L.: conceptualization, supervision, funding acquisition, equipment access, and writing—original draft, review and editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41539-025-00337-y.

Correspondence and requests for materials should be addressed to Ping Li.