

# Driving and suppressing the human language network using large language models

Received: 6 May 2023

Accepted: 10 November 2023

Published online: 3 January 2024

 Check for updates

Greta Tuckute  , Aalok Sathe  , Shashank Srikant<sup>3,4</sup>, Maya Taliaferro<sup>1,2</sup>, Mingye Wang<sup>1,2</sup>, Martin Schrimpf  , Kendrick Kay<sup>7</sup> & Evelina Fedorenko  

Transformer models such as GPT generate human-like language and are predictive of human brain responses to language. Here, using functional-MRI-measured brain responses to 1,000 diverse sentences, we first show that a GPT-based encoding model can predict the magnitude of the brain response associated with each sentence. We then use the model to identify new sentences that are predicted to drive or suppress responses in the human language network. We show that these model-selected novel sentences indeed strongly drive and suppress the activity of human language areas in new individuals. A systematic analysis of the model-selected sentences reveals that surprisal and well-formedness of linguistic input are key determinants of response strength in the language network. These results establish the ability of neural network models to not only mimic human language but also non-invasively control neural activity in higher-level cortical areas, such as the language network.

Reading and understanding this sentence engages a set of left-lateralized frontal and temporal brain regions. These interconnected areas (or the ‘language network’<sup>1–4</sup>) support both comprehension and production of spoken, written and signed linguistic utterances<sup>2,5–7</sup> across diverse languages<sup>8</sup>. These regions are highly selective for language relative to diverse non-linguistic inputs (see ref. 9 for a review), are sensitive to linguistic structure at many levels<sup>2,7,9–11</sup>, and are causally important for language such that damage to these regions leads to linguistic deficits<sup>12,13</sup>. However, many aspects of the representations and algorithms that support language comprehension remain unknown.

Over the past few years, artificial neural networks for language have emerged as *in silico* models of language processing. These large language models (LLMs) can generate coherent text, answer questions,

translate between languages and perform sophisticated language comprehension tasks<sup>14,15</sup>. Strikingly, although the LLMs were not developed with the goal of modelling human language processing, some of these models (especially the unidirectional Transformer architectures<sup>14</sup>) have a remarkable capacity to mimic human language behaviour<sup>16,17</sup> and predict brain activity during language processing<sup>18–23</sup>. But despite LLMs being today’s most quantitatively accurate models of language processing, there has been no attempt to test whether LLMs can causally control language responses in the brain. By ‘causal control’, we mean using models to make quantitative predictions about a neural target (a cell or a brain area/network) and subsequently using those predictions to successfully modulate neural activity in the target in a closed-loop manner.

<sup>1</sup>Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>2</sup>McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>3</sup>Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>4</sup>MIT-IBM Watson AI Lab, Cambridge, MA, USA. <sup>5</sup>Quest for Intelligence, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>6</sup>Neuro-X Institute, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. <sup>7</sup>Center for Magnetic Resonance Research, University of Minnesota, Minneapolis, MN, USA. <sup>8</sup>Program in Speech and Hearing Bioscience and Technology, Harvard University, Cambridge, MA, USA.

 e-mail: [gretatu@mit.edu](mailto:gretatu@mit.edu); [evelina9@mit.edu](mailto:evelina9@mit.edu)

Recent work in visual neuroscience has shown that artificial neural network models for image recognition can causally intervene in the non-human primate visual system by generating visual stimuli that modulate activity in different regions of the ventral visual pathway<sup>24,25</sup>. In this work, we ask whether similar model-based control is feasible for the higher-level cognitive domain of language: can we leverage the predictive power of LLMs to identify new stimuli to maximally drive or suppress brain responses in the language network of new individuals? This question taps into two key aspects of the generalization ability of LLMs. First, do LLMs capture features of language representations that generalize across humans? Second, do LLMs have the capacity to predict brain responses to model-selected stimuli that extend beyond the distribution of naturally occurring linguistic input? We demonstrate that model-selected stimuli drive and suppress brain responses in the language network of new individuals, establishing the ability of brain-aligned LLMs to non-invasively control areas implicated in higher-level cognition. We then leverage sentence-level brain responses to a broad distribution of linguistic input to ask what kinds of linguistic input the language network is most responsive to. In a large-scale behavioural experiment, we collect rating norms for ten sentence properties and use these norms to characterize the language network's preferred stimuli.

## Results

### Model-selected sentences control language network responses

Our aim was to test whether current models of the human language network are capable of driving and suppressing brain responses in these higher-level cognitive brain areas. We developed an encoding model of the left hemisphere (LH) language network in the human brain with the goal of identifying new sentences that would activate the language network to a maximal or minimal extent. The model takes as input last-token sentence embeddings from GPT2-XL<sup>14</sup> (previously identified as the most brain-aligned language model<sup>20</sup>; layer 22, see Supplementary Information section 6a for the cross-validated analysis that led to this choice) and was trained, via ridge regression, to predict the average LH language network's (functionally defined<sup>2</sup>) blood-oxygen-level-dependent (BOLD) response (also referred to as the language network response; see 'Definition of ROIs' in Methods). The BOLD responses were acquired from five train participants who read a set of 1,000 diverse, corpus-extracted sentences (baseline set) (two sessions each,  $n = 10$  sessions total; see 'Encoding model development' in Methods) (Fig. 1a). The encoding model achieved a prediction performance of  $r = 0.38$  (the noise ceiling (NC) is  $r = 0.56$ ; Supplementary Information section 5) when evaluated on held-out sentences within the baseline set (s.e. over five splits, 0.016; all five  $P$  values,  $<0.001$ ; Supplementary Information section 6a). To ensure that the encoding model performance did not hinge on specific experimental decisions, we confirmed that the model maintained high predictivity performance on held-out sentences when changing the procedure for obtaining sentence embeddings (the average of all tokens in the sentence; Supplementary Information section 6b) and even using sentence embeddings from a different LLM architecture (a bidirectional-attention Transformer model, BERT-large; Supplementary Information section 6c). Furthermore, the encoding model also achieved relatively high predictive performance on anatomically, rather than functionally, defined language regions, although predictivity was lower (Supplementary Information section 6d).

To identify sentences that would elicit a desired (high or low) level of activation in the language network, we searched across ~1.8 million sentences from nine diverse large-scale text corpora (Fig. 1b). We identified a set of 250 sentences that were predicted to elicit maximally strong activity in the language network (drive sentences; for example, 'Turin loves me not, nor will.' or 'People on Insta Be Like, "Gross!"') and 250 sentences that were predicted to elicit minimal activity in the

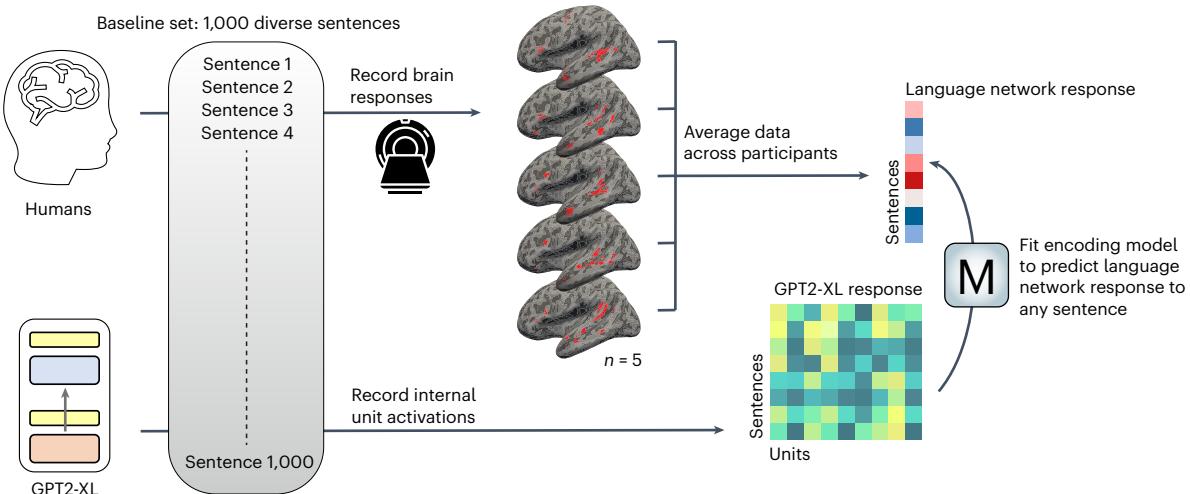
language network (suppress sentences; for example, 'We were sitting on the couch.' or 'Inside was a tiny silver sculpture.'). We evaluated our encoding model by recording brain responses to these new drive and suppress sentences in new participants (denoted as evaluation participants) (note the fully independent procedure using both new stimuli and new participants; for evidence that the new drive and suppress sentences differ from the baseline sentences, see Supplementary Information sections 10 and 11).

We collected functional MRI (fMRI) responses to the drive and suppress sentences in an event-related, single-trial paradigm with three new participants (three sessions each,  $n = 9$  sessions total; see 'Encoding model evaluation' in Methods). The drive and suppress sentences were randomly interspersed among the 1,000 baseline sentences. Figure 2b shows the average responses for the  $n = 3$  evaluation participants for the drive, suppress and baseline sentence conditions. The drive sentences yielded significantly higher responses than the suppress sentences ( $\beta = 0.57$ ,  $t = 15.93$ ,  $P < 0.001$  using linear mixed effects (LME) modelling; see 'Statistical analyses' in Methods and Supplementary Information section 18). The drive sentences also yielded significantly higher responses than the baseline sentences ( $\beta = 0.27$ ,  $t = 9.72$ ,  $P < 0.001$ ), with the evoked BOLD signal being 85.7% higher for the drive condition than for the baseline (quantified using non-normalized BOLD responses; Supplementary Information section 12a). Finally, the suppress sentences yielded lower responses than the baseline sentences ( $\beta = -0.29$ ,  $t = -10.44$ ,  $P < 0.001$ ), with the evoked BOLD responses being 97.5% lower for the suppress condition than for the baseline (Supplementary Information section 12a). In summary, we trained an encoding model to generate predictions about the magnitude of activation in the language network for a new set of sentences and then 'closed the loop' by collecting brain data for these new sentences from new participants to demonstrate that these sentences modulate brain responses as predicted. We note that although we trained the encoding model using the responses in the LH language network as a whole, the five individual LH language functional regions of interest (fROIs) showed highly correlated responses across the baseline set (Supplementary Information section 4 and 'Language regions exhibit high stimulus-related activity') and similar condition-level responses to the drive, suppress and baseline sentences (Supplementary Information section 15f) (see Supplementary Information section 15g for evidence that this pattern of responses to drive, suppress and baseline sentences is not ubiquitously present across the brain). These inter-fROI similarities align well with past work showing similar modulation of the different language areas by diverse linguistic manipulations<sup>7,8,11,26–29</sup>.

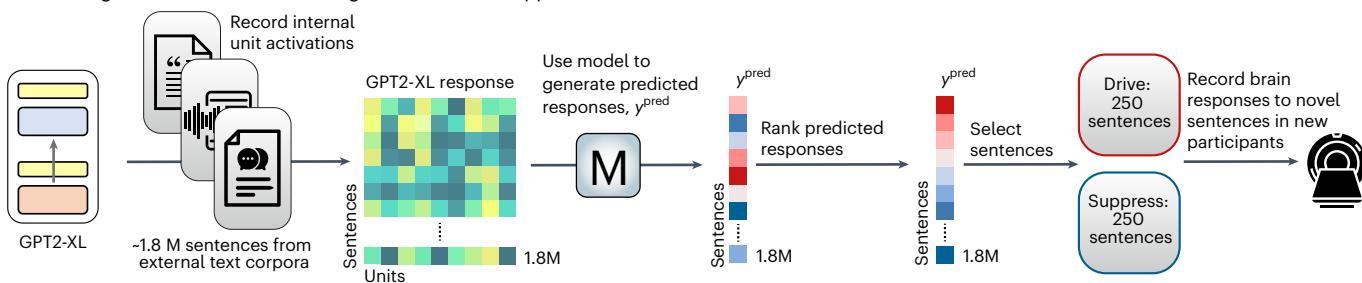
To further validate the robustness of responses to the drive and suppress sentences, we collected brain data for a large subset of the drive, suppress and baseline stimuli in a traditional blocked fMRI design, where drive, suppress and baseline sentences were blocked into groups, with four additional participants (one session each,  $n = 4$  sessions total; see 'fMRI experiments' in Methods). The results mirrored those from the event-related experiment: the drive sentences yielded the highest response, followed by the baseline sentences (the evoked BOLD response was 12.9% higher for drive than for baseline and 56.6% lower for suppress than for baseline; Supplementary Information section 12b). Hence, independent of experimental design (event-related versus blocked) and modelling procedure (single-trial modelling versus condition-level modelling), the brain responses to the drive sentences were higher than those to the baseline sentences, and the responses to the suppress sentences were lower than those to the baseline sentences.

For a final examination of model-guided stimulus selection, we explored an alternative approach to selecting drive/suppress sentences: the 'modify' approach, where, instead of searching within existing text corpora, we used gradient-based modifications to transform a random sentence into a novel sentence predicted to elicit high or low fMRI responses (Supplementary Information section 16a) and

**a Encoding model development: fitting a predictive model of the language network**



**b Encoding model evaluation: selecting novel drive and suppress sentences**



**Fig. 1 | Overview of the procedure for encoding model development and stimulus selection for evaluation.** **a**, We developed an encoding model (M) of the LH language network in the human brain with the goal of identifying novel sentences that activate the language network to a maximal or minimal extent (see ‘Encoding model development’ in Methods). Five participants (train participants) read a large sample ( $n = 1,000$ ) of six-word corpus-extracted sentences, the baseline set (sampled to maximize linguistic diversity; Supplementary Information section 1), in a rapid, event-related design while their brain activity was recorded using fMRI. BOLD responses from voxels in the LH language network were averaged within each train participant and averaged across participants to yield an average language network response to each of the 1,000 baseline set sentences. We trained a ridge regression model from the representations of the unidirectional-attention Transformer language model, GPT2-XL (identified as the most brain-aligned language base model in Schrimpf et al.<sup>20</sup>), to the 1,000 averaged fMRI responses. Given that GPT2-XL can generate a representation for any sentence, the encoding model (M) can predict the LH language network response for arbitrary sentences. To select the top-performing layer for our encoding model, we evaluated all 49 layers of GPT2-XL and selected the layer that had highest predictivity performance on brain responses to held-out baseline set sentences (layer 22; Supplementary Information section 6a).

**b**, To evaluate the encoding model (M), we identified a set of sentences to activate the language network to a maximal extent (drive sentences) or a minimal extent (suppress sentences) (see ‘Encoding model evaluation’ in Methods). To do so, we obtained GPT2-XL embeddings for ~1.8 million sentences from diverse, large text corpora, generated predicted language network responses and ranked these responses to select the sentences that are predicted to increase or decrease brain responses relative to the baseline set. Finally, we collected brain responses to these novel sentences in new participants (evaluation participants).

collected responses to these novel sentences from two participants (event-related design,  $n = 6$  sessions total). We found that this exploratory modify approach was able to drive responses by 57% relative to the baseline but failed to suppress responses, most likely because the resulting modify stimuli were often akin to word lists, which the encoding model was not trained on (Supplementary Information section 16b).

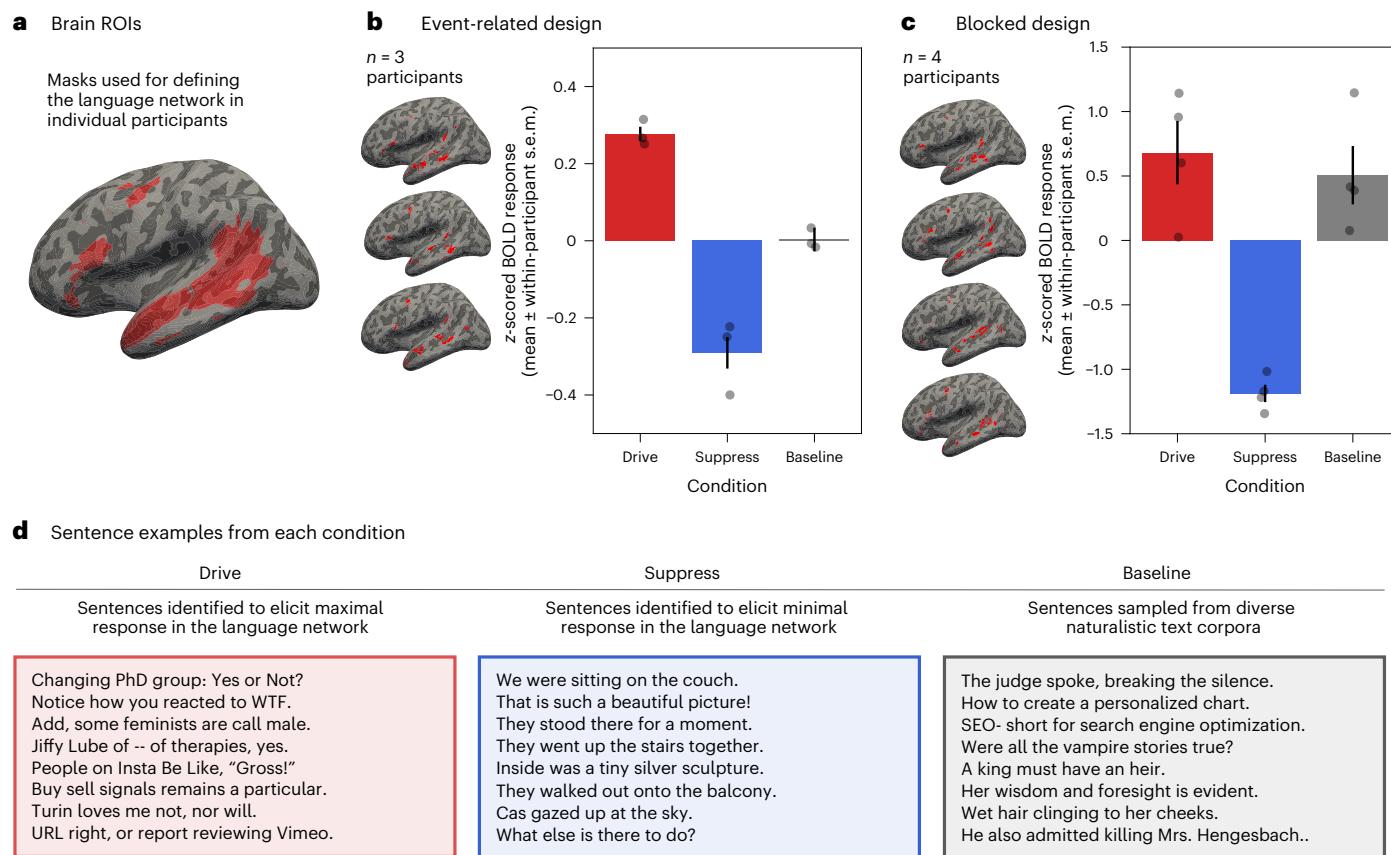
### The model captures most explainable variance in new participants

In the previous section, we examined predictivity at the condition level (drive versus suppress versus baseline). Here we sought to evaluate the accuracy of the predictions from the encoding model at the level of individual sentences. To do so, we turned to the event-related experiment (Fig. 2b), which allows us to estimate sentence-level brain responses to 1,500 sentences for each of the three evaluation participants.

Figure 3 shows the model-predicted versus observed brain responses in the language network ( $n = 3$  evaluation participants). These participants were not used to train the encoding model and hence allowed us to estimate encoding model predictivity performance

in held-out participants and held-out sentences. Across the full set of 1,500 baseline, drive and suppress sentences, we obtained a Pearson correlation of 0.43 (d.f. = 1,498,  $P < 0.001$ ,  $t = 18.60$ , s.e. = 0.02) between predicted and observed brain responses. Because the drive and suppress sentences were designed to elicit high or low brain responses, respectively, one might expect that the correlation might be unduly driven by these two conditions. We therefore isolated the set of  $n = 1,000$  naturalistic, corpus-extracted baseline sentences and obtained a correlation of 0.30 (d.f. = 998,  $P < 0.001$ ,  $t = 9.88$ , s.e. = 0.03). Hence, the encoding model was able to predict a substantial and statistically significant amount of variance in brain responses in new participants to both naturalistic sentences that fall within the distribution of the training (baseline) set and out-of-distribution sentences (drive/suppress set), for which encoding model predictions (the x axis in Fig. 3) extend far beyond the training set distribution.

To better interpret the accuracy of the sentence-level predictions, we quantified the maximum possible prediction performance by treating inter-participant variability as ‘noise’ that cannot be predicted by a computational model. The goal here is to assess how



**Fig. 2 | Model-selected sentences successfully drive and suppress responses in the language network.** **a**, We used our encoding model to select sentences that would elicit maximal response (drive sentences) or minimal response (suppress sentences) in the functionally defined language network. To define fROIs, we used demarcations ('language parcels'; shown on the surface-inflated Montreal Neurological Institute (MNI) 152 template brain) within which most or all individuals in prior studies showed activity for the language localizer contrast in large samples (for example, refs. 2,4). We defined the LH language network as regions within the borders of these five parcels that were activated (top 10%) in the functional localizer acquired for each participant (see the brain visualizations in **b** and **c**). **b**, The average language network fMRI response across 250 drive, 250 suppress and 1,000 baseline sentences for *n* = 3 evaluation participants, collected in an event-related, single-trial fMRI paradigm. In both **b** and **c**, the individual points show the average of each condition per participant. fMRI responses were z-scored session-wise (see Supplementary Information section 12a for the responses without normalization; no key patterns are affected). The

evoked BOLD response was 85.7% higher for drive than for baseline and 97.5% lower for suppress than for baseline (Supplementary Information section 12a). The error bars show the within-participant standard error of the mean. The brain illustrations show the functionally defined language network in the participants of interest on the surface-inflated brain, visualized in Freeview. For the surface projections, volumetric data (in MNI IIXI549Space; SPM12 (ref. 100)) were registered to FreeSurfer's CVS35 (combined volumetric and surface-based) in the MNI152 space using mri\_volsurf in FreeSurfer v.7.3.2 (ref. 101) with a projection distance of 1.5 mm and otherwise default parameters. **c**, The average language network fMRI response across 240 drive, 240 suppress and 240 baseline sentences (randomly sampled from the superset of 250 drive, 250 suppress and 1,000 baseline sentences) for *n* = 4 evaluation participants, collected in a blocked fMRI paradigm. The evoked BOLD response was 12.9% higher for drive than for baseline and 56.6% lower for suppress than for baseline (Supplementary Information section 12b). **d**, Example sentences from each condition.

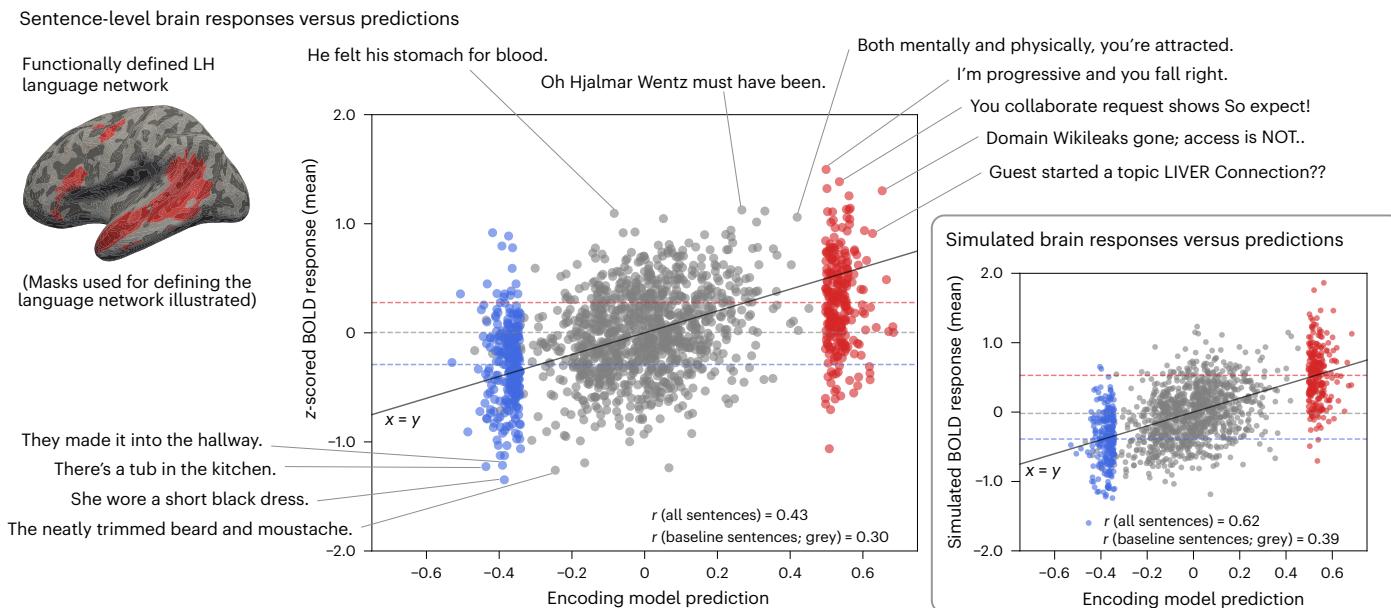
well our model predicts brain activity at the group level, taking into account irreducible variance due to inter-participant variability and measurement noise. First, we computed the empirical variability in participants' responses to the 1,500 sentences. Next, we simulated response noise for each participant using the empirical variability across participants (drawing samples from a Gaussian distribution with zero mean and the empirical inter-participant standard deviation). For each sentence, simulated response noise was added to the encoding model's predicted response (the *x* axis in Fig. 3), and the responses were then averaged across participants. This simulation provides an estimate of the maximum possible prediction performance of the encoding model.

The inset in Fig. 3 shows these simulated brain responses versus the predicted responses. In these simulations, the Pearson correlation was 0.62 (d.f. = 1,498, *P* < 0.001, *t* = 30.85, s.e. = 0.02) across all 1,500 sentences (observed: *r* = 0.43—that is, 69.4% of the theoretically obtainable correlation) and 0.39 (d.f. = 998, *P* < 0.001, *t* = 13.32, s.e. = 0.03)

across the 1,000 baseline sentences (observed: *r* = 0.30—that is, 76.9% of the theoretically obtainable correlation). These results show that due to inter-participant variability in fMRI measurements, even a perfect model can achieve only *r* = 0.62 predictive performance. Although our model is not perfect, the performance level suggests that the model successfully captures much of the neurally relevant variance in responses to individual sentences.

### Language regions exhibit high stimulus-related activity

Having established that model-selected stimuli could indeed drive and suppress brain responses in the language network of new individuals (Figs. 2 and 3), our next goal was to investigate what kinds of linguistic input the LH language network is most responsive to. Before delving into that investigation, however, we wanted to assess whether the LH language regions show reliable responses to and track properties of linguistic stimuli. We also wanted to assess the similarity among the language fROIs in their fine-grained linguistic preferences to decide



**Fig. 3 | The encoding model maintains high predictive performance for brain responses from three new participants to out-of-distribution sentences.**

Sentence-level brain responses as a function of the predicted responses along with sentence examples. Predicted brain responses were obtained from the encoding model ( $x$  axis). The observed brain responses ( $y$  axis) are the average of  $n = 3$  evaluation participants' language network responses (illustrated for individual participants in Supplementary Information section 13). The blue points represent the suppress sentences, the grey points represent the baseline sentences and the red points represent the drive sentences. The suppress and

drive sentences were selected to yield low or high brain responses, respectively, and are therefore clustered on the low and high ends of the prediction axis ( $x$  axis). The dashed horizontal lines show the mean of each condition. The inset shows the simulated sentence-level brain responses as a function of predicted responses. Predicted brain responses were obtained from the encoding model ( $x$  axis). The simulated brain responses ( $y$  axis) were obtained by sampling from a noise distribution representing the empirical inter-participant variability. This plot illustrates the maximum possible predictive performance, given inter-participant variability and fMRI measurement noise.

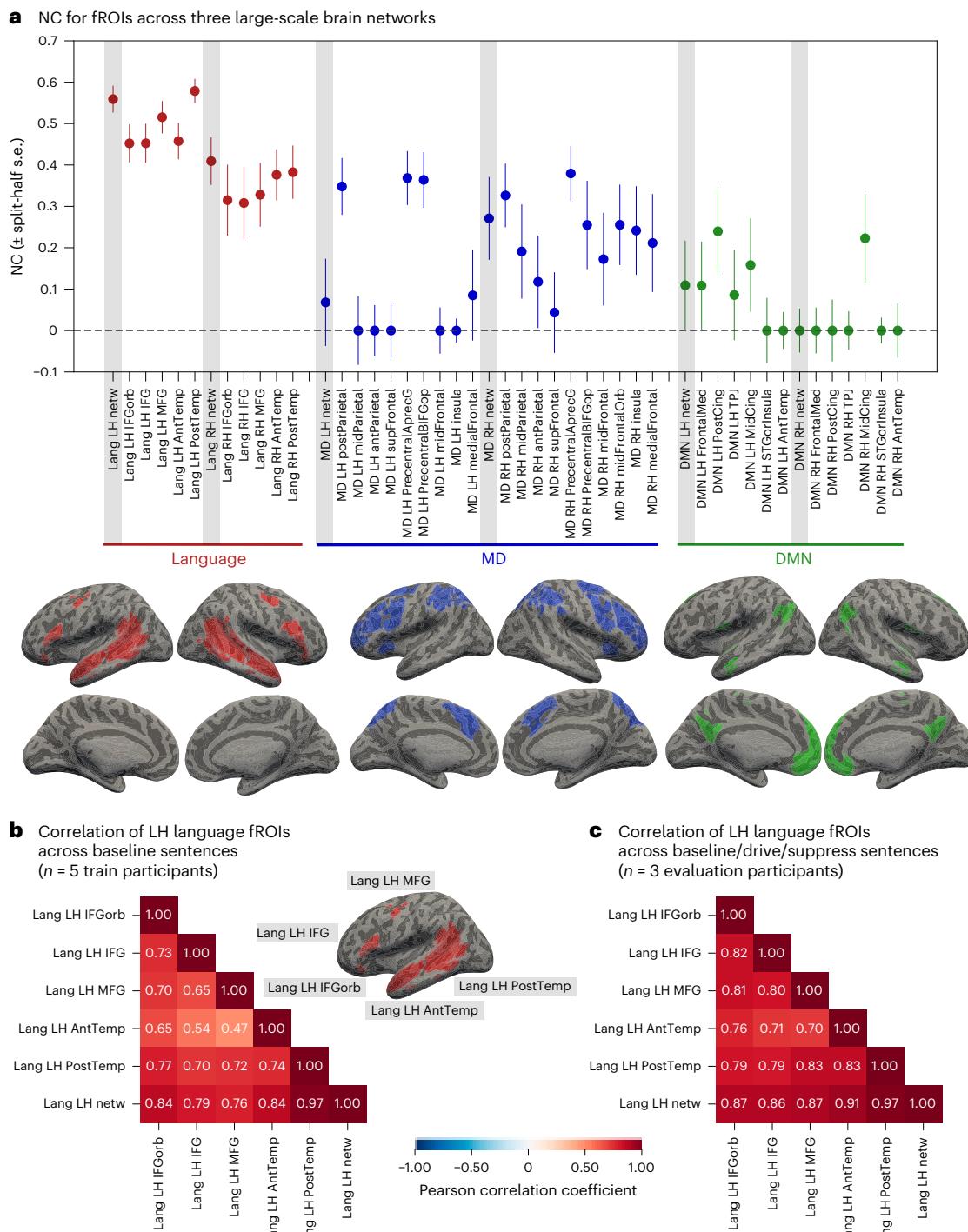
whether it may be worth examining the fROIs separately in addition to examining the language network as a whole.

First, we quantified NCs for the language regions along with a set of control brain regions (Fig. 4a). An NC for a brain region is a measure of stimulus-related response reliability and is typically expressed in terms of the fraction of variance that can be attributed to the stimulus rather than to measurement noise. Standard approaches for NC estimation leverage repeated stimulus presentations, with the core assumption that repeated presentations should yield the same brain response<sup>30,31</sup>. Because in the current study, each sentence was presented only once to a given participant (for the motivation for this design choice and details of the procedure, see the Methods and Discussion), we developed a procedure for NC estimation that uses the repeated presentations of the same sentence across participants, allowing for estimation of reliability in single-repetition paradigms (Supplementary Information section 5). Using this procedure, we computed NCs on the basis of the brain responses to the 1,000 baseline sentences for the  $n = 5$  train participants in the language regions and a set of control regions (Fig. 4a). In particular, we examined two large-scale brain networks that have been linked to high-level cognitive processing—the multiple demand (MD) network<sup>32</sup> and the default mode network (DMN)<sup>33</sup>—which we defined using independent functional localizers (see Supplementary Information section 15 for the details) (Fig. 4a). For additional comparison, we examined a set of anatomical parcels<sup>34</sup> that cover a large fraction of the cortical surface (Supplementary Information section 8).

Prior studies have demonstrated high consistency of responses in language regions across participants using naturalistic story-listening paradigms<sup>35–37</sup>. In line with those studies, we found that in our single-sentence paradigm, language regions were also characterized by high NCs. The ceiling values were higher than those observed in the two other functional networks (Fig. 4a) and in anatomical areas across the brain (including anatomical areas that fall in spatially similar

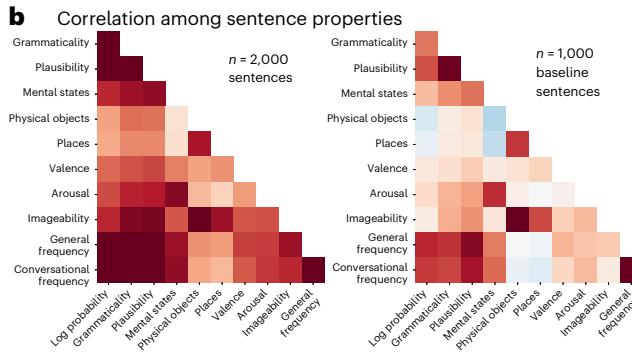
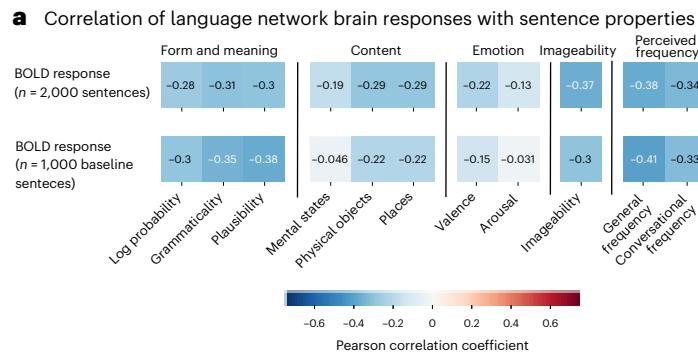
locations to the language areas, which provides further evidence for the advantages of functional localization<sup>38</sup>; Supplementary Information section 8). In particular, for the LH language areas, the NC was estimated to be  $r = 0.56$  (split-half s.e., 0.03); that is, -31% of the variance in the responses of these areas at the group level can be considered ‘true’, stimulus-related signal. For comparison, for the MD network, the NC was estimated to be  $r = 0.07$  (s.e. = 0.11) (for the LH MD areas) and  $r = 0.27$  (s.e. = 0.10) (for the right hemisphere (RH) MD areas; see ref. 37 for convergent evidence from a different approach), and for the DMN, the NC was estimated to be  $r = 0.11$  (s.e. = 0.11) (for the LH DMN areas) and  $r = 0$  (s.e. = 0.05) (for the RH DMN areas). The LH language network NC values were significantly higher than the NC in each of these four networks—LH and RH MD and DMN (d.f. = 3,998, all four  $P < 0.001$ , all four  $t > 126$  via Bonferroni-corrected two-sided independent  $t$ -tests using split-half bootstrap NC values). Thus, other brain regions implicated in high-level cognition (MD and DMN) were not as reliable as the language regions in their responses to linguistic stimuli (and similarly not as well predicted by GPT2-XL features; Supplementary Information section 8). In summary, the high NCs of the language regions show that these regions process stimulus-related information in a similar way across participants (see also refs. 35–37), opening the door to investigations of what stimulus properties affect neural responses (see the next section).

Second, we examined whether the five regions that comprise the LH language network are similar in their responses at the fine-grained level of single sentences. Prior work has demonstrated that the LH language regions exhibit similar functional response profiles in terms of their selectivity for language relative to non-linguistic inputs (see ref. 9 for a review) and similar sensitivity to diverse linguistic manipulations<sup>7,11,29</sup>, as well as highly correlated time courses during naturalistic paradigms<sup>8,27,28,39</sup>. Here we investigated whether the five LH language regions have similar preferences for some sentences over others across  $n = 1,000$  or  $n = 1,500$  sentences.

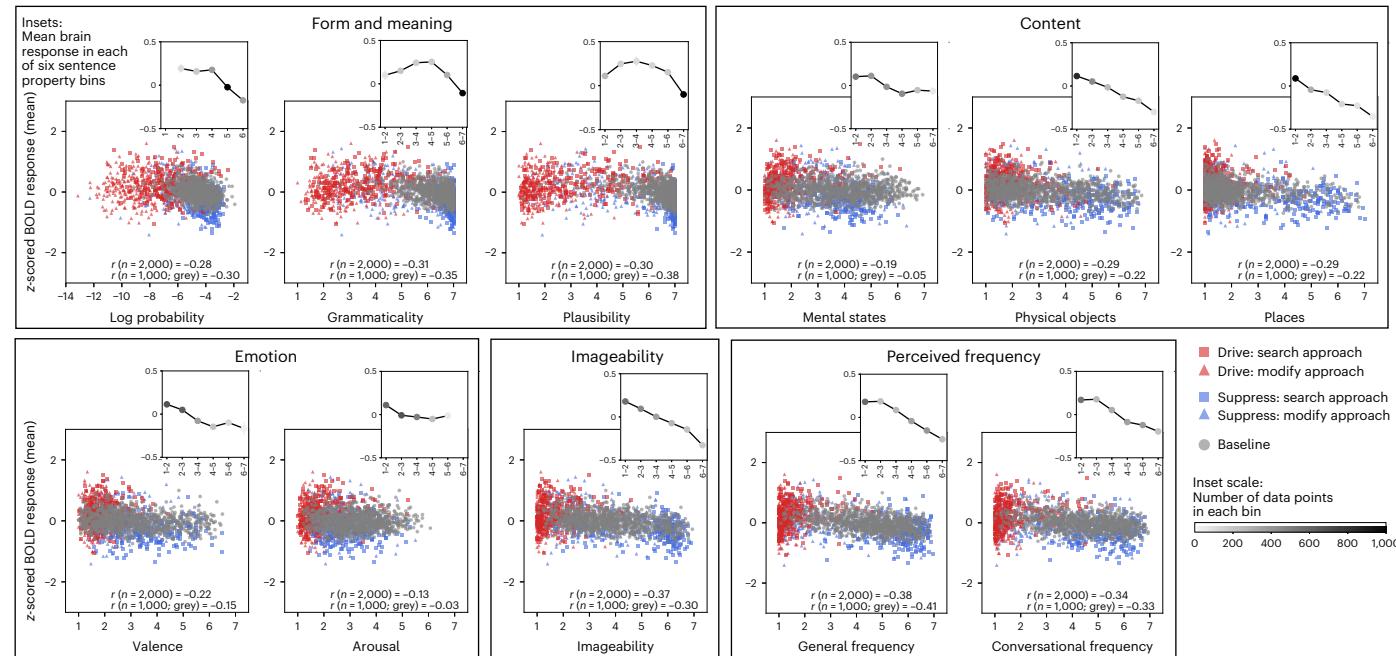


**Fig. 4 | LH language regions show a high degree of stimulus-related activity for linguistic input relative to other brain areas, and the LH language regions show functionally similar responses.** **a**, We quantified the noise ceiling (NC), a measure of stimulus-related response reliability, across all functionally defined ROIs in the language network (red), the MD network (blue) and the DMN (green). For the language network, we defined 10 such fROIs (along with the ‘Language LH/RH network’ fROI, which is the mean across all voxels in the fROIs within the network and hemisphere, yielding 12 ROIs in total); for the MD network, we defined 21 fROIs (note that one participant did not show a response to the MD localizer in the MD LH midFrontalOrb ROI, and hence this ROI was excluded in the NC computation); and for the DMN network, we defined 14 fROIs. The grey shaded areas indicate the network-level fROIs. The points show the NC estimate computed across  $n = 1,000$  baseline sentences across  $n = 5$  train participants, for each of the ROIs. The error bars show the NC reliability quantified as the standard

error over NC values computed from 1,000 splits of the data (Supplementary Information section 5b). The brain illustrations show the anatomical parcels (demarcations) that were used to constrain the selection of participant-specific fROIs for each network on the surface-inflated MN152 template brain. **b**, The Pearson correlation matrix computed over  $n = 1,000$  baseline sentences for the average of  $n = 5$  train participants. The first five rows/columns show the five core LH language fROIs (IFGorb, IFG, MFG, AntTemp and PostTemp; see ‘Definition of ROIs’ in Methods). The sixth row/column shows the full LH language network consisting of the average of the voxels from the five fROIs; these values show how representative the language network as a whole is of each of the five fROIs. **c**, Same as in **b**, but for the  $n = 1,500$  drive/suppress/baseline sentences for the average of  $n = 3$  evaluation participants (derived using the main, search approach). Correlation matrices for individual participants are shown in Supplementary Information section 4.



**c Scatterplots of brain responses versus sentence properties**



**Fig. 5 | Surprisal and several other sentence properties modulate responses in the language network.** **a**, Correlation of the LH language network response with 11 sentence properties (columns) in five categories for all  $n = 2,000$  sentences (first row; drive, suppress and baseline sentences averaged across  $n = 5$  train and  $n = 5$  evaluation participants) and for  $n = 1,000$  baseline sentences (second row; similarly averaged across  $n = 5$  train and  $n = 5$  evaluation participants). **b**, Correlation among the sentence properties shown for either  $n = 2,000$  sentences (left) or  $n = 1,000$  sentences (right). The colour scale is the same as in **a**. **c**, Sentence-level brain responses as a function of sentence property. The brain responses (y-axis) were averaged across  $n = 5$  train and  $n = 5$  evaluation participants. The sentence properties were derived from behavioural norming experiments in independent participants (besides the ‘log probability’ feature, which was estimated using GPT2-XL). The inset line graphs show the average brain response with each property grouped into six uniformly spaced bins. The error bars show the standard error of the mean across items in each bin (often not visible given the large number of data points). For the behavioural norms, the bins were defined according to the rating scale—that is, [1,2], [2,3], [3,4], [4,5], [5,6] and [6,7]. For log probability, the bins were similarly uniformly spaced, but according to the range of surprisal values: [-13.1, -11.3], [-11.3, -9.4], [-9.4, -7.5], [-7.5, -5.7], [-5.7, -3.8] and [-3.8, -1.9] (omitted in the x-axis label). The shade of the points in these graphs denotes the amount of data in each bin (darker points correspond to larger amounts of data; bins containing less than 1% of the data (that is, 20 responses) were omitted from the line graphs). Statistical comparisons accompanying the inset plots can be found in Supplementary Information section 24.

brain response with each property grouped into six uniformly spaced bins. The error bars show the standard error of the mean across items in each bin (often not visible given the large number of data points). For the behavioural norms, the bins were defined according to the rating scale—that is, [1,2], [2,3], [3,4], [4,5], [5,6] and [6,7]. For log probability, the bins were similarly uniformly spaced, but according to the range of surprisal values: [-13.1, -11.3], [-11.3, -9.4], [-9.4, -7.5], [-7.5, -5.7], [-5.7, -3.8] and [-3.8, -1.9] (omitted in the x-axis label). The shade of the points in these graphs denotes the amount of data in each bin (darker points correspond to larger amounts of data; bins containing less than 1% of the data (that is, 20 responses) were omitted from the line graphs). Statistical comparisons accompanying the inset plots can be found in Supplementary Information section 24.

Figure 4b shows the Pearson correlation across the  $n = 1,000$  baseline sentences for LH fROIs from the average of  $n = 5$  train participants. Correspondingly, Fig. 4c shows the correlation across the  $n = 1,500$  drive/suppress/baseline sentences for LH fROIs from the average of  $n = 3$  evaluation participants. Both plots show high inter-fROI correlations for the LH language network (correlation range, 0.47–0.83), which suggests that even in their fine-grained preferences for particular linguistic stimuli, the LH language fROIs show a high degree of similarity. Along with the prior body of evidence noted above, these high correlations motivated our decision to investigate what kinds of linguistic input engage this network as a whole (see the next section).

**Sentence complexity modulates language network responses**

To gain an understanding of what sentence properties modulate brain responses in the language network, we obtained a set of 11 features to characterize our experimental materials ( $n = 2,000$  sentences: 1,000 baseline, 250 drive and 250 suppress sentences from the search approach, and 250 drive and 250 suppress sentences from the exploratory modify approach; Supplementary Information section 16) and correlated these features with sentence-level brain responses (see ‘Sentence properties that modulate brain responses’ in Methods). The choice of features was inspired by past work in linguistics/psycholinguistics and cognitive neuroscience of language. First, building on prior evidence that surprisal (the degree of contextual predictability, which is

typically estimated as the negative log probability) modulates language processing difficulty in both behavioural psycholinguistic work<sup>40–42</sup> and brain imaging investigations<sup>43–47</sup>, we computed sentence-level log probability estimates for each of 2,000 sentences using GPT2-XL (see ‘Sentence properties that modulate brain responses’ in Methods). Second, we collected ten behavioural rating norms from a total of  $n = 3,600$  participants (on average, 15.23 participants per sentence per norm; minimum, 10; maximum, 19). The norms spanned five broad categories and were selected on the basis of prior behavioural (for example, refs. 40,48–50) and neural studies (for example, refs. 11,51–53). The first category targeted two core aspects of sentences: grammatical well-formedness (how much does the sentence obey the rules of English grammar?) (for details of the instructions, see Supplementary Information section 22c) and plausibility (how much sense does the sentence make?). Because sentence surprisal (log probability), as estimated with GPT2-XL, is likely to capture both of these aspects to some extent<sup>54–56</sup>, we grouped these two norms with surprisal in the analyses. Furthermore, because surprisal probably captures diverse aspects of form and meaning more generally, we examined the norm–brain relationships for all other norms after factoring out variance due to surprisal. Inspired by work on distributed neural representation of meaning, including across the language network<sup>57,58</sup>, the next three norms probed different aspects of the sentence’s content: how much does the sentence make you think about (1) others’ mental states, (2) physical objects and their interactions, and (3) places and environments? The latter two have to do with the physical world, and the former, with internal representations; the physical-versus-social distinction is one plausible organizing dimension of meaning<sup>59,60</sup>. Two norms probed emotional dimensions of the sentences: valence (how positive is the sentence’s content?) and arousal (how exciting is the sentence’s content?). One norm targeted visual imagery (how visualizable is the sentence’s content?). Finally, the last two norms probed people’s perception of how common the sentence is, in general versus in conversational contexts.

Figure 5a shows the correlation between the language network response and each of the 11 sentence properties across the five categories. The sentences spanned a broad range of brain responses, as evidenced in the sentence-level scatter plots in Fig. 5c (yaxis). Importantly, this broad range was made possible by our approach of specifically designing stimuli to drive and suppress neural responses. Notice how the drive and suppress sentences cover parts of the linguistic space that are barely covered by the set of naturalistic baseline sentences (for comparisons of linguistic properties among conditions, see Supplementary Information section 19).

In terms of the effects of different sentence properties on neural responses, first, we found that less probable (that is, more surprising) sentences elicited higher brain responses (Fig. 5c) ( $r = -0.30$  for the  $n = 1,000$  baseline sentences, d.f. = 998,  $P < 0.001$ ,  $t = -9.83$ , s.e. = 0.03; see Fig. 5c for the correlation values for the full set of  $n = 2,000$  sentences and Supplementary Information section 20 for robustness to model choice to derive surprisal). This result aligns with previous evidence for a positive effect of surprisal on brain responses in MEG/EEG<sup>45,47</sup> and fMRI<sup>43,44,46</sup>. Similarly, for the predictors related to a sentence’s grammatical and plausibility, sentences that were rated as less grammatical or plausible elicited higher responses ( $r = -0.31$ ,  $r = -0.30$ , d.f. = 998,  $t = -11.92$ ,  $t = -12.79$ , both  $P < 0.001$ , both s.e. = 0.03; the two norms were correlated with each other at  $r = 0.74$ ). To understand whether grammaticality or plausibility explained variance above and beyond surprisal and each other, we fitted LME models with different sets of sentence properties as predictors and compared these using likelihood ratio tests (see ‘Statistical analyses’ in Methods and Supplementary Information section 23). Plausibility explained variance beyond surprisal and grammaticality ( $\chi^2 = 17.86$ ,  $P < 0.001$ ; all likelihood ratio statistics are reported on the baseline set). Similarly, grammaticality explained variance beyond surprisal and plausibility ( $\chi^2 = 12.97$ ;  $P < 0.001$ ), albeit to a lesser extent. Interestingly,

a finer-grained examination of the relationship between these features and neural responses reveals a nonlinearity, such that sentences in the mid-range of grammaticality and plausibility elicit stronger responses than sentences on the lower and higher ends of the scales (Fig. 5c). This pattern suggests that two effects may be at play: an increase in neural response is seen (1) for sentences that better adhere to form and meaning regularities of language (similar to the previously reported stronger responses to sentences than to lists of words<sup>2,61,62</sup>) and (2) for sentences that may have greater processing costs due to their unexpected form and/or meaning (for example, see refs. 44,63 for evidence of a strong relationship between behavioural processing difficulty and the strength of neural response in the language areas).

For the properties that relate to the sentence content, we found no increase in explained variance (beyond surprisal) related to whether the sentence concerned others’ mental states ( $\chi^2 = 0.69$ ,  $P = 0.407$ ). This finding aligns with evidence that the language network does not support mental state inference and is robustly dissociated from the theory of mind network (for example, ref. 28), and challenges claims that the language areas are modulated by social content (for example, refs. 64,65). However, whether the sentence’s content concerned physical objects or places correlated negatively with brain responses (both  $r = -0.22$ , d.f. = 998, both  $P < 0.001$ ,  $t = -7.04$  and  $t = -7.11$ , both s.e. = 0.03) and explained variance beyond surprisal (physical objects:  $\chi^2 = 74.26$ ,  $P < 0.001$ ; places:  $\chi^2 = 63.47$ ,  $P < 0.001$ ; the two norms were correlated with each other at  $r = 0.53$ ). Note, however, that these two aspects of the sentence content were also strongly correlated with imageability (discussed below), which may be the underlying driver of these effects.

For the properties that relate to the emotional aspects of sentences, we found that valence correlated negatively with brain responses, such that more positive sentences elicited a lower response ( $r = -0.15$ , d.f. = 998,  $P < 0.001$ ,  $t = -4.69$ , s.e. = 0.03), and it explained some variance beyond surprisal ( $\chi^2 = 16.53$ ,  $P < 0.001$ ). In contrast, whether the sentence was exciting did not explain additional variance beyond surprisal ( $r = -0.03$ , d.f. = 998,  $P = 0.329$ ,  $t = -0.98$ , s.e. = 0.03; likelihood ratio,  $\chi^2 = 0.18$ ,  $P = 0.668$ ).

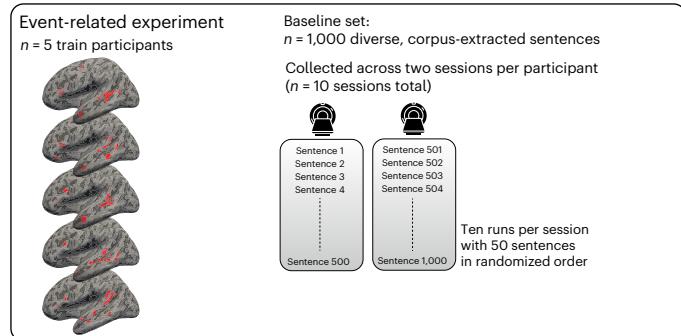
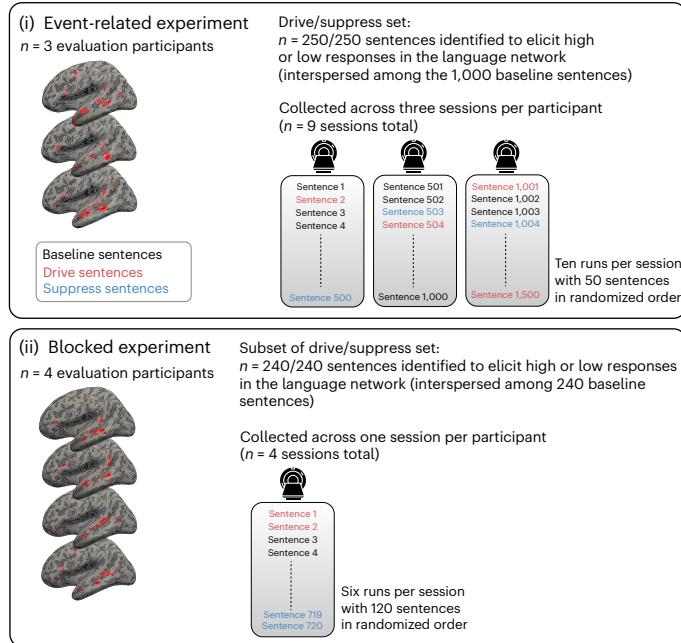
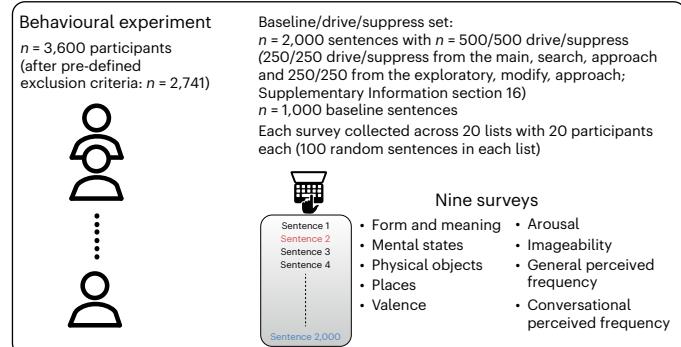
Imageability—whether sentences are easy to visualize—was strongly correlated with whether the sentence’s content concerned physical objects ( $r = 0.75$ ) and places ( $r = 0.49$ ). Imageability strongly modulated brain responses, such that sentences rated as more imageable elicited a lower response ( $r = -0.30$ , d.f. = 998,  $P < 0.001$ ,  $t = -10.04$ , s.e. = 0.03), and it explained variance beyond surprisal ( $\chi^2 = 93.03$ ,  $P < 0.001$ ).

Finally, for perceived frequency, we found that sentences that are perceived as more frequent (either in general or in conversational settings; these two norms were correlated with each other at  $r = 0.77$ ) elicited lower responses ( $r = -0.41$  and  $r = -0.33$ , d.f. = 998, both  $P < 0.001$ ,  $t = -14.14$  and  $t = -10.89$ , both s.e. = 0.03), with additional variance explained beyond surprisal (general perceived frequency:  $\chi^2 = 96.63$ ,  $P < 0.001$ ; conversational perceived frequency:  $\chi^2 = 44.46$ ,  $P < 0.001$ ).

To summarize the findings in this section, sentences that are surprising, fall in the middle of the grammaticality and plausibility range and are perceived as not very frequent elicit a stronger response in the language network. In contrast, sentences that have positive content, talk about physical objects and places, and, more generally, are easy to visualize elicit a lower response in the language network (Fig. 5). These patterns were highly similar across individual LH language fROIs and anatomically defined language ROIs but showed some differences from the RH language network, in line with some past claims (Supplementary Information section 21).

## Discussion

We provide a demonstration of non-invasive neural activity control in areas that are implicated in higher-level cognition: a brain-aligned Transformer model (GPT2-XL) can be used to drive and suppress brain responses in the language network of new individuals. We also provide

**a** Encoding model development**b** Encoding model evaluation**c** Sentence properties that modulate brain responses

a rich characterization of stimulus properties that modulate neural responses in the language network and find that less probable sentences generally elicit higher responses, with additional contributions from several form- and meaning-related features.

A number of studies have now shown that representations extracted from neural network models of language can capture neural responses to language in human brains, as recorded with fMRI or intracranial methods<sup>18–23,66–68</sup>. These studies have been conducted in an ‘open-loop’ manner: brain responses are simply acquired to a set of stimuli without any attempt to achieve specific levels of brain activity according to quantitative predictions. These stimulus sets have been limited to naturally occurring sentences, which cover a restricted portion of the space of linguistic/semantic variation. Furthermore, the

**Fig. 6 | Experimental overview.** **a**, Encoding model development. We curated a large set ( $n = 1,000$ ) of diverse, corpus-extracted sentences (the baseline set) and collected brain responses in  $n = 5$  train participants in an event-related fMRI design across two sessions per participant. **b**, Encoding model evaluation. We identified a set of sentences to activate the language network to a maximal extent (250 drive sentences) or a minimal extent (250 suppress sentences) by searching across ~1.8 million sentences (search approach). We collected responses to these 500 drive/suppress sentences randomly interspersed among the baseline sentences in  $n = 3$  new participants (evaluation participants) across three sessions per participant in the event-related design (i). Moreover, we collected responses to a large subset of the drive, suppress and baseline sentences (240 from each condition, a total of 720 sentences) in  $n = 4$  new participants in a blocked fMRI design within one session for each participant (ii). **c**, Sentence properties that modulate brain responses. To understand what sentence properties modulate brain responses in the language network, we collected ten behavioural rating norms (across nine surveys) to characterize our experimental materials ( $n = 2,000$  sentences: 1,000 baseline, 250 drive and 250 suppress sentences from the search approach, and 250 drive and 250 suppress sentences from the exploratory modify approach; see Supplementary Information section 16) across  $n = 3,600$  participants.

encoding model is typically trained and tested on data from the same participant (for example, refs. 18–23, but see ref. 68), potentially making it overly reliant on patterns of participant-specific idiosyncrasies. Prior work has thus established similarity between LLMs and humans on a narrow distribution of linguistic input and using within-participant evaluation in an open-loop fashion. In this work, we went beyond these studies by taking inspiration from closed-loop stimulus design in visual systems neuroscience<sup>24,25</sup>: we evaluated the ability of an LLM-based encoding model to modulate the strength of neural responses in new individuals via new model-selected stimuli. Unlike typical encoding or representational similarity approaches to testing neural networks as models of the brain, we here used their predictive power to generate stimuli that would maximally drive or suppress responses of the language network. We emphasize that although using LLMs to identify new stimuli requires similarity to the human brain, this similarity need not hold at the implementation level, only at the level of representations. We, and others, acknowledge that the hardware of LLMs differs in many ways from human neural circuits (but see ref. 69). These hardware differences, possibly coupled with factors such as training data and objective, could explain why LLMs sometimes diverge from human-level performance for common linguistic phenomena such as negation and quantifier use<sup>70,71</sup>. Nevertheless, in spite of these differences, LLMs and the human language system appear to arrive at a similar representational space (see ref. 72 for similar findings in vision), making LLMs currently the most predictive models of the human language network at the granularity of fMRI voxels and intracranial recordings<sup>20,22</sup>, and allowing us to modulate brain responses via targeted stimulus selection.

A priori, one might expect this model-based stimulus selection approach to not be feasible within the domain of language because of at least two reasons. First, unlike largely bottom-up brain systems such as the ventral visual stream<sup>73</sup>, the language system extracts abstract meaning representations from linguistic sequences, which makes these representations further removed from the stimulus proper and thus more divergent across individuals, especially for more abstract meanings<sup>74</sup>. Second, language processing requires attentional engagement<sup>75</sup>, and such engagement is difficult to sustain for an extended period, especially if stimuli are repeated. One recent approach to combat fatigue/boredom has been to turn to rich naturalistic stimuli, such as stories, podcasts or movies, and to collect massive amounts of data (sometimes many hours’ worth) from a small number of individuals<sup>31,57</sup>—what is often referred to as the ‘deep data’ approach<sup>76</sup>. However, such stimuli plausibly do not sample the space of linguistic and/or semantic variation well (see Supplementary Information section 10 for evidence).

and consequently do not allow for testing models on stimuli that differ substantially from those used during training. We solved these methodological challenges by collecting neural responses to each of 1,000 semantically, syntactically and stylistically diverse sentences for each participant in rapid, event-related fMRI, presented once to maximize engagement (Fig. 6). We extended existing state-of-the-art methods for single-trial modelling<sup>30</sup> and reliability estimation (for example, ref. 31) to obtain robust neural responses to each sentence. Even with robust neural data, it was unclear whether encoding model performance is contingent on features that are specific to the stimulus set and/or to the participant at hand, which would limit generalization to (1) stimuli that differ from the ones in the training set and/or (2) brain data from new individuals. By showing that model-selected stimuli successfully modulate brain responses in new individuals in ways predicted by the model, we established that LLM representations contain information that can be utilized for causal perturbation of language responses in the human brain in a general, participant-independent manner.

We identified sentences that would push neural activity towards the edges of the stimulus–response distribution (driving and suppressing) using quantitative model-based predictions. Obtaining neural responses that span a wide range of activation levels enabled us to ask which stimulus properties maximally (or minimally) engage the language network in the human brain, bringing us closer to understanding the representations and computations that support language comprehension. This general approach dates back to the pioneering work of Hubel and Wiesel<sup>77</sup> that provided an understanding of visual cortical computations by examining what stimuli cause each neuron to respond the most. Because linguistic input is extremely rich and language-responsive neuronal populations could, in principle, be tuned to many (possibly interacting) dimensions related to lexical, syntactic, semantic or other linguistic properties, including ones that were not hypothesized in advance, we here identified target drive and suppress sentences using model predictions, thus removing experimenter bias.

Of course, a predictive model can be developed using features from any quantitative representation of sentences, including hidden states from an LLM (as we do here) but also much simpler univariate measures of different linguistic properties. Following a reviewer's suggestion, we explicitly compared the predictivity performance of our encoding model, which uses GPT2-XL hidden states as features, to the performance of three encoding models that use univariate measures of surprisal (we focused on surprisal given its prominence in theorizing and empirical work on language<sup>40–47</sup>). The encoding models based on univariate surprisal estimates perform substantially lower than the encoding model based on GPT2-XL hidden states (Supplementary Information section 17). Importantly, however, our motivation for using GPT2-XL representations goes beyond predictivity performance. LLMs allow for an assumption-neutral and multi-faceted approach to stimulus identification. Because LLMs are optimized for next-word prediction, their representations contain information about linguistic regularities at all levels, from word-level properties (including both word forms and their meanings), to syntactic structure, to semantic compositional meanings<sup>78–80</sup>. This is because all of these properties can inform what word is likely to come next. By virtue of its assumption neutrality, this approach allows for bottom-up discovery. Surprisal models (for example, based on *n*-grams or structure probabilities in a PCFG parser; Supplementary Information section 20) have the advantage of being interpretable but can only be used for testing specific hypotheses. Neural network language models can also be leveraged to test specific hypotheses but additionally enable bottom-up discoveries of features that may not have been hypothesized in advance.

Indeed, we identified drive sentences that we could not have come up with in advance. These sentences were unusual on various dimensions related to their linguistic properties (Supplementary Information section 19) and highly distinct from the naturalistic baseline sentences

(Supplementary Information section 11; note that the suppress sentences were more akin to naturalistic sentences), making these sentences *a priori* unlikely to be created or selected by experimenters and unlikely to be present in naturalistic stimuli, such as stories or movies (Supplementary Information section 10). Yet these stimuli were able to drive responses in the language network.

To understand what stimulus properties modulate neural responses, we examined the effects of 11 sentence properties on the brain responses to the linguistically diverse set of 2,000 sentences. In line with much past work<sup>43–47</sup>, we found that surprisal has a strong effect on neural activity, with less probable sentences eliciting higher responses. However, a number of other properties explained variance beyond surprisal, including grammatical well-formedness and plausibility. Examining responses to a highly diverse set of sentences revealed a nonlinearity in neural response in the form of an inverted-U shape. Sentences in the mid-range of well-formedness and plausibility elicit the highest response. This response is higher than the response to sentences in the low range, similar to the previously reported findings of stronger responses to phrases and sentences than to lists of unconnected words<sup>2,61,62</sup>. The response is also higher than the response to sentences in the high range—sentences that are highly plausible and use common grammatical structures—which are easy to process<sup>41</sup>. Put differently, it appears that to elicit a strong response in the language network, a stimulus has to sufficiently resemble the kind of input we encounter in our experiences with language, given that our experiences presumably tune the language network to those kinds of stimuli. However, once some minimal level of language-likeness is reached, neural responses are modulated by processing difficulty, which depends on a combination of lexical, syntactic and semantic features. Finally, one contribution of this work relative to past brain imaging studies is that we show sensitivity to these different linguistic properties at the fine-grained level of individual sentences (unlike standard blocked or event-related designs where groups of sentences are compared). In this way, we believe that this rich dataset powerfully complements and extends prior evidence<sup>43,44,46,51,63,81</sup> and allows for the testing of new hypotheses about linguistic/semantic properties affecting neural responses.

A few limitations and future directions are worth noting. First, we here studied the language network—comprising three frontal and two temporal areas—as a whole. As discussed earlier, there are good reasons to adopt this approach: the different regions of this network have similar functional response profiles, with respect to both their selectivity for language<sup>9,26</sup> and their responses to linguistic manipulations<sup>7,11,29</sup>, and they exhibit highly correlated time courses during naturalistic cognition paradigms<sup>8,27,28,39</sup>. However, some functional heterogeneity has been argued to exist within the language network<sup>61,82</sup>. Future efforts using an approach like the one adopted here may discover functional differences within the language network (by searching for stimuli that would selectively drive particular regions in the network) as well as between the core LH language network and the RH homotopic areas and other language-responsive cortical, subcortical and cerebellar areas. Second, the current results are limited to English but can be extended to other languages given the advances in multi-lingual language models. Third, we have here relied on fMRI—a method with an inherently limited temporal resolution. Data from fMRI could be fruitfully supplemented with data from intracranial recordings, which would allow for model representations to be related to neural activity in a temporally resolved, word-by-word fashion and potentially uncover functional dissociations that are obscured when activity is averaged across adjacent words. Finally, novel ways of quantifying properties of linguistic input (for example, based on the LLM representational space<sup>83</sup>) hold great potential to further understand how certain sentences modulate responses in the mind and brain.

In conclusion, we demonstrate modulation of brain responses in the language network in new individuals in a closed-loop manner.

This work has far-reaching implications for neuroscientific research and clinical applications. In particular, an accurate model-to-brain encoding model can serve as a quantitative, assumption-neutral tool for deriving experimental materials aimed at understanding the functional organization of the language network and putatively downstream areas that support abstract knowledge and reasoning<sup>84–86</sup>. Moreover, accurate encoding models can be used as a ‘virtual language network’ to simulate experimental contrasts *in silico*<sup>87</sup>. In particular, the model-selected sentences can be queried in a high-throughput manner to analyse the response properties of the language network in detail, providing the ability to rapidly generate novel hypotheses about language processing that can then be tested in a closed-loop manner. For prospective clinical application, stimuli can be optimized for eliciting a strong response, thus allowing for efficient identification of language circuits, which may be especially important for individuals with brain disorders and other special populations, or in circumstances where time is of essence (for example, neurosurgical planning and intraoperative testing). Finally, integrating the rapid advancements of artificial neural network models with larger and/or time-resolved measures of neural activity opens the door to even more fine-grained control of areas implicated in higher-level cognition.

## Methods

All experiments were performed with ethical approval from the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology (MIT) (protocol number 2010000243). All participants gave informed written consent before starting the experiments.

We developed an encoding model to predict brain responses in the language network to arbitrary new sentences and evaluated this model by (1) identifying novel sentences that are predicted to activate the language network to a maximal (or minimal) extent and (2) collecting brain responses to these sentences in new participants. We then investigated which stimulus properties drive the responses in the language network (see Fig. 6 for an overview of the study).

## Encoding model development

**General approach and data collection.** We developed an encoding model of the LH language network in the human brain. Developing an encoding model requires brain responses to a broad range of linguistic input. We therefore curated a large set of diverse, corpus-extracted six-word sentences ( $n=1,000$ , baseline set); collected brain responses while five participants (train participants) read each sentence in an event-related, condition-rich fMRI paradigm (each sentence equals a condition) across two sessions each; and modelled those responses using a recently developed single-trial modelling framework<sup>30</sup>, which we adapted for no-repeats designs ('fMRI experiments' and Supplementary Information section 3). The baseline set consisted of two subsets: the first subset ( $n=534$  sentences) aimed to maximize semantic diversity to cover a broad range of topics, and the second subset ( $n=466$  sentences) was selected from across diverse genres and styles (newspaper text, web media, transcribed spoken language and so on) (Supplementary Information section 1). In five train participants, we recorded brain responses to the sentences in the baseline set across two scanning sessions (Fig. 6a). The participants were instructed to read attentively and think about the sentence’s meaning. To encourage engagement with the stimuli, prior to the session, the participants were informed that they would be asked to perform a short memory task after the session ('fMRI experiments'). The sentences were presented one at a time for two seconds with a four-second inter-stimulus interval (ISI). Each run contained 50 sentences (5:36 minutes), and the sentence order was randomized across participants.

The language network was defined functionally in each participant using an extensively validated localizer task<sup>24</sup> ('Definition of ROIs'). Although the network consists of five areas (two in the temporal lobe

and three in the frontal lobe), we treat it here as a functionally integrated system given the similarity among the five regions in their functional response profiles across dozens of experiments (for example, refs. 11,29,46; see Fig. 4b,c and Supplementary Information section 4 for evidence of similar preferences for the baseline set in the current data) and high inter-regional correlations during naturalistic cognition paradigms<sup>8,27,28,36,39</sup>. To mitigate the effect of collecting data across multiple scanning sessions and to equalize response units across voxels and participants, the BOLD responses were z-scored session-wise per voxel. BOLD responses from the voxels in the LH language network were averaged within each train participant ('Definition of ROIs') and averaged across participants to yield an average language network response to each of the 1,000 baseline set sentences.

**Encoding model.** To develop an encoding model of the language network, we fitted a linear model from the representations of an LLM to brain responses (an encoding approach). The brain data that were used to fit the encoding model were the averaged LH language network responses from the  $n=5$  train participants. To map from LLM representations to brain responses, we used a linear mapping model. Note that the term ‘mapping model’ refers to the regression model from LLM representations to brain activity, while the term ‘encoding model’ encompasses both the LLM used to transform a sentence into an embedding representation as well as the mapping model.

The mapping model was an L2-regularized ('ridge') regression model, which can be seen as placing a zero-mean Gaussian prior on the regression coefficients<sup>88</sup>. Introducing the L2 penalty on the weights results in a closed-form solution to the regression problem, which is similar to the ordinary least-squares regression equation:

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X} + \alpha I)^{-1} \mathbf{X}^T \mathbf{y} + w_0$$

where  $\mathbf{X}$  is a matrix of regressors ( $n$  stimuli by  $d$  regressors). The regressors are unit activations from the sentence representations derived by exposing an LLM to the same stimuli as the human participant was exposed to, and  $d$  refers to the number of units in the LLM embedding representation ('hidden size').  $\mathbf{y}$  is an  $n$ -length column vector containing the relevant brain ROI's mean response to each stimulus.  $I$  is the identity matrix ( $d$  by  $d$ ).  $\mathbf{w}$  is a  $d$ -length column vector with the weights learned for each regressor.  $w_0$  is the intercept term.

$\alpha$  is the regularization parameter that trades off between the fit to the data and the penalty for weights with high coefficients. To select this regularization parameter, we used leave-one-out cross-validation implemented using the scikit-learn Python library function RidgeCV (ref. 89; v.0.24.2). Specifically, for each of 60 logarithmically spaced  $\alpha$  regularization parameter values ( $1 \times 10^{-30}, 1 \times 10^{-29}, \dots, 1 \times 10^{28}, 1 \times 10^{29}$ ), we measured the squared error in the resulting prediction of the left-out stimulus using regression weights derived from the other stimuli in the data. We computed the average of this error (across the stimuli) for each of the 60 potential  $\alpha$  regularization parameter values. We then selected the  $\alpha$  regularization parameter that minimized this mean squared error ( $\alpha=10,000$ ). When cross-validation was performed, the  $\alpha$  regularization parameter was always selected using the stimuli in the train split, and with the  $\alpha$  parameter selected, the regression model using that parameter was used on the test split.

**Encoding model performance.** We obtained an unbiased estimate of encoding model performance using three different approaches: (1) cross-validated predictivity performance on held-out sentences (Supplementary Information section 6), (2) cross-validated predictivity performance on held-out participants within the train participants (Supplementary Information section 7) and (3) held-out prediction performance on new participants (evaluation participants) and sentences (see 'Encoding model evaluation' in Methods and 'The model captures most explainable variance in new participants' in Results).

**Sentence representations from LLMs.** To obtain sentence representations for the encoding model, we used the unidirectional-attention Transformer LLM GPT2-XL<sup>14</sup>, which was identified as the most brain-aligned language base model in prior work<sup>20</sup> and which was the largest unidirectional OpenAI GPT model available on HuggingFace<sup>90</sup> at the time of the experiments (summer 2021). (Supplementary analyses were performed using BERT-large (Supplementary Information section 6c).) We used the pretrained model available via the HuggingFace library<sup>90</sup> (transformers v.4.11.3; <https://huggingface.co/gpt2-xl>). GPT2-XL has 48 layers (that is, Transformer blocks) in addition to the embedding layer. The embedding dimension is 1,600. We obtained model representations by tokenizing each sentence using the model's standard tokenizer (GPT2TokenizerFast) and passing each sentence through the model. We retrieved model representations for each model layer (that is, at the end of each Transformer block). Given that human participants were exposed to the whole sentence at once, we similarly computed a sequence summary representation for each sentence. We obtained the representation of the last sentence token, given that unidirectional models aggregate representations of the preceding context (that is, earlier tokens in the sentence). Furthermore, to ensure that the results were robust to this choice of summary representation, we also obtained a sequence summary representation by computing the arithmetic mean of the representations associated with each token in each sentence (Supplementary Information section 6b). The resulting features were used as regressors in the LLM–brain comparisons. Each LLM layer (the model stage for which representations were extracted—that is, Transformer blocks) was treated as a separate set of regressors in the LLM–brain comparisons. Layer 22 features were selected as regressors in the encoding model on the basis of cross-validated model performance evaluation (Supplementary Information section 6a).

### Encoding model evaluation

Using our trained encoding model, we identified a set of new sentences to activate the language network to a maximal extent (drive sentences) or a minimal extent (suppress sentences). To do so, we searched across ~1.8 million sentences to identify sentences predicted to elicit high or low fMRI responses (250 sentences of each kind) (Fig. 6b and Supplementary Information section 9). We collected brain responses to these novel sentences from three new participants (across three sessions each). The drive and suppress sentences were randomly interspersed among the 1,000 baseline sentences (for a total of  $n = 1,500$  sentences), collected across three scanning sessions per participant ( $n = 9$  sessions total). (In a more exploratory component of the study, we complemented the search approach with another approach—the modify approach—where we used gradient-based modifications to transform a random sentence into a novel sentence/string predicted to elicit high or low fMRI responses. We collected brain responses to these novel sentences in two new participants (see Supplementary Information section 16 for the details of methods and the results)).

To ensure that the results were robust and generalizable to different experimental paradigms, we additionally collected fMRI responses to a large subset of the drive and suppress sentences along with the baseline sentences in a traditional blocked design with four independent participants (one scanning session each). The participants for the blocked experiment were exposed to a total of 720 unique sentences (from the baseline, drive and suppress conditions, 240 per condition, which were randomly sampled for each participant). The sentences were grouped into blocks of five sentences from the same condition and were presented on the screen one at a time for 2 s with a 400 ms ISI. Each run contained 120 sentences in 24 blocks (5:36 minutes).

### fMRI experiments

**Participants.** A total of 14 neurotypical adults (9 female), aged 21 to 31 (mean, 25.3; s.d., 3), participated for payment between October 2021 and December 2022. The sample size was based on those used for

previous fMRI semantic decoding experiments<sup>57,58</sup>. All participants had normal or corrected-to-normal vision and no history of neurological, developmental or language impairments. Twelve participants (~86%) were right-handed, as determined by self-report and the Edinburgh handedness inventory<sup>91</sup>, and two (~14%) were left-handed. All participants had a left-lateralized/bilateral language network as determined by the examination of the activation maps for the language localizer<sup>2</sup>. All participants were native speakers of English. Each scanning session lasted between one and two hours. All participants gave informed written consent in accordance with the requirements of MIT's Committee on the Use of Humans as Experimental Subjects (protocol number 2010000243). The participants were compensated for their time (US\$30 per hour). To err on the conservative side, no participants were excluded from the study on the basis of data quality considerations.

**Critical fMRI tasks.** Sentence-reading task: event-related design. We developed a paradigm to collect brain responses to as many individual sentences as possible (similar to recent paradigms in visual neuroscience—for example, the Natural Scenes Dataset<sup>31</sup>). The participants passively read each sentence once, in a condition-rich, event-related fMRI design (each sentence is effectively a condition). The sentences were presented in black font on a light-grey background one at a time for 2 s with a 4 s ISI consisting of a fixation cross. Each run contained 50 unique sentence trials and three 12 s fixation blocks (in the beginning, middle (that is, after 25 sentences) and end of each run). Each run lasted 336 s (5:36 minutes).

The participants were instructed to read attentively and think about the sentence's meaning. To encourage engagement with the stimuli, prior to the session, the participants were informed that they would be asked to perform a short memory task after the session (outside of the scanner).

The first five participants (train participants) were exposed to the set of  $n = 1,000$  baseline sentences and therefore completed 20 experimental runs (across two scanning sessions). The sentences were randomly assigned to runs for each participant (that is, the participants were exposed to different presentation orders).

The next five participants (evaluation participants) were exposed to  $n = 250$  drive and  $n = 250$  suppress sentences interspersed among the set of  $n = 1,000$  baseline sentences—a total of  $n = 1,500$  sentences—and therefore completed 30 runs of the experiment (across three scanning sessions). The  $n = 1,500$  sentences were randomly assigned to experimental runs for each participant while ensuring a balanced distribution of baseline, drive and suppress sentences in each run, leading to the following distribution of baseline/drive/suppress sentences in the three scanning sessions: 333/84/83, 333/83/84 and 334/83/83.

Sentence-reading task: blocked design. To evaluate the robustness of brain responses to the drive and suppress sentences, we additionally presented a subset of the drive, suppress and baseline sentence materials in a traditional blocked design.

The sentences were grouped into blocks of five sentences from the same condition (baseline, drive or suppress) and were presented on the screen (in black font on a light-grey background) one at a time for 2 s with a 400 ms ISI consisting of a fixation cross (for a total block duration of 12 s). Each run consisted of 24 blocks with 8 blocks (40 sentences) per condition. There were five 12 s fixation blocks: in the beginning and end of each run, as well as after 6, 12 and 18 blocks. Each run lasted 348 s (5:48 minutes).

As in the event-related experiment, the participants were instructed to read attentively and think about the sentence's meaning. Prior to the session, the participants were informed that they would be asked to perform a short memory task after the session (outside of the scanner).

The participants for the blocked experiment were exposed to a total of 720 unique sentences (from the baseline, drive and suppress conditions; 240 per condition) across six runs in a single scanning

session. These sentences were sampled randomly without replacement from the full set of materials (consisting of 250 drive, 250 suppress and 1,000 baseline stimuli from the search approach). The sentences were randomly sampled and assigned to runs for each participant (that is, the participants were exposed to different presentation orders of different subsets of the materials). The condition order was counterbalanced across runs and participants.

**Memory task for the sentence-reading task.** For both the event-related and blocked critical sentence-reading experiments, the participants completed a memory task at the end of each scanning session (outside of the scanner) to incentivize attention throughout the session. The participants were informed ahead of time that they would be asked to perform a memory task after the scanning session.

The participants were presented with a set of sentences, one at a time, and asked to decide whether they had read it during the scanning session. For both the event-related and blocked experiments, the memory task consisted of 30 sentences: 20 sentences from the set used in the scanning session and 10 foil sentences. For the event-related experiment, the 20 correct targets were randomly sampled without replacement from each of the 10 runs in that session, 2 from each run. For the blocked experiment, the 20 correct targets were randomly sampled without replacement from each of the 6 runs in that session, 3 from each run, with an additional 2 sentences from random runs.

The 10 foil sentences were randomly sampled without replacement from a set of 100 sentences. These 100 foil sentences were manually selected from the same corpora that were used to construct the baseline stimulus set (15 sentences from each of the three genres from the Toronto Book Corpus—45 in total—and 55 sentences from the additional eight corpora; Supplementary Information section 1).

The average accuracy (the sum of correct responses divided by the total number of responses; chance level is 50%) was 70.4% (s.d. across sessions, 11.4%) for the event-related participants ( $n = 24$  sessions—responses for one session were not saved due to an error in the script) and 61.7% (s.d. across sessions, 10%) for the blocked participants ( $n = 4$  sessions).

**fMRI data acquisition, preprocessing and first-level analysis.** fMRI data acquisition. Structural and functional data were collected on the whole-body 3 Tesla Siemens Prisma scanner with 32-channel head coil, at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. T1-weighted, magnetization prepared rapid gradient echo (MP-RAGE) structural images were collected in 176 sagittal slices with 1 mm isotropic voxels (repetition time (TR): 2,530 ms; echo time (TE): 3.48 ms; inversion time (TI): 1,100 ms; flip, 8 degrees). Functional BOLD data were acquired using an SMS EPI sequence (with a 90-degree flip angle and using a slice acceleration factor of 2), with the following acquisition parameters: 52 2-mm-thick near-axial slices acquired in the interleaved order (with 10% distance factor), 2 mm  $\times$  2 mm in-plane resolution, a field of view in the phase-encoding ( $A \gg P$ ) direction of 208 mm and a matrix size of 104  $\times$  104, a TR of 2,000 ms and a TE of 30 ms, and a partial Fourier of 7/8. The first 10 s of each run were excluded to allow for steady-state magnetization.

**fMRI preprocessing.** The fMRI data were preprocessed using SPM12 (release 7487), the CONN EvLab module (release 19b) and custom MATLAB scripts. Each participant's functional and structural data were converted from DICOM to NIfTI format. All functional scans were coregistered and resampled using B-spline interpolation to the first scan of the first session. Potential outlier scans were identified from the resulting subject-motion estimates as well as from BOLD signal indicators using the default thresholds in the CONN preprocessing pipeline (five standard deviations above the mean in global BOLD signal change, or framewise displacement values above 0.9 mm)<sup>92</sup>. Note that the identification of outlier scans was leveraged in the blocked first-level modelling but not in the data-driven event-related first-level

modelling. Functional and structural data were independently normalized into a common space (the MNI template; IXI549Space) using the SPM12 unified segmentation and normalization procedure<sup>93</sup> with a reference functional image computed as the mean functional data after realignment across all time points omitting outlier scans. The output data were resampled to a common bounding box between MNI-space coordinates (-90, -126, -72) and (90, 90, 108), using 2 mm isotropic voxels and fourth-order spline interpolation for the functional data, and 1 mm isotropic voxels and trilinear interpolation for the structural data. Last, the functional data were smoothed spatially using spatial convolution with a 4 mm full width at half maximum (FWHM) Gaussian kernel.

**First-level modelling of event-related experiments.** The critical, event-related experiment was analysed using GLMsingle<sup>30</sup>, a framework for obtaining accurate response estimates in quick event-related single-trial fMRI designs. Modelling such responses is challenging due to temporal signal autocorrelation, participant head motion and scanner instabilities. The GLMsingle framework introduces three main steps to combat noise in a data-driven manner. The first is the choice of haemodynamic response function (HRF) to convolve with the design matrix: an HRF is identified from a library of 20 candidate functions (derived from independent fMRI data<sup>31</sup>) as the best fitting for each voxel separately. The second step is noise regressors: a set of voxels that are unrelated to the experimental paradigm are identified, and these voxels' time courses are used to derive an optimal set of noise regressors by performing principal component analysis. The third step is the regularization of voxel responses: instead of an ordinary least-squares regression, GLMsingle uses fractional ridge regression<sup>94</sup> to model voxel responses to dampen the noise inflation in a standard ordinary least-squares regression due to correlated predictors from rapid, successive trials.

Using this framework, a general linear model (GLM) was used to estimate the beta weights that represent the BOLD response amplitude evoked by each individual sentence trial (fixation was modelled implicitly, such that all time points that did not correspond to one of the conditions (sentences) were assumed to correspond to a fixation period). The data from different scanning sessions for a given participant were analysed together. The 'sessionIndicator' option in GLMsingle was used to specify how different input runs were grouped into sessions. For each voxel, the HRF that provided the best fit to the data was identified (on the basis of the amount of variance explained). The data were modelled using a fixed number of noise regressors (five) and a fixed ridge regression fraction (0.05) (these parameters were determined empirically using an extensive joint data modelling and data evaluation framework; Supplementary Information section 3).

By default, GLMsingle returns beta weights in units of percent signal change by dividing by the mean signal intensity observed at each voxel and multiplying by 100. Hence, the beta weight for each voxel can be interpreted as a change in BOLD signal for a given sentence trial relative to the fixation baseline. To mitigate the effect of collecting data across multiple scanning sessions, the beta values were z-scored session-wise per voxel ('Definition of ROIs').

**First-level modelling of blocked experiments.** The blocked experiments were analysed using standard analysis procedures using SPM12 (release 7487) and the CONN EvLab module (release 19b). Effects were estimated using a GLM in which the beta weight associated with each experimental condition was modelled with a boxcar function convolved with the canonical HRF (fixation was modelled implicitly, such that all time points that did not correspond to one of the conditions were assumed to correspond to a fixation period). Temporal autocorrelations in the BOLD signal time series were accounted for by a combination of high-pass filtering with a 128-second cutoff and whitening using an AR(0.2) model (a first-order autoregressive model linearized

around the coefficient  $a = 0.2$ ) to approximate the observed covariance of the functional data in the context of restricted maximum likelihood estimation. In addition to experimental condition effects, the GLM design included first-order temporal derivatives for each condition (to model variability in the HRF delays), as well as nuisance regressors to control for the effects of slow linear drifts, subject-motion parameters and potential outlier scans on the BOLD signal.

### Definition of ROIs

**Language ROIs.** Language ROIs were defined in individual participants using functional localization<sup>2,38</sup>. This approach is crucial because many functional regions do not exhibit a consistent mapping onto macro-anatomical landmarks<sup>95</sup>, and this variability is problematic when functionally distinct regions lie close to each other, as is the case with both frontal and temporal language areas (see ref. 9 for a discussion of this issue for ‘Broca’s area’).

For each participant, fROIs were defined by combining two sources of information<sup>2</sup>: (1) the participant’s activation map for the localizer contrast of interest (*t*-map) and (2) group-level constraints (‘parcels’) that delineated the expected gross locations of activations for the relevant contrast and were sufficiently large to encompass the extent of variability in the locations of individual activations (all parcels are available for download from <https://evlab.mit.edu/funcloc/download-parcels>).

**Language network localizer task.** The task used to localize the language network was a reading task contrasting sentences (for example, ‘THE SPEECH THAT THE POLITICIAN PREPARED WAS TOO LONG FOR THE MEETING’) and lists of unconnected, pronounceable non-words (for example, ‘LAS TUPING CUSARISTS FICK PRELL PRONT CRE POME VILLPA OLP WORNETIST CHO’) in a standard blocked design with a counterbalanced condition order across runs (introduced in Fedorenko et al.<sup>2</sup>). The sentences > non-words contrast targets higher-level aspects of language, including lexical and phrasal semantics, morphosyntax, and sentence-level pragmatic processing, to the exclusion of perceptual (speech- or reading-related) processes. The areas identified by this contrast are strongly selective for language relative to diverse non-linguistic tasks (for example, ref. 26; see Fedorenko and Blank<sup>9</sup> for a review). This paradigm has been extensively validated and shown to be robust to variation in the materials, modality of presentation, language and task (for example, refs. 2,8). Furthermore, a network that corresponds closely to the localizer contrast emerges robustly from whole-brain task-free data—voxel fluctuations during rest<sup>39</sup>.

Each stimulus consisted of 12 words/non-words. The stimuli were presented in the centre of the screen, one word/non-word at a time, at the rate of 450 ms per word/non-word. Each stimulus was preceded by a 100 ms blank screen and followed by a 400 ms screen showing a picture of a finger pressing a button, and then a blank screen for another 100 ms, for a total trial duration of 6 s. Experimental blocks lasted 18 s (with three trials per block), and fixation blocks lasted 14 s. Each run (consisting of 5 fixation blocks and 16 experimental blocks) lasted 358 s. The participants completed two runs. The participants were instructed to read attentively (silently) and press a button on the button box whenever they saw the picture of a finger pressing a button on the screen. The button-pressing task was included to help the participants remain alert.

The materials and scripts are available from the Fedorenko Lab website (<https://evlab.mit.edu/funcloc>).

**Language network fROIs.** The language fROIs were defined using the sentences > non-words contrast from the language localizer collected in each participant’s first scanning session (see, for example, Mahowald and Fedorenko<sup>96</sup>, for evidence that localizer maps are highly stable within individuals over time, including across sessions). This contrast targets higher-level aspects of language, to the exclusion of perceptual

(speech/reading) and motor–articulatory processes (for a discussion, see Fedorenko and Thompson-Schill<sup>3</sup>).

To define the language fROIs, each individual sentences > non-words *t*-map was intersected with a set of ten binary parcels (five in each hemisphere). These parcels were derived from a probabilistic activation overlap map using watershed parcellation, as described by Fedorenko et al.<sup>2</sup>, for the sentences > non-words contrast in 220 independent participants and covered extensive portions of the lateral frontal, temporal and parietal cortices. Specifically, five language fROIs were defined in the dominant hemisphere: three on the lateral surface of the frontal cortex (in the inferior frontal gyrus, IFG, and its orbital part, IFCorb, as well as in the middle frontal gyrus, MFG) and two on the lateral surface of the temporal and parietal cortex (in the anterior temporal cortex, AntTemp, and posterior temporal cortex, PostTemp). Following prior work (for example, ref. 27), to define the RH fROIs, the LH language parcels were transposed onto the RH, allowing the LH and RH homotopic fROIs to differ in their precise locations within the parcels.

Within each of these ten parcels, the 10% of voxels with the highest *t* values for the sentences > non-words contrast were selected (see Supplementary Information section 15e for the number of voxels in each fROI).

**Control ROIs.** In addition to language regions, we examined two large-scale brain networks linked to high-level cognitive processing—the MD network<sup>32</sup> and the DMN<sup>33</sup>—which, similar to the language regions, were functionally defined using independent localizer tasks in each participant. We also examined a set of anatomical parcels<sup>34</sup> in an effort to cover the entire cortex (see Supplementary Information section 15 for the details).

**Aggregation of voxels within each ROI.** The voxels belonging to each fROI (language, MD and DMN) and each anatomical Glasser ROI were aggregated by averaging. For the fMRI data reported in the main text, each voxel was *z*-scored session-wise prior to averaging, to minimize potential non-stationarities that exist across different scanning sessions and to equalize response units across voxels. In Supplementary Information sections 12 and 14, we report fMRI data without any normalization (the key patterns of the results are not affected).

On average, we extracted responses from 10 language fROIs (s.d. = 0), 19.43 MD fROIs (s.d. = 1.28), 12 DMN fROIs (s.d. = 0) and 353.71 anatomical Glasser parcels (s.d. = 10.34) across  $n = 14$  participants (5 train participants, 5 evaluation participants in the event-related fMRI design from the search and modify approaches, and 4 evaluation participants in the blocked fMRI design). In a few cases, (f)ROIs could not be extracted due to a negative *t* statistic for the contrast of interest or lack of coverage in our functional acquisition sequence.

### Sentence properties that modulate brain responses

**General approach.** To shed light on what property or properties make some sentences elicit stronger responses in the language network, we collected an extensive set of norms to characterize the full set of sentences in this study ( $n = 2,000$ : 1,000 baseline sentences, 250 drive and 250 suppress sentences from the search approach, and 250 drive and 250 suppress sentences from the exploratory modify approach) (Fig. 6c) and examined the relationship between these properties and fMRI responses. First, building on the body of evidence for surprisal modulating language processing difficulty (in both behavioural psycholinguistic work<sup>40–42</sup> and brain imaging investigations<sup>43–47</sup>), we computed the average log probability for each sentence using GPT2-XL (surprisal is the negative log probability; Surprisal features). Second, we collected ten behavioural rating norms across a total of  $n = 3,600$  participants (on average, 15.23 participants per sentence per rating norm; minimum, 10; maximum, 19). The norms spanned five broad categories and were all motivated by prior work in linguistics and psycholinguistics (‘Behavioural norms’).

**Surprisal features.** We estimated the log probability of a word given its context for the words in each sentence. The negative log probability of a word/sentence is known as ‘surprisal’<sup>97</sup>. The log probability of each sentence was computed using the pre-trained unidirectional-attention language model GPT2-XL<sup>14</sup> from the HuggingFace library<sup>90</sup> (transformers v.4.11.3). GPT2-XL was trained on 40 GB of web text from various domains (WebText dataset). Each sentence was tokenized using the model’s standard tokenizer (GPT2Tokenizer) and the special token, [EOS], was prepended to each sentence. Punctuation was retained. We obtained the sentence-level surprisal by taking the mean of the token-level surprisals.

For supplementary analyses, we obtained surprisal estimates from an *n*-gram model and a probabilistic context-free grammar model in addition to GPT2-XL (Supplementary Information section 20).

**Behavioural norms.** *Participants.* Participants were recruited using crowd-sourcing platforms: Prolific (*n* = 8 surveys) and Amazon Mechanical Turk (*n* = 1 survey). For Prolific, the study was restricted to workers with English as their first language and their most fluent language, USA as their location and a submission approval rate greater than or equal to 90%. For Amazon Mechanical Turk, the study was restricted to ‘Mechanical Turk Masters’ workers. A total of 3,600 participants took part in the experiment across the nine surveys (400 participants for each survey; see Supplementary Information section 22c for the details). After we applied pre-defined exclusion criteria, 2,741 participants remained (Supplementary Information section 22a). The experiments were conducted with approval from and in accordance with MIT’s Committee on the Use of Humans as Experimental Subjects (protocol number 2010000243). The participants gave informed consent before starting each experiment and were compensated for their time (minimum US\$12 per hour).

**Materials, design and procedure.** The *n* = 2,000 sentences were randomly assigned to 20 unique sets containing 100 sentences each. For each survey, the participants first provided informed consent. They then answered several demographic questions (whether English is their first language, which country they are from and what age bracket they fall into); they were explicitly told that payment was not contingent on their answers to these questions. Finally, they were presented with the survey-specific instructions and the following warning: ‘There are some sentences for which we expect everyone to answer in a particular way. If you do not speak English or do not understand the instructions, please do not do this hit—you will not get paid.’

One survey targeted two core aspects of sentences: grammatical well-formedness (how much does the sentence obey the rules of English grammar?) (for details of the instructions, see Supplementary Information section 22c) and plausibility (how much sense does the sentence make?). Three surveys probed different aspects of the sentence’s content: how much does the sentence make you think about (1) others’ mental states, (2) physical objects and their interactions, and (3) places and environments? The latter two have to do with the physical world, and the former, with internal representations; the physical-versus-social distinction is one plausible organizing dimension of meaning<sup>59,60</sup>. Two surveys probed emotional dimensions of the sentences: valence (how positive is the sentence’s content?) and arousal (how exciting is the sentence’s content?). One survey targeted visual imagery (how visualizable is the sentence’s content?). Finally, the last two surveys probed people’s perception of how common the sentence is, in general versus in conversational contexts. The first survey (with two questions per sentence) took 25.01 minutes on average; the remaining surveys took 14.25 minutes on average. After the participants answered the rating question(s) for the 100 sentences (the order was randomized separately for each participant), they were asked to complete six sentence preambles (for example, ‘When I was younger, I would often...’; see Supplementary Information section

22b for the full set), which were used post hoc to evaluate English proficiency. See Supplementary Information section 22 for the details on the experimental procedures.

## Statistical analyses

LME models (implemented using the lmer function from the R package lme4 (ref. 98) v.1.1–31) were used to evaluate the statistical significance of the differences in the BOLD response among the sentence conditions (baseline, drive and suppress) and of the effect of sentence properties on the BOLD response. The critical variable of interest (either condition or sentence property) was modelled as a fixed effect. As additional effects, we modelled other variables that could modulate the BOLD response but that were not our critical variables of interest, including item (sentence), run order within a session (1–10) and sentence order within a run (1–50):

$$\text{BOLD response} \sim \text{variable of interest} + (1|\text{sentence}) + \\ \text{run within session} + \text{trial within run}.$$

(Note that because the BOLD responses were z-scored session-wise, there was no additional variance to explain by including session number or participant as a model term.)

The models were fitted using maximum likelihood estimation and used the Satterthwaite method for estimating degrees of freedom. For each LME model reported, we provide (in Supplementary Information sections 18 and 23) a table with model formulae, effect size estimates, standard error estimates, *t* statistics, *P* values, degrees of freedom and *R*<sup>2</sup> values. We evaluated the statistical significance of differences between pairs of conditions using estimated marginal means (implemented using the emmeans function from the R package emmeans<sup>99</sup> v.1.8.4-1) using Tukey’s multiple comparison method. Finally, we evaluated the statistical significance of differences between pairs of LME models using likelihood ratio tests using the chi-squared value,  $\chi^2$ , as the test statistic (implemented using the anova function from the R package lme4).

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The data are publicly available and can be downloaded via the following repository: [https://github.com/gretatuckute/drive\\_suppress\\_brains](https://github.com/gretatuckute/drive_suppress_brains).

## Code availability

The code is publicly available in the following repository: [https://github.com/gretatuckute/drive\\_suppress\\_brains](https://github.com/gretatuckute/drive_suppress_brains).

## References

1. Binder, J. R. et al. Human brain language areas identified by functional magnetic resonance imaging. *J. Neurosci.* **17**, 353–362 (1997).
2. 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).
3. Fedorenko, E. & Thompson-Schill, S. L. Reworking the language network. *Trends Cogn. Sci.* **18**, 120–126 (2014).
4. Lipkin, B. et al. Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. *Sci. Data* **9**, 529 (2022).
5. MacSweeney, M. et al. Neural systems underlying British Sign Language and audio-visual English processing in native users. *Brain J. Neurol.* **125**, 1583–1593 (2002).
6. Deniz, F., Nunez-Elizalde, A. O., Huth, A. G. & Gallant, J. L. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *J. Neurosci.* **39**, 7722–7736 (2019).

7. Hu, J. et al. Precision fMRI reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cereb. Cortex* **33**, 4384–4404 (2022).
8. Malik-Moraleda, S. et al. An investigation across 45 languages and 12 language families reveals a universal language network. *Nat. Neurosci.* **25**, 1014–1019 (2022).
9. Fedorenko, E. & Blank, I. A. Broca's area is not a natural kind. *Trends Cogn. Sci.* **24**, 270–284 (2020).
10. Bautista, A. & Wilson, S. M. Neural responses to grammatically and lexically degraded speech. *Lang. Cogn. Neurosci.* **31**, 567–574 (2016).
11. Fedorenko, E., Blank, I. A., Siegelman, M. & Mineroff, Z. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition* **203**, 104348 (2020).
12. Mesulam, M.-M. Primary progressive aphasia. *Ann. Neurol.* **49**, 425–432 (2001).
13. Wilson, S. M. et al. Language mapping in aphasia. *J. Speech Lang. Hear. Res.* **62**, 3937–3946 (2019).
14. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. *Improving Language Understanding by Generative Pre-training* Technical Report (OpenAI, 2018).
15. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT 2019* (eds Burstein, J. et al.) 4171–4186 (Association for Computational Linguistics, 2019); <https://doi.org/10.18653/v1/N19-1423>
16. Wilcox, E. G., Gauthier, J., Hu, J., Qian, P. & Levy, R. On the predictive power of neural language models for human real-time comprehension behavior. In *Proc. 42nd Annual Meeting of the Cognitive Science Society* (eds Denison, S. et al.) 1707–1713 (Cognitive Science Society, 2020).
17. Shain, C., Meister, C., Pimentel, T., Cotterell, R. & Levy, R. P. Large-scale evidence for logarithmic effects of word predictability on reading time. Preprint at *PsyArXiv* <https://doi.org/10.31234/osf.io/4hyna> (2022).
18. Toneva, M. & Wehbe, L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems 32* (NeurIPS 2019) (eds Wallach, H. et al.) 14954–14964 (Curran Associates, Inc., 2019).
19. Gauthier, J. & Levy, R. Linking artificial and human neural representations of language. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (eds Inui, K. et al.) 529–539 (Association for Computational Linguistics, 2019); <https://doi.org/10.18653/v1/D19-1050>
20. Schrimpf, M. et al. The neural architecture of language: integrative modeling converges on predictive processing. *Proc. Natl Acad. Sci. USA* **118**, e2105646118 (2021).
21. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5**, 134 (2022).
22. Goldstein, A. et al. Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).
23. 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).
24. Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).
25. Ponce, C. R. et al. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell* **177**, 999–1009 (2019).
26. Fedorenko, E., Behr, M. K. & Kanwisher, N. Functional specificity for high-level linguistic processing in the human brain. *Proc. Natl Acad. Sci. USA* **108**, 16428–16433 (2011).
27. Blank, I., Kanwisher, N. & Fedorenko, E. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *J. Neurophysiol.* **112**, 1105–1118 (2014).
28. Paukov, A. M., Blank, I. A. & Fedorenko, E. Functionally distinct language and Theory of Mind networks are synchronized at rest and during language comprehension. *J. Neurophysiol.* **121**, 1244–1265 (2019).
29. Blank, I. A. & Fedorenko, E. No evidence for differences among language regions in their temporal receptive windows. *NeuroImage* **219**, 116925 (2020).
30. Prince, J. S. et al. Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife* **11**, e77599 (2022).
31. Allen, E. J. et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* **25**, 116–126 (2022).
32. Duncan, J. The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends Cogn. Sci.* **14**, 172–179 (2010).
33. Buckner, R. L., Andrews-Hanna, J. R. & Schacter, D. L. The brain's default network: anatomy, function, and relevance to disease. *Ann. N. Y. Acad. Sci.* **1124**, 1–38 (2008).
34. Glasser, M. F. et al. A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
35. Lerner, Y., Honey, C. J., Silbert, L. J. & Hasson, U. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* **31**, 2906–2915 (2011).
36. Honey, C. J., Thompson, C. R., Lerner, Y. & Hasson, U. Not lost in translation: neural responses shared across languages. *J. Neurosci.* **32**, 15277–15283 (2012).
37. Blank, I. A. & Fedorenko, E. Domain-general brain regions do not track linguistic input as closely as language-selective regions. *J. Neurosci.* **37**, 9999–10011 (2017).
38. Nieto-Castañón, A. & Fedorenko, E. Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *NeuroImage* **63**, 1646–1669 (2012).
39. Braga, R. M., DiNicola, L. M., Becker, H. C. & Buckner, R. L. Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. *J. Neurophysiol.* **124**, 1415–1448 (2020).
40. Demberg, V. & Keller, F. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* **109**, 193–210 (2008).
41. Smith, N. J. & Levy, R. The effect of word predictability on reading time is logarithmic. *Cognition* **128**, 302–319 (2013).
42. Brothers, T. & Kuperberg, G. R. Word predictability effects are linear, not logarithmic: implications for probabilistic models of sentence comprehension. *J. Mem. Lang.* **116**, 104174 (2021).
43. Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P. & van den Bosch, A. Prediction during natural language comprehension. *Cereb. Cortex* **26**, 2506–2516 (2016).
44. 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).
45. Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P. & de Lange, F. P. A hierarchy of linguistic predictions during natural language comprehension. *Proc. Natl Acad. Sci. USA* **119**, e2201968119 (2022).
46. Shain, C., Blank, I. A., van Schijndel, M., Schuler, W. & Fedorenko, E. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia* **138**, 107307 (2020).

47. Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K. & Coulson, S. Strong prediction: language model surprisal explains multiple N400 effects. *Neurobiol. Lang.* [https://doi.org/10.1162/nola\\_00105](https://doi.org/10.1162/nola_00105) (2023).
48. Rayner, K. & Duffy, S. A. Lexical complexity and fixation times in reading: effects of word frequency, verb complexity, and lexical ambiguity. *Mem. Cogn.* **14**, 191–201 (1986).
49. Brysbaert, M., Warriner, A. B. & Kuperman, V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res. Methods* **46**, 904–911 (2014).
50. Arfè, B., Delatorre, P. & Mason, L. Effects of negative emotional valence on readers' text processing and memory for text: an eye-tracking study. *Read. Writ.* **36**, 1743–1768 (2022).
51. Kuchinke, L. et al. Incidental effects of emotional valence in single word processing: an fMRI study. *NeuroImage* **28**, 1022–1032 (2005).
52. Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T. & Medler, D. A. Distinct brain systems for processing concrete and abstract concepts. *J. Cogn. Neurosci.* **17**, 905–917 (2005).
53. Ferstl, E. C. & von Cramon, D. Y. Time, space and emotion: fMRI reveals content-specific activation during text comprehension. *Neurosci. Lett.* **427**, 159–164 (2007).
54. Lau, J. H., Clark, A. & Lappin, S. Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cogn. Sci.* **41**, 1202–1241 (2017).
55. Hu, J., Gauthier, J., Qian, P., Wilcox, E. & Levy, R. P. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (eds Jurafsky, D. et al.) 1725–1744 (Association for Computational Linguistics, 2020).
56. Kauf, C. et al. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cogn. Sci.* **47**, e13386 (2023).
57. 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).
58. Anderson, A. J. et al. Multiple regions of a cortical network commonly encode the meaning of words in multiple grammatical positions of read sentences. *Cereb. Cortex* **29**, 2396–2411 (2019).
59. Baron-Cohen, S., Wheelwright, S., Spong, A., Scahill, V. & Lawson, J. Are intuitive physics and intuitive psychology independent? A test with children with Asperger syndrome. *J. Dev. Learn. Disord.* **5**, 47–78 (2001).
60. Jack, A. I. et al. fMRI reveals reciprocal inhibition between social and physical cognitive domains. *NeuroImage* **66**, 385–401 (2013).
61. Pallier, C. & Devauchelle, A.-D. Cortical representation of the constituent structure of sentences. *Proc. Natl Acad. Sci. USA* **108**, 2522–2527 (2011).
62. Diachek, E., Blank, I., Siegelman, M., Affourtit, J. & Fedorenko, E. 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).
63. Wehbe, L. et al. Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. *Cereb. Cortex* **31**, 4006–4023 (2021).
64. Mellem, M. S., Jasmin, K. M., Peng, C. & Martin, A. Sentence processing in anterior superior temporal cortex shows a social-emotional bias. *Neuropsychologia* **89**, 217–224 (2016).
65. Redcay, E., Veloskey, K. R. & Rowe, M. L. Perceived communicative intent in gesture and language modulates the superior temporal sulcus. *Hum. Brain Mapp.* **37**, 3444–3461 (2016).
66. Wehbe, L. et al. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS ONE* **9**, e112575 (2014).
67. Jain, S. & Huth, A. G. Incorporating context into language encoding models for fMRI. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)* (eds Bengio, S., et al.) 6628–6637 (Curran Associates, Inc., 2018).
68. Toneva, M., Mitchell, T. M. & Wehbe, L. Combining computational controls with natural text reveals aspects of meaning composition. *Nat. Comput. Sci.* **2**, 745–757 (2022).
69. Kozachkov, L., Kastanenka, K. V. & Krotov, D. Building transformers from neurons and astrocytes. *Proc. Natl Acad. Sci. USA* **120**, e2219150120 (2023).
70. Jang, J., Ye, S. & Seo, M. Can large language models truly understand prompts? A case study with negated prompts. In *Proc. 1st Transfer Learning for Natural Language Processing Workshop* (eds Albalak A. et al.) 52–62 (PMLR, 2023).
71. Michaelov, J. A. & Bergen, B. K. Rarely a problem? Language models exhibit inverse scaling in their predictions following few-type quantifiers. In *Findings of the Association for Computational Linguistics: ACL 2023* (eds Rogers, A. et al.) 14162–14174 (Association for Computational Linguistics, 2023).
72. Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A. & Konkle, T. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? Preprint at bioRxiv <https://doi.org/10.1101/2022.03.28.485868> (2023).
73. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
74. Wang, X. & Bi, Y. Idiosyncratic Tower of Babel: individual differences in word-meaning representation increase as word abstractness increases. *Psychol. Sci.* **32**, 1617–1635 (2021).
75. Cohen, L., Salondy, P., Pallier, C. & Dehaene, S. How does inattention affect written and spoken language processing? *Cortex* **138**, 212–227 (2021).
76. Gratton, C. & Braga, R. M. Editorial overview: deep imaging of the individual brain: past, practice, and promise. *Curr. Opin. Behav. Sci.* **40**, iii–vi (2021).
77. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154 (1962).
78. Tenney, I., Das, D. & Pavlick, E. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (eds Korhonen, A. et al.) 4593–4601 (Association for Computational Linguistics, 2019).
79. Li, B. Z., Nye, M. & Andreas, J. Implicit representations of meaning in neural language models. In *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Vol. 1: Long Papers)* (eds Zong, C. et al.) 1813–1827 (Association for Computational Linguistics, 2021); <https://doi.org/10.18653/v1/2021.acl-long.143>
80. Unger, L. & Fisher, A. V. The emergence of richly organized semantic knowledge from simple statistics: a synthetic review. *Dev. Rev.* **60**, 100949 (2021).
81. Keller, T. A., Carpenter, P. A. & Just, M. A. The neural bases of sentence comprehension: a fMRI examination of syntactic and lexical processing. *Cereb. Cortex* **11**, 223–237 (2001).
82. Regev, T. I. et al. Neural populations in the language network differ in the size of their temporal receptive windows. Preprint at bioRxiv <https://doi.org/10.1101/2022.12.30.522216> (2023).
83. Kim, B. et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning (ICML 2018)* (eds Dy, J. & Krause, A.) 2673–2682 (Proceedings of Machine Learning Research, 2018).
84. Saxe, R. & Kanwisher, N. People thinking about thinking people: the role of the temporo-parietal junction in 'theory of mind'. *NeuroImage* **19**, 1835–1842 (2003).

85. Baldassano, C., Hasson, U. & Norman, K. A. Representation of real-world event schemas during narrative perception. *J. Neurosci.* **38**, 9689–9699 (2018).
86. Deen, B. & Freiwald, W. A. Parallel systems for social and spatial reasoning within the cortical apex. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.09.23.461550> (2022).
87. Jain, S., Vo, V. A., Wehbe, L. & Huth, A. G. Computational language modeling and the promise of *in silico* experimentation. *Neurobiol. Lang.* [https://doi.org/10.1162/nol\\_a\\_00101](https://doi.org/10.1162/nol_a_00101) (2023).
88. Hoerl, A. E. & Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
89. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
90. Wolf, T. et al. Transformers: state-of-the-art natural language processing. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds Liu, Q. & Schlangen, D.) 38–45 (Association for Computational Linguistics, 2020); <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
91. Oldfield, R. C. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* **9**, 97–113 (1971).
92. Nieto-Castanon, A. *Handbook of Functional Connectivity Magnetic Resonance Imaging Methods in CONN* (Hilbert, 2020); <https://doi.org/10.56441/hilbertpress.2207.6598>
93. Ashburner, J. & Friston, K. J. Unified segmentation. *NeuroImage* **26**, 839–851 (2005).
94. Rokem, A. & Kay, K. Fractional ridge regression: a fast, interpretable reparameterization of ridge regression. *GigaScience* **9**, giaa133 (2020).
95. Vázquez-Rodríguez, B. et al. Gradients of structure–function tethering across neocortex. *Proc. Natl Acad. Sci. USA* **116**, 21219–21227 (2019).
96. Mahowald, K. & Fedorenko, E. Reliable individual-level neural markers of high-level language processing: a necessary precursor for relating neural variability to behavioral and genetic variability. *NeuroImage* **139**, 74–93 (2016).
97. Hale, J. A probabilistic Earley parser as a psycholinguistic model. In *2nd Meeting of the North American Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, 2001).
98. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
99. Lenth, R. V. emmeans: Estimated marginal means, aka least-squares means. R package version 1.8.4-1 (2023).
100. Friston, K., Ashburner, J., Kiebel, S., Nichols, T. & Penny, W. *Statistical Parametric Mapping: The Analysis of Functional Brain Images* (Elsevier, 2006).
101. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage* **9**, 179–194 (1999).

## Acknowledgements

This work was supported by the Amazon Fellowship from the Science Hub (administered by the MIT Schwarzman College of Computing) (G.T.); the International Doctoral Fellowship from the American

Association of University Women (G.T.); the K. Lisa Yang ICoN Center Graduate Fellowship (G.T.); the MIT-IBM Watson AI Lab (S.S.); NIH award nos R01-DC016607 (E.F.), R01-DC016950 (E.F.) and U01-NS121471 (E.F.); and funds from the McGovern Institute for Brain Research (E.F.), the Simons Center for the Social Brain (E.F.) and the Brain and Cognitive Sciences Department (E.F.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank C. Casto, E. Lee and E. Gibson for their help on the project; C. Shain for comments on an earlier draft of the manuscript; and N. Kanwisher, N. A. R. Murty, N. Zaslavsky, K. Kar, J. Prince, J. McDermott, B. Lipkin, A. Ivanova and U.-M. O'Reilly for valuable discussions. We would also like to acknowledge the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, and its support team (Steve Shannon and Atsushi Takahashi).

## Author contributions

Conceptualization: G.T., M.S. and E.F. Methodology: G.T., M.S., K.K. and E.F. Software: G.T., A.S. and S.S. Validation: G.T., A.S., S.S. and M.W. Formal analysis: G.T., A.S., S.S. and M.W. Investigation (data collection): G.T., A.S. and M.T. Data curation: G.T. and M.T. Writing—original draft: G.T. and E.F. Writing—review and editing: G.T., A.S., S.S., M.T., M.W., M.S., K.K. and E.F. Visualization: G.T. Supervision: E.F. and K.K. Project administration: E.F. Funding acquisition: E.F.

## 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/s41562-023-01783-7>.

**Correspondence and requests for materials** should be addressed to Greta Tuckute or Evelina Fedorenko.

**Peer review information** *Nature Human Behaviour* thanks Surampudi Bapiraju and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2024

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give P values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

|                 |   |
|-----------------|---|
| Data collection | For fMRI data collection, the materials and scripts to localize the language and multiple demand system are available from the Fedorenko Lab website ( <a href="https://evlab.mit.edu/funcloc">https://evlab.mit.edu/funcloc</a> ). The software used to collect fMRI data was MATLAB 2014b.  |
| Data analysis   | <p>fMRI preprocessing was performed using SPM12 (release 7487) and custom MATLAB scripts, some of which build on the CONN toolbox (CONN EvLab module, release 19b).</p> <p>Event-related fMRI first-level modeling was performed using GLMsingle (Prince et al., 2022) in a Python environment (version 3.8.12). The code is available at the following repository: <a href="https://github.com/gretatuckute/GLMsingle">https://github.com/gretatuckute/GLMsingle</a></p> <p>Blocked fMRI first-level modeling was performed using SPM12 (release 7487) and custom MATLAB scripts, some of which build on the CONN toolbox (CONN EvLab module, release 19b).</p> <p>For further data analysis, code was written in Python (version 3.8.11), making heavy use of the numpy (Harris et al., 2020; version 1.21.2), scipy (Virtanen et al., 2020; version 1.7.3), scikit-learn (Pedregosa et al., 2011; version 0.24.2), pandas (McKinney et al., 2010; version 1.4.2) and transformers (Wolf et al., 2020; version 4.11.3) libraries.</p> <p>For statistical analyses, linear mixed effects models were implemented using the lmer function from the lme4 package (Bates et al., 2015; version 1.1-31) in R (version 4.2.2).</p> <p>Analysis code is available at the following repository: <a href="https://github.com/gretatuckute/drive_suppress_brains">https://github.com/gretatuckute/drive_suppress_brains</a></p> |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Data can be downloaded via the following repository: [https://github.com/gretatuckute/drive\\_suppress\\_brains](https://github.com/gretatuckute/drive_suppress_brains)

## Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender Information about participants' sex was collected and reported. Gender information was not collected.

Reporting on race, ethnicity, or other socially relevant groupings Information about race, ethnicity, or other socially relevant groupings is not reported.

Population characteristics A total of 14 neurotypical adults (9 female), aged 21 to 31 (mean 25.3; std 3). 12 participants (~86%) were right-handed, as determined by the Edinburgh handedness inventory and self-report 2 (~14%) were left-handed. All participants were native speakers of English.

Recruitment Participants were recruited from MIT and the surrounding Cambridge/Boston, MA community.

Ethics oversight The protocol for these studies was approved by MIT's Committee on the Use of Humans as Experimental Subjects (COUHES). All participants gave written informed consent in accordance with the requirements of this protocol.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description The study contains fMRI neuroimaging data and behavioral data. Data were quantitative.

Research sample Participants were from MIT and the surrounding Cambridge/Boston, MA community. Participants were native English speakers.

Sampling strategy The sample size was based on those used for previous fMRI semantic decoding experiments (Huth et al., 2016; Pereira et al., 2018).

Data collection Neuroimaging data were collected at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT. Behavioral data were collected using the Prolific and Amazon Mechanical Turk crowd-sourcing platforms.

Timing Neuroimaging data were collected between October 2021 and December 2022. Behavioral data were collected between February 2023 and April 2023.

Data exclusions For neuroimaging data, no participants were excluded from the study based on data quality considerations. For behavioral data, participants were excluded based on the following pre-defined criteria:  
 1. Native speaker status: Participants were excluded based on their native speaker status self report as well as the Prolific/mTurk language and location filters.  
 2. Sentence completions: Participants were excluded if their sentence completions were ungrammatical, contained spelling errors (that were not obvious typos) or if the completions were deeply nonsensical.  
 3. Average response time: Participants were excluded if the average response time per question was less than 3 seconds (i.e., for the survey that contained two questions, the threshold was 6 seconds).  
 4. Lack of variance in ratings: Participants were excluded if they only used a total of 2 unique rating values (out of 7) for all items in the survey. In addition, for the 2-question "form and meaning" survey, participants were excluded if they always provided the same rating for two questions across all items.

**5. Correlation with other participants:** Participants were excluded if the average Pearson correlation with the ratings of remaining participants fell below 2 standard deviations below the mean inter-participant correlation. The inter-participant correlated was computed by correlating a vector of responses for a given participant with the vector of responses for each of the remaining participants and taking the average of these pairwise correlation values.

## Non-participation

No participants dropped out.

## Randomization

Participants were not allocated to experimental groups. There was no randomization procedure for participant selection or enrollment.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

- |                                     |                               |
|-------------------------------------|-------------------------------|
| n/a                                 | Involved in the study         |
| <input checked="" type="checkbox"/> | Antibodies                    |
| <input checked="" type="checkbox"/> | Eukaryotic cell lines         |
| <input checked="" type="checkbox"/> | Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | Animals and other organisms   |
| <input checked="" type="checkbox"/> | Clinical data                 |
| <input checked="" type="checkbox"/> | Dual use research of concern  |
| <input checked="" type="checkbox"/> | Plants                        |

### Methods

- |                                     |                        |
|-------------------------------------|------------------------|
| n/a                                 | Involved in the study  |
| <input checked="" type="checkbox"/> | ChIP-seq               |
| <input checked="" type="checkbox"/> | Flow cytometry         |
| <input type="checkbox"/>            | MRI-based neuroimaging |

## Magnetic resonance imaging

### Experimental design

## Design type

Localizer experiments were blocked.

The critical experiment consisted of both an event-related design and a blocked design. n=10 participants took part in the event-related experiment (5 participants completed two sessions each, and 5 participants completed 3 sessions each). n=4 participants took part in the blocked experiment (one session each).

## Design specifications

For the language localizer task: Blocked design with a counterbalanced condition order across runs. Each stimulus consisted of 12 words/nonwords. Stimuli were presented in the center of the screen, one word/nonword at a time, at the rate of 450ms per word/nonword. Each stimulus was preceded by a 100ms blank screen and followed by a 400ms screen showing a picture of a finger pressing a button, and a blank screen for another 100ms, for a total trial duration of 6s. Experimental blocks lasted 18s (with 3 trials per block), and fixation blocks lasted 14s. Each run (consisting of 5 fixation blocks and 16 experimental blocks) lasted 358s. Participants completed 2 runs.

For the multiple demand localizer task: MD localizer task: Blocked design with a counterbalanced condition order across runs. The runs consisted of easy and hard arithmetic conditions. The arithmetic task (numbers) were presented in the center of the screen for 1,450ms, followed by the response choices presented for 1,450ms and an inter-stimulus interval of 100ms. Experimental blocks lasted 15 s (with 5 trials per block), and fixation blocks lasted 15s. Each run consisted of 16 experimental blocks—8 blocks per condition—and 5 fixation blocks; a fixation block appeared at the beginning of the run and after each set of four experimental blocks, and lasted 315s. Participants completed 2 runs. For the critical event-related experiment: Participants passively read each sentence once, in a condition-rich, event-related fMRI design (each sentence is effectively a condition). Sentences were presented (in black font) on a light grey background one at a time for 2s with a 400ms ISI consisting of a fixation cross. Each run contained 50 unique sentence trials and three 12s fixation blocks (in the beginning, middle (i.e., after 25 sentences) and end of each run). Each run lasted 336s (5:36 minutes). Participants completed either a total of 20 runs (n=5 participants across two sessions) or 30 runs (n=5 participants across three sessions).

For the critical blocked experiment: Sentences were grouped into blocks of 5 sentences from the same condition (baseline, drive, suppress) and were presented on the screen (in black font on a light grey background) one at a time for 2s with a 400ms ISI consisting of a fixation cross (for a total block duration of 12s). Each run consisted of 24 blocks with 8 blocks (40 sentences) per condition. There were five 12s fixation blocks: in the beginning and end of each run, as well as after 6, 12, and 18 blocks. Each run lasted 348s (5:48 minutes). Participants (n=4) completed a total of six runs.

## Behavioral performance measures

Button press in the language localizer task. Forced-choice in the multiple demand localizer task. Behavioral performance measures were not analyzed for the localizers.

For the critical experiment, no task was performed during data acquisition. However, participants completed a memory task at the end of each scanning session (outside of the scanner) to incentivize attention throughout the session. Participants were informed ahead of time that they would be asked to perform a memory task after the scanning session. Behavioral responses were analyzed: The average accuracy (sum of correct responses divided by total number of responses) was 70.4% (SD across sessions: 11.4%) for the event-related participants (n=24 sessions – responses for

one session were not saved due to an error in the script), and 61.7% (SD across sessions: 10%) for the blocked participants (n=4 sessions).

## Acquisition

Imaging type(s)

Structural and functional.

Field strength

3T.

Sequence & imaging parameters

T1-weighted, Magnetization Prepared RApid Gradient Echo (MP-RAGE) structural images were collected in 176 sagittal slices with 1 mm isotropic voxels (TR = 2,530 ms, TE = 3.48 ms, TI = 1100 ms, flip = 8 degrees). Functional, blood oxygenation level dependent (BOLD), data were acquired using an SMS EPI sequence (with a 90 degree flip angle and using a slice acceleration factor of 2), with the following acquisition parameters: fifty-two 2 mm thick near-axial slices acquired in the interleaved order (with 10% distance factor) 2 mm × 2 mm in-plane resolution, FoV in the phase encoding (A ≫ P) direction 208 mm and matrix size 104 × 104, TR = 2,000 ms and TE = 30 ms, and partial Fourier of 7/8. The first 10 s of each run were excluded to allow for steady state magnetization.

Area of acquisition

Whole brain.

Diffusion MRI

Used

Not used

## Preprocessing

Preprocessing software

SPM12 and custom MATLAB scripts.

Normalization

All functional scans were coregistered and resampled using B-spline interpolation to the first scan of the first session.

Normalization template

SPM12 default Montreal Neurological Institute (MNI) template.

Noise and artifact removal

For the blocked experiments, data were high-pass filtered at 128s. For the event-related experiments, the data-driven analysis method GLMdenoise and the statistical technique of ridge regression were used. These methods can account for a variety of sources of noise (e.g., physiological, motion, scanner artifacts, effects of collinearity).

Volume censoring

For the blocked experiments, time points classified as outliers based on the motion data each had a regressor included in the GLM but were not removed (outliers were identified from the resulting subject-motion estimates as well as from BOLD signal indicators using default thresholds in CONN preprocessing pipeline: 5 standard deviations above the mean in global BOLD signal change, or framewise displacement values above 0.9 mm; Nieto-Castanon, 2020). For the event-related experiment, no volume censoring was performed.

## Statistical modeling & inference

Model type and settings

Single-trial BOLD response amplitudes were estimated for individual voxels in individual participants. Voxels were aggregated by averaging across voxels in pre-defined regions of interest (see "Anatomical location(s)" below). Univariate and predictive (encoding model) analyses were performed on these ROI-level responses.

Effect(s) tested

Main effects were estimated using LME models with the following formulae:  
 $BOLD\ response \sim variable\_of\_interest + (1 | sentence) + run\_within\_session + trial\_within\_run$   
 Where "variable\_of\_interest" is either the condition (drive, suppress, baseline) or the behavioral sentence property of interest.

Specify type of analysis:  Whole brain  ROI-based  Both

Anatomical location(s)

Functionally-located language regions were the primary brain ROIs. These were defined as follows: Each individual map for the sentences > nonwords contrast from the language localizer was intersected with a set of 10 binary parcels (both hemispheres) derived from a probabilistic activation overlap map for the same contrast in a large set of participants (n=220) using watershed parcellation (an approach developed in Fedorenko et al., 2010).

For supplementary analyses, we further defined ROIs based on the multiple demand localizer, also intersected with a set of binary parcels derived using the approach described in Fedorenko et al., 2010. Finally, in an attempt to cover a large part of the cortex, we obtained ROIs from the Glasser parcellation (Glasser et al., 2016).

Statistic type for inference

(See [Eklund et al. 2016](#))

The Eklund paper concerns traditional group-level random effect analyses, in the current paper all the analyses are performed within individuals and then the extracted responses are analyzed with linear mixed effects models and correlation measures.

Correction

Not applicable, as voxel-wise statistical significance inferences are not included in this paper.

## Models & analysis

n/a Involved in the study

Functional and/or effective connectivity

Graph analysis

Multivariate modeling or predictive analysis

### Multivariate modeling and predictive analysis

For the encoding model analyses, we mapped from language model representations to brain responses using a ridge regression mapping model (L2 regularized). The regularization parameter was estimated using leave-one-out cross-validation implemented using the scikit-learn Python library function RidgeCV (Pedregosa et al., 2011; version 0.24.2). No dimensionality reduction was performed.