Driving and suppressing the human language network using large language models

Greta Tuckute[1,2], Aalok Sathe[1,2], Shashank Srikant[3,4], Maya Taliaferro[1,2], Mingye Wang[1,2], Martin Schrimpf[5], Kendrick Kay[6], Evelina Fedorenko[1,2,7]

1. Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139 USA
2. McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139 USA
3. Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA
4. MIT-IBM Watson AI Lab, Cambridge, MA 02142, USA
5. Quest for Intelligence, Massachusetts Institute of Technology, Cambridge, MA 02139 USA
6. Center for Magnetic Resonance Research, University of Minnesota, Minneapolis, MN 55455 USA
7. The Program in Speech and Hearing Bioscience and Technology, Harvard University, Cambridge, MA 02138 USA

Corresponding authors: Greta Tuckute (gretatu@mit.edu) and Evelina Fedorenko (evelina9@mit.edu).

Abstract

Transformer language models are today's most accurate models of language processing in the brain. Here, using fMRI-measured brain responses to 1,000 diverse sentences, we develop a GPT-based encoding model to identify new sentences that are predicted to drive or suppress responses in the human language network. We demonstrate that these model-selected 'out-of-distribution' sentences indeed drive and suppress activity of human language areas in new individuals (85.7% increase and 97.5% decrease relative to the diverse naturalistic sentences). 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 accurate models of the brain to noninvasively control neural activity in higher-level cortical areas, like the language network.

Introduction

Reading and understanding these words engages a set of left-lateralized frontal and temporal brain regions. These interconnected areas, or the 'language network' (e.g., [1–3]), a) support both comprehension and production of spoken, written, and signed linguistic utterances (e.g., [4,2,5–10] across diverse languages [11,12]; b) are highly selective for language relative to diverse non-linguistic inputs (e.g., [13–17]; see [18] for a review); c) are sensitive to linguistic structure at many levels (e.g., [2,14,6,19–21,18,10]); and d) are causally important for language such that their damage leads to linguistic deficits [22–28]. However, many aspects of the representations and algorithms that support language comprehension remain unknown.

Past accounts of language comprehension in the brain have been dominated by verbal descriptive models—expressed in plain English—where different brain regions are coarsely tied to cognitive constructs like 'syntactic structure building', 'lexical access', and 'semantic composition' (e.g., [29–34]). Although interpretability and intuitive-level understanding is a key goal of scientific models, this class of models leave much of the detail of neural responses to language unexplained. Another key goal of scientific models is high generality and predictive power. That is, a model should allow us to predict, with high accuracy, responses to any arbitrary stimulus. Until recently, we have lacked adequate computationally precise and stimulus-computable accounts of how meaning can be extracted from language. However, a few years ago, this barrier was lifted: increases in computing power, availability of vast text corpora, and machine learning advances have led to engineering breakthroughs in artificial intelligence, including language. Neural network models have emerged that can generate coherent text, answer questions, translate between languages, and perform at human or near-human levels on diverse language tasks (e.g., [35–41]).

Impressed by these models' successes, language researchers began using these large language models (LLMs) as models of human language processing. Strikingly, despite the fact that these models were not developed with the goal of modeling human language, some of these models (especially the unidirectional Transformer architectures; [36]) appear to be able to capture human data, both behavioral (e.g., [42–45]) and neural (e.g., [46–58]). As such, this new class of computationally precise large language models (LLMs) has allowed the field of neurobiology of language to move beyond verbal descriptive models of language processing with promise of major advances in our understanding of human language processing.

This work has two goals. The first goal is to put the new class of models, LLMs, to a strong test as models of language processing. Prior claims about the apparent similarity between LLMs and the brain have relied on neural responses to naturally occurring sentences, which cover a restricted portion of the space of linguistic/semantic variation. Further, the mapping model (from LLM representations to human neural data) is typically trained and tested on data from the same participant (cf. [56,59]). These two limitations have left it unclear whether the similarity between LLM and humans i) only holds within a narrow distribution of linguistic input and/or ii) is overly reliant on patterns of participant-specific idiosyncrasies. Here, we leverage the predictive power of LLMs to identify new stimuli that would maximally drive or suppress brain responses in

the language network. These model-selected stimuli constitute the edges of the stimulus-response distribution and enable comparative model-brain analyses that go beyond previous studies of naturally occurring sentences. Critically, we collect brain data to these new model-selected stimuli in new participants in a 'closed-loop' manner (e.g., [60–62]), testing whether LLMs capture features of language representations that generalize across humans. The second goal of this work is to obtain intuitive-level understanding of language processing by characterizing—across a broad distribution of linguistic input and associated brain responses—the stimulus properties that the language network is most responsive to.

Results

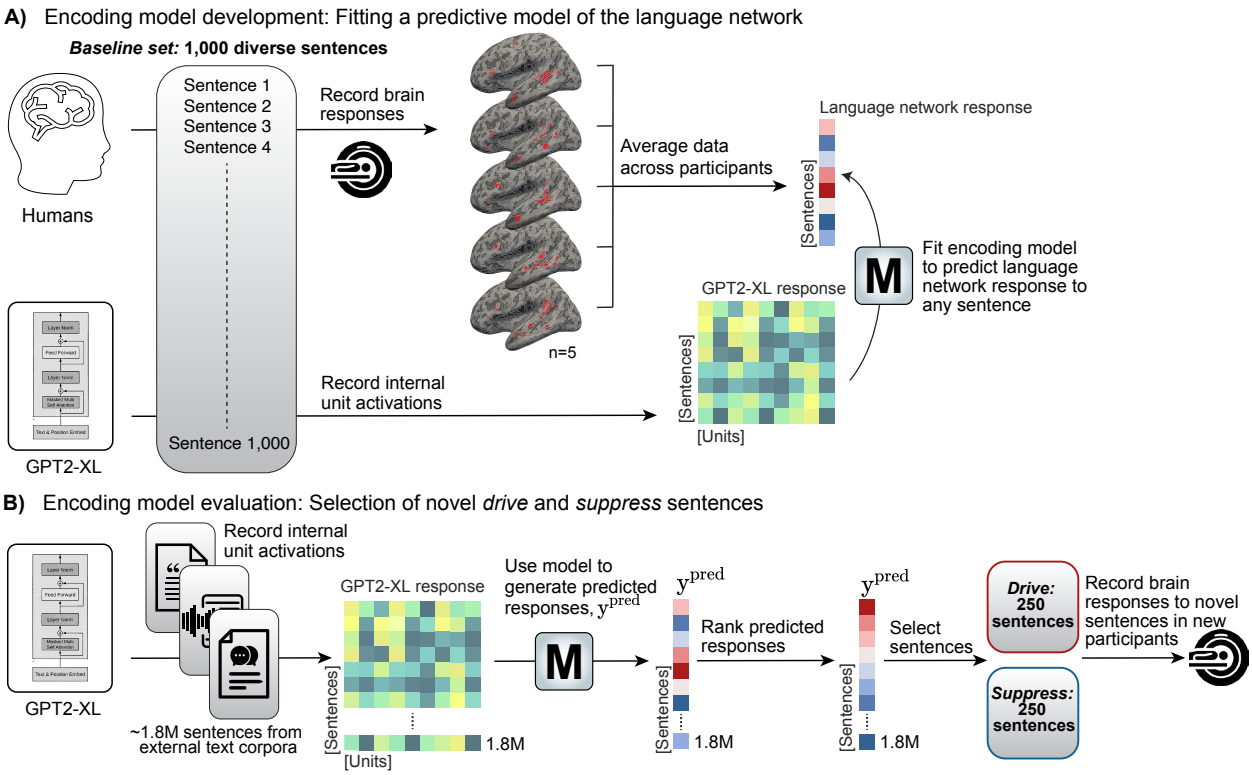

Figure 1. Overview of the procedure for encoding model development and stimulus selection for evaluation.

A) We developed an encoding model (**M**) of the left hemisphere (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 (Methods; Encoding model development). Five participants (*train* participants) read a large sample (n=1,000) of 6-word corpus-extracted sentences, the *baseline set* (sampled to maximize linguistic diversity; Extended Methods 1.1), in a rapid, event-related design while their brain activity was recorded using fMRI. Blood-oxygen-level-dependent (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 model in Schrimpf et al. [48]), to the 1,000 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; SI 5A).
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 minimal extent (*suppress* sentences) (Methods; Encoding model evaluation). 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).

Model-selected sentences successfully drive and suppress brain responses in the language network of new individuals

We wanted 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 [36] (layer 22, see SI 5A for the cross-validated analysis that led to this choice) and was trained to predict the average LH language network's (functionally defined; [2]) BOLD response (henceforth language network's response; Extended Methods 2.4) from 5 *train* participants who read a set of 1,000 diverse, corpus-extracted sentences (*baseline set*) (2 sessions each, n=10 sessions total; Extended Methods 2.2.1) (**Figure 1A**). The encoding model achieved a prediction performance of r=0.38 (noise ceiling is r=0.56; SI 4) when evaluated on held-out sentences within the *baseline set* (SI 5A). To ensure encoding model robustness, 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; SI 5B) and even using sentence embeddings from a different LLM architecture (a bidirectional-attention Transformer model, BERT-large; SI 5C). Finally, the encoding model also achieved relatively high predictivity performance on anatomically, rather than functionally, defined language regions, although predictivity was lower (SI 5D).

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 9 diverse large-scale text corpora (**Figure 1B**). We identified a set of 250 sentences that were predicted to elicit maximally strong activity in the language network (*drive* sentences; e.g., "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; e.g., "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 participants).

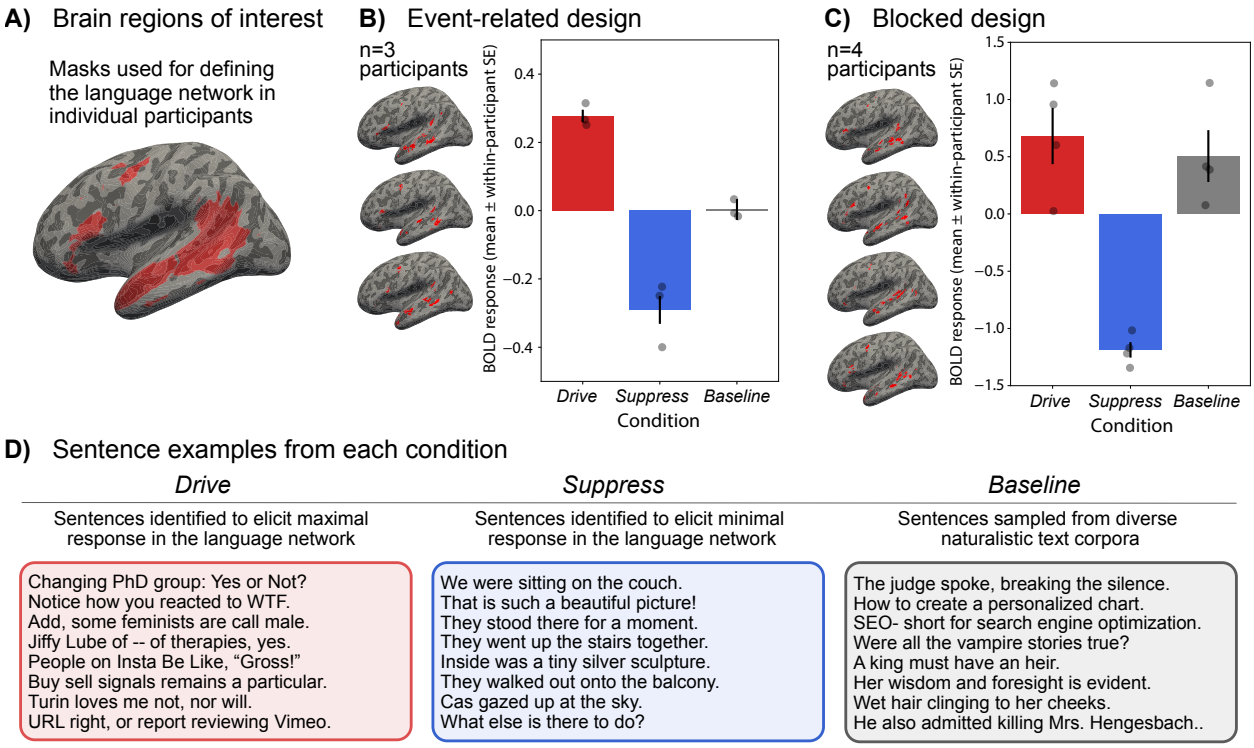

Figure 2. Model-selected sentences successfully drive and suppress responses in the language network.

A) We identified sentences to elicit maximal response (*drive* sentences) or minimal response (*suppress* sentences) in the functionally defined language network. To define functional ROIs, we used demarcations (‘language parcels’; shown on the surface-inflated MNI152 template brain) within which most or all individuals in prior studies showed activity for the language localizer contrast in large samples (e.g., [2,3]). 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 brain visualizations in B and C).

B) The average language network fMRI response across, respectively, 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, individual points show the average of each condition per participant. The average BOLD response was 85.70% higher for *drive* relative to *baseline* and 97.51% lower for *suppress* relative to *baseline* (SI 9A). Error bars show 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.

C) The average language network fMRI response across, respectively, 240 *drive,* 240 *suppress*, and 240 *baseline* sentences (randomly sampled from the superset of 250/250 *drive/suppress* sentences and 1,000 *baseline* sentences) for n=4 *evaluation* participants, collected in a blocked fMRI paradigm. The average BOLD response was 12.87% higher for *drive* relative to *baseline* and 56.64% lower for *suppress* relative to *baseline* (SI 9B).

D) Example sentences from each condition.

We collected fMRI responses to the *drive* and *suppress* sentences in an event-related, single-trial paradigm in three new participants (3 sessions each, n=9 sessions total; Extended Methods 2.2.1). 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 (β=0.57, t=15.94, p<.0001 using linear mixed effect modeling, see Methods; Statistical analyses and SI 14). The *drive* sentences also yielded significantly higher responses than the *baseline* sentences (β=0.27, t=9.72, p<.0001) with the average BOLD signal being 85.70% higher for *drive* relative to *baseline* (SI 9A). Finally, the *suppress* sentences yielded lower responses than the baseline sentences (β=-0.29, t=-10.44, p<.0001) with the average responses being 97.51% lower for *suppress* relative to *baseline* (SI 9A). 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 in new participants to demonstrate that these sentences modulate brain responses as predicted (see SI 12F for evidence that this pattern of responses to *drive, suppress*, and *baseline* sentences is not ubiquitously present across the brain).

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 are blocked into groups (cf. randomly interspersed), in four additional participants (1 session each, n=4 sessions total; Extended Methods 2.2.2). The results mirrored those from the event-related experiment: the *drive* sentences yielded the highest response followed by *baseline* sentences (average increase of 12.87% from *baseline* to *drive*; average decrease of 56.64% from *baseline* to *suppress*; SI 9B). Hence, independent of experimental design (event-related vs. blocked) and b) modeling procedure (single-trial modeling vs. condition-level modeling), the brain responses to the *drive* sentences were high relative to the baseline sentences, and the responses to the *suppress* sentences were low relative to the baseline sentences.

For a final examination of the robustness of the results, we explored an alternative approach to identifying *drive*/*suppress* sentences (the *modify* approach: gradient-based modifications to transform a random sentence into a novel sentence predicted to elicit high or low fMRI responses; SI 13A) and collected responses to these novel sentences in two participants (event-related design, n=6 sessions total). We found that this exploratory *modify* approach was able to drive responses by 56.97% relative to *baseline*, but failed to suppress responses (SI 13B).

Model predictions for new sentences capture most of the explainable variance in new participants

In the previous section, we examined predictivity at the condition level (*drive* vs. *suppress* vs. *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 (**Figure 2B**), which allows us to estimate sentence-level brain responses to 1,500 sentences for each of the three *evaluation* participants.

Figure 3A shows the observed versus predicted brain responses in the language network. The predictions were derived from the encoding model, and the actual responses were obtained from the 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 ($p<.0001$) 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. Therefore, we isolated the set of n=1,000 naturalistic, corpus-extracted *baseline* sentences and obtained a correlation of 0.30 ($p<.0001$). 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, within-distribution sentences (*baseline set*) and out-of-distribution sentences (*drive/suppress* set).

To better interpret the accuracy of sentence-level predictions, we quantified the maximal possible prediction performance by treating inter-participant variability as “noise” that cannot be predicted by a computational model that has no access to measurements from a new participant. The goal here is to assess how well our model predicts brain activity at the group level, taking into account irreducible variance due to inter-participant variability and/or measurement noise. First, we computed the empirical variability in participants’ responses across all 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 (x-axis in **Figure 3A**) and averaged across participants. This simulation provides an estimate of the maximum possible encoding model prediction performance.

Figure 3 inset shows these simulated brain responses versus predicted responses. In these simulations, the Pearson correlation was 0.62 across all 1,500 sentences (observed: r=0.43, i.e., 69.4% of the theoretically obtainable correlation), and 0.39 across the 1,000 *baseline* sentences (observed: r=0.30, i.e., 76.9% of the theoretically obtainable correlation). These results indicate that the ability of our modeling approach to predict brain responses in novel participants is actually quite high. We have shown that given inter-participant variability in fMRI measurements, even a perfect model can achieve only r=0.62 predictive performance. Although our model is not perfect (r=0.43, i.e., 69.4% of the theoretically obtainable correlation), the performance level suggests that the model is successfully capturing much of the neurally relevant variance in responses to individual sentences.

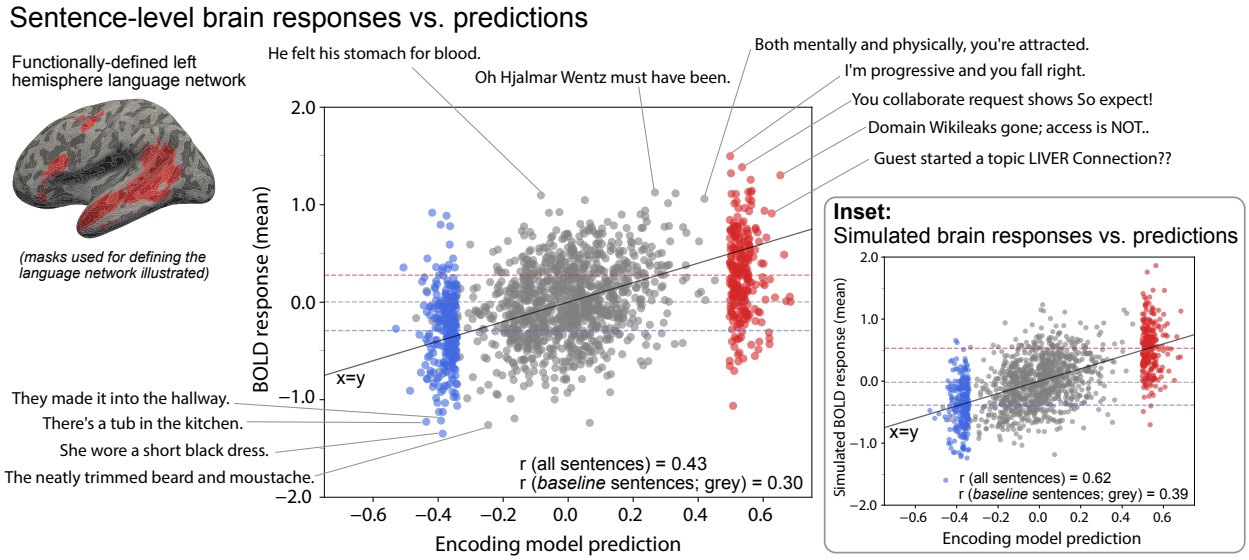

Figure 3. The encoding model maintains high predictive performance for brain responses from three new participants.
Sentence-level brain responses as a function of the predicted responses along with sentence examples. Predicted brain responses were obtained from the encoding model (x-axis). The actual brain responses (y-axis) are the average of n=3 *evaluation* participants' language network responses (illustrated for individual participants in SI 10). 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 respectively low or high brain responses and are therefore clustered on the low and high end of the prediction axis (x-axis). Dashed horizontal lines show the mean of each condition. Individual *drive* sentences had a BOLD increase up to 428.83% relative to the mean of the *baseline* condition (SI 11).
Inset: 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.

Language regions exhibit a high degree of stimulus-related activity relative to other brain regions, and this activity is well-accounted for by LLMs

We evidenced that an encoding model using features from GPT2-XL was predictive of brain responses in the language network for held-out sentences (SI 5A,B) and held-out participants (**Figure 3**; SI 6; SI 13B). To understand whether this prediction performance was spatially restricted to the language network, we first examined the amount of variance that any external predictor can theoretically explain (i.e., noise ceiling) in the language network versus other brain regions, and then evaluated encoding models that were trained to predict brain responses in the language network versus other brain regions.

First, we quantified noise ceilings for the language regions along with a set of control brain regions (**Figure 4A,B**). A noise ceiling (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 typically leverage repeated stimulus presentations, with the core assumption that repeated presentations should yield the same brain response (e.g., [63–65]). 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 Extended Methods 2.2.1 and Discussion), we developed a procedure for NC estimation by making use of the repeated presentations of the same sentence *across participants*, allowing for estimation of reliability in single-repetition paradigms (SI 4). Using this procedure, we computed NCs in language regions and a set of control regions based on the brain responses to the 1,000 *baseline* sentences for the n=5 *train* participants (**Figure 4A,B**). In addition to language regions, we examined i) two large-scale brain networks that have been linked to high-level cognitive processing—the multiple demand (MD) network [66–70] and the default mode network (DMN) [71–75]—which we defined using independent functional localizers (see SI 12 for details) (**Figure 4A**) and ii) a set of anatomical parcels [76] that cover a large fraction of the cortical surface (**Figure 4B**).

We found that language regions were characterized by the highest NC compared to both other functional areas and anatomical areas across the brain (including anatomical parcels that fall in spatially similar locations to the language areas, which provides further evidence for the advantages of functional localization; [77,78]). In particular, for the LH language areas, the NC was estimated to be r=0.56 (split-half standard error (SE)=0.03), i.e., ~31% of the variance in the responses of these areas can be considered “true”, stimulus-related, signal. For comparison: for the MD network, the NC was estimated to be r=0.07 (SE=0.11) (for the LH areas) and r=0.27 (SE=0.10) (for the RH areas; see [79] for convergent evidence from a different approach), and for the DMN, the NC was estimated to be r=0.11 (SE=0.11) (for the LH areas) and r=0 (SE=0.05) (for the RH areas) ($p<.0001$ for comparison of the LH language network against each of four networks—LH and RH MD and DMN—via Bonferroni-corrected t-tests using split-half bootstrap). 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 [79–84]), opening the door to investigations of what stimulus properties affect neural responses (see next section).

Second, we examined how well GPT2-XL features can explain brain responses in language regions versus other brain regions. To do so, we trained encoding models to predict responses in the control brain regions (defined functionally and anatomically) using features from GPT2-XL on the participant-averaged data from n=5 *train* participants on the *baseline set* (as in the main encoding model approach, see Extended Methods 3.2). **Figure 4C** shows the prediction performance on held-out sentences plotted against the NC for the functionally defined control brain regions across all 49 layers of GPT2-XL (see SI 7 for predictivity versus NC for the anatomical regions). Language regions were most reliable and well-predicted (red points in the upper right part of the plot). For example, the GPT2-XL features explained 67.86% of the variance in the LH language network (layer 22). The MD and DMN regions were clustered in the lower left part of the plot, characterized by relatively low NC and predictivity values. Hence, the

language regions were accounted for well by GPT-XL features in a spatially specific manner: other brain areas, including those that support high-level cognition, were not as reliable or well-predicted as language areas. (See SI 12F for the average brain responses to the *drive, suppress,* and *baseline* conditions for the functionally defined LH/RH language, MD and DMN networks.)

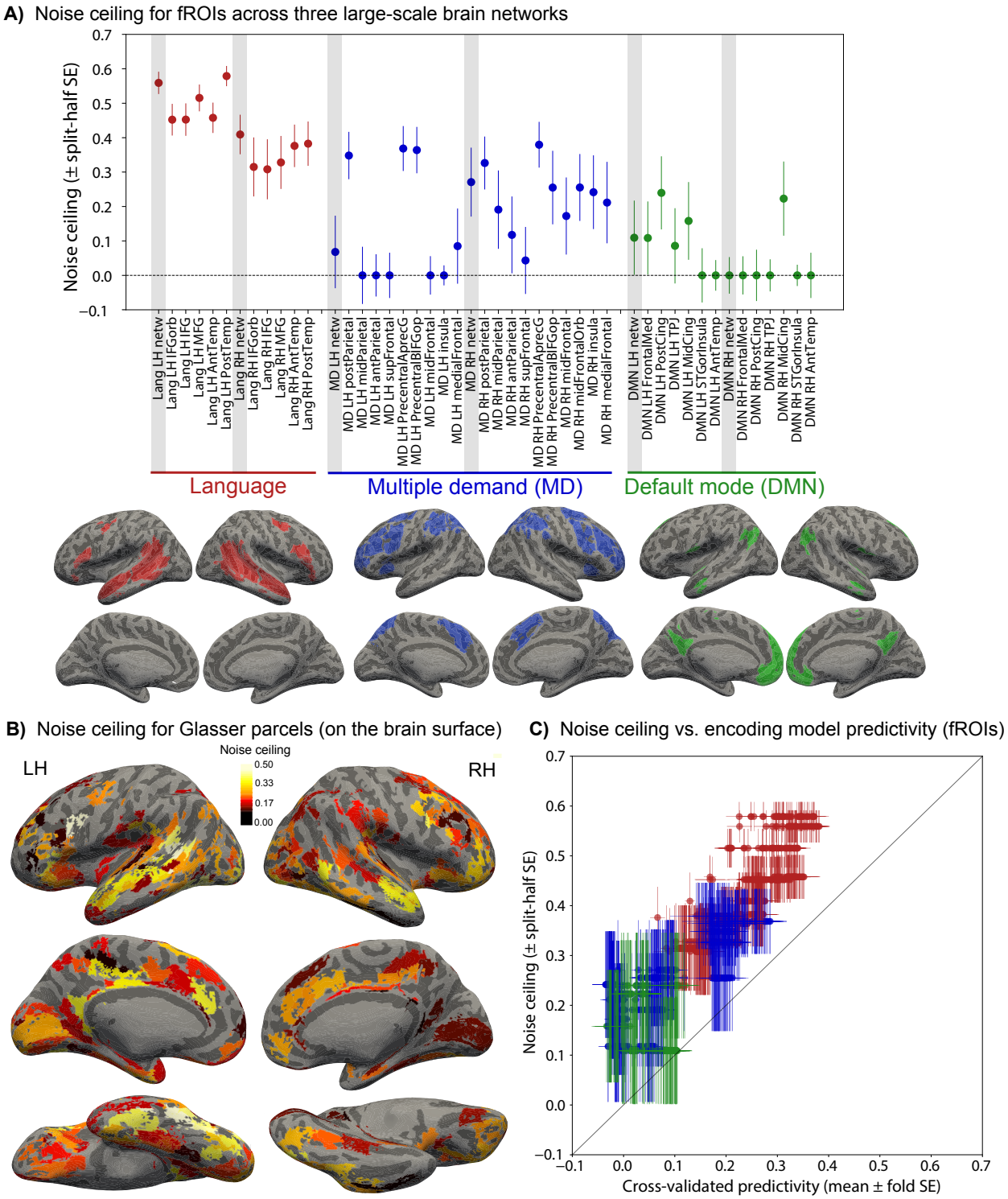

Figure 4. Left-hemisphere language regions show a high degree of stimulus-related activity for linguistic input relative to other brain areas and are well-accounted for by LLM features.
A) We quantified the noise ceiling (NC), a measure of stimulus-related response reliability, across all functionally defined ROIs in the language network (red), multiple demand (MD) network (blue), and the default mode network (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, i.e., 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. Error bars show the split-half NC reliability (SI 4). Grey shaded areas indicate the network-level fROIs.
B) We quantified the NC for a set of anatomical parcels that cover the whole brain [76], here illustrated on the surface-inflated MNI152 template brain. The color shows the NC with lighter colors indicating higher values. Anatomical ROIs where the NC split-half reliability overlapped with zero are not shown (leaving 189 anatomical ROIs: 100 in LH; 89 in RH).
C) NC (y-axis) as a function of encoding model performance (x-axis) for the functional ROIs in A for which the NC reliability did not overlap with zero (30 fROIs in total; red points: language fROIs; blue points: MD fROIs; green points: DMN fROIs). Encoding model performance was obtained using 5-fold cross-validation (SI 5). The error bar on the x-axis shows the standard error of the mean across cross-validation folds. Performance for all layers of GPT2-XL is shown (49 layers). The error bar on the y-axis shows the split-half NC reliability.

Uncovering the sentence properties that drive neural responses in the language network

In order to understand 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, SI 13) and correlated these features with sentence-level brain responses (Methods; Investigating sentence properties that modulate brain responses). The choice of features was inspired by past work in linguistics/psycholinguistics and cognitive neuroscience of language. First, building on the body of evidence that surprisal (the degree of contextual predictability, which is typically estimated as negative log probability) modulates language processing difficulty, in both behavioral psycholinguistic work (e.g., [85–89]) and brain imaging investigations (e.g., [90–98]), we computed sentence-level log probability estimates for each of 2,000 sentences using GPT2-XL (Extended Methods 4.1). And second, we collected 10 behavioral rating norms from a total of n=3,600 participants (on average, 15.23 participants per sentence per norm, min: 10, max: 19). The norms spanned five broad categories and were selected based on prior work examining the effects of various properties of words and sentences on language processing, in behavioral studies (e.g., [99–101,86,102–104]) and neural ones (e.g., [105–112,21]). 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 SI 17C) 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 (e.g., [113–118]), we grouped these two norms with surprisal in the analyses. Furthermore, because more generally, surprisal likely captures diverse aspects of form and meaning, 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 (e.g., [119–122]), the next three norms probed different aspects of the sentence content: how much does the sentence make you think about i) others' mental states, ii) physical objects and their interactions, and iii) places and environments. The latter two have to do with the physical world,

and the former — with internal representations; the physical vs. social distinction is one plausible organizing dimension of meaning [123,124]. 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 vs. in conversational contexts.

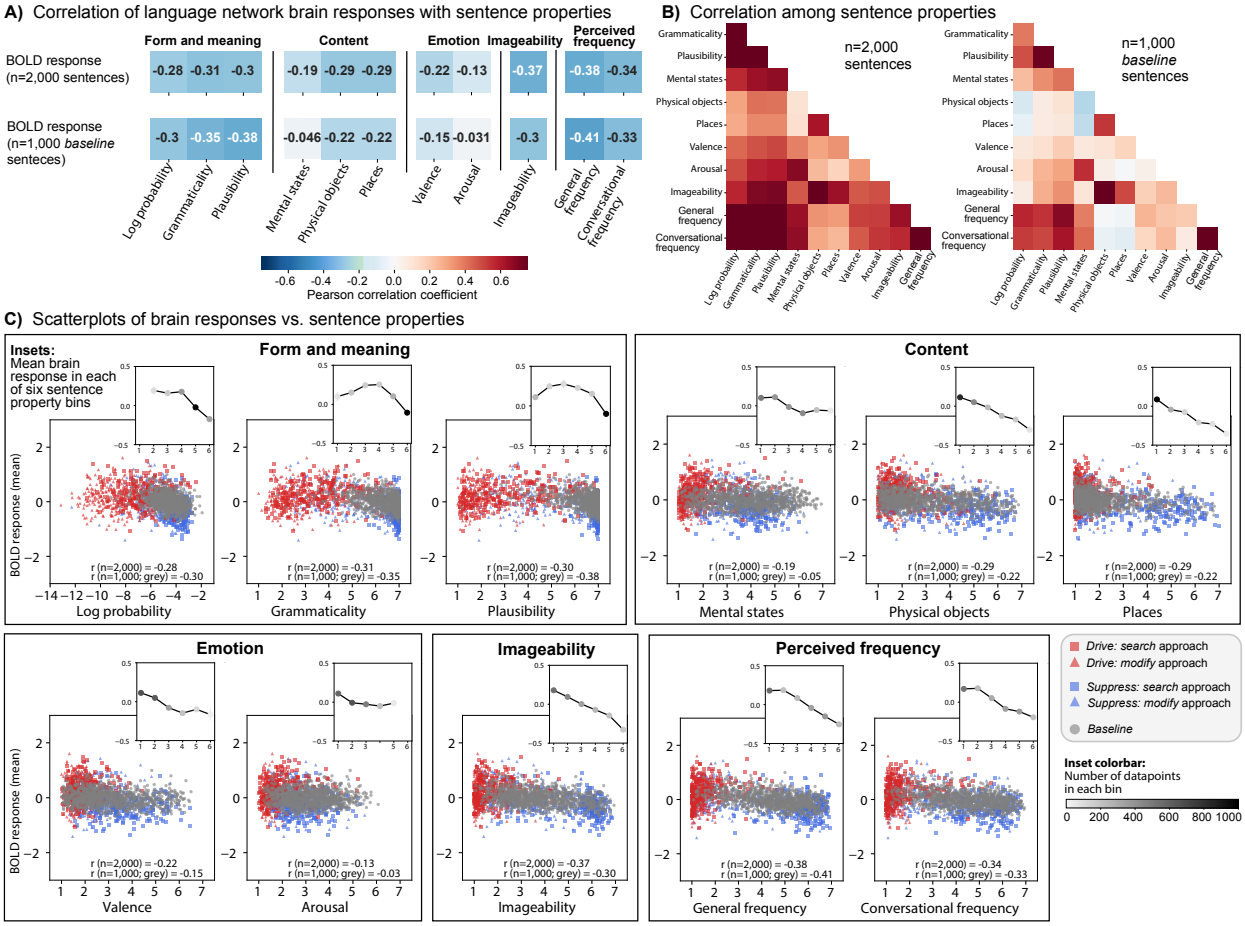

Figure 5. Surprisal and several other sentence properties modulate responses in the language network.

A) Correlation of the LH language network's response with 11 sentence properties (columns) within 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 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 matrix) or n=1,000 sentences (right matrix). Color scale 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 behavioral norming experiments in independent participants (besides the "Log probability" feature which was obtained from GPT2-XL). The inset line graphs show the average brain response with each property grouped into six uniformly sized bins. The color of the points in these graphs denotes the amount of data

in each bin (darker dots correspond to larger amounts of data; bins containing less than 1% of the data, i.e., 20 responses, were omitted from the line graphs).

Figure 5A shows the correlation between the language network's response and each of 11 sentence properties across the five categories. First, we found that less probable, i.e., more surprising, sentences elicited higher brain responses (**Figure 5C**) (r=-0.30 for the n=1,000 *baseline* sentences, p<.0001; see **Figure 5C** for the correlation values for the full set of n=2,000 sentences and SI 16 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 (e.g., [90,96]) and fMRI (e.g., [91–93,97]). Similarly, for the predictors related to a sentence's **grammaticality and plausibility**, sentences that were rated as less grammatical or plausible elicited higher responses (r=-0.31, r=-0.30, each p<.0001; 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 linear mixed effect models and compared these using likelihood ratio tests (see Methods; Statistical analyses and SI 18). Plausibility explained variance beyond surprisal and grammaticality ($\mathcal{X}^2$=17.86, p<.0001; all likelihood ratio statistics reported on the *baseline set*). Similarly, grammaticality explained variance beyond surprisal and plausibility ($\mathcal{X}^2$=12.97; p=.0003), albeit to a lesser extent. Interestingly, a finer-grained examination of the relationship between these features and neural responses reveals a non-linearity, 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 (**Figure 5D**; note also that the response to surprisal asymptotes at the higher end of the surprisal scale such that more surprising sentences no longer lead to stronger responses). This pattern suggests that two effects may be at play: an increase in neural response is seen i) for sentences that better adhere to form and meaning regularities of language (similar to the previously reported stronger responses to sentences than lists of words; e.g., [2,125]), and ii) for sentences that may have greater processing costs due to their unexpected form and/or meaning (e.g., see [92,126] for evidence of a strong relationship between behavioral processing difficulty and the strength of neural response in the language areas).

For the properties that relate to the sentence **content**, we evidenced no increase in explained variance (beyond surprisal) related to whether the sentence concerned others' mental states ($\mathcal{X}^2$=0.69, p=0.41). 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 (e.g., [127,83,128]) and challenges claims that the language areas are modulated by social content (e.g., [129–131]). However, whether the sentence's content concerned physical objects or places correlated negatively with brain responses (both r=-0.22) and explained variance beyond surprisal (physical objects: $\mathcal{X}^2$=74.26, p<.0001; places: $\mathcal{X}^2$=63.47, p<.0001; the two norms were correlated with each other at r=0.54). 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) and explained some variance beyond surprisal ($\mathcal{X}^2$=16.53, p<.0001). In contrast, whether the sentence was exciting did not explain additional variance beyond surprisal (r=-0.03, $\mathcal{X}^2$=0.18, p=0.67).

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) and explained variance beyond surprisal ($\mathcal{X}^2$=93.03, p=<.0001).

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), with additional variance explained beyond surprisal (general perceived frequency: $\mathcal{X}^2$=96.63, p<.0001; conversational perceived frequency: $\mathcal{X}^2$=44.46, p<.0001).

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 areas. 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 areas (**Figure 5**).

Discussion

We provide the first 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. By 'control', we mean the use of models to make quantitative predictions about a neural target of interest and subsequently using those predictions to successfully modulate neural activity in the target in a 'closed-loop' manner. We also richly characterize stimulus properties that modulate neural responses in the language network and find that less probable sentences generally elicit higher responses, with additional contribution from several form- and meaning-related features.

As foreshadowed in the introduction, in explaining the world, scientists strive for models that both have high predictive power and are interpretable. Similarly, in neuroscience, understanding can come in the form of quantitatively accurate models that can predict how a brain cell, region, or network of regions will respond to a particular stimulus. But understanding can also come in the form of intuitive descriptions of stimulus properties that lead to stronger or weaker neural responses (e.g., [132,133]). Here, we have coalesced these two kinds of understanding: we put predictive models to a strong test—assessing their ability to drive neural activity up or down—while, simultaneously, characterizing our stimuli in detail to gain insights into linguistic properties that modulate neural responses in the target brain areas.

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 (e.g., [134–136,46–52,55,58]). This study goes beyond prior work in four ways.

First, to develop our encoding model, we turned to a novel experimental approach. Robust encoding models require *reliable* neural responses to a *wide range* of stimuli. High reliability of neural responses is typically achieved through multiple repetitions of the same stimulus and averaging across repetitions. However, unlike largely bottom-up brain systems such as the ventral visual stream [137], which process stimuli in a similar way across multiple repetitions (e.g., [138,139]), repeating language stimuli like phrases or sentences is not effective because participants do not engage with the same sentence in the same way and/or to the same extent (unpublished data from the Fedorenko lab). Moreover, unlike a non-human primate with chronically implanted electrodes, where it is possible to obtain neural responses to thousands of stimuli (over the course of days or weeks; e.g., [140]), for humans, we are generally more limited in how much data is feasible to collect, both because of general constraints on participants' time and boredom/cognitive fatigue. One recent approach to combat the latter has been to turn to rich naturalistic stimuli, like stories, podcasts, or movies and to collect massive amounts of data (sometimes, many hours' worth) from a small number of individuals (e.g., [119,141,142])—what is often referred to as the 'deep data' approach (e.g., [143–148]). However, such stimuli do not sample the space of linguistic and/or semantic variation well, and consequently, do not allow for testing the model on stimuli that differ substantially from those used for training. Here, we solved these challenges by developing a single-repetition experimental paradigm with careful stimulus design at scale. Our stimulus set consisted of 1,000 semantically, syntactically, and stylistically diverse linguistic stimuli from multiple corpora (the diversity of the stimuli was confirmed by our behavioral norming experiments), making it the largest single-sentence dataset available to date. Another novel aspect of the experimental approach lies within the experimental design and analyses techniques: critically, we collected neural responses to each individual sentence for each participant in rapid, event-related fMRI, presented once to maximize engagement (cf. blocked designs and condition-level modeling). We extended existing state-of-the-art methods for single-trial modeling [149] to obtain robust estimates to each sentence. Similarly, we adapted methods (e.g., [65]) for noise ceiling estimation to quantify reliability in this novel single-trial regime. Hence, we acquired a large-scale, single-trial fMRI dataset spanning the linguistic space for optimal encoding model development in the domain of language.

Second, we evaluated our encoding model on new individuals. Most studies that have used language models to model human neural responses to language (e.g., [135,136,46,48–52,55,53,54,57,58]; cf. [56,59]) fit an encoding model using a given participant's data and then use that model to predict responses to held-out stimuli in that same participant. Consequently, such encoding models may learn a different function for each brain unit in each participant. Here we instead fitted an encoding model using data from five individuals, and then tested this model on new individuals. This approach ensures that the encoding model learns features about language processing that *generalize across human brains*.

Third and critically, we put our encoding model to a strong test by evaluating its ability to modulate the strength of neural responses via new model-selected stimuli. This predictive encoding modeling approach takes inspiration from adaptive stimulus design in systems neuroscience [150] which has recently been merged with the predictive power of neural networks in the domain of vision [60–62,139]. Unlike typical encoding or representational similarity approaches to testing neural networks as models of the brain, this approach exploits neural networks to generate 'optimal' stimuli for specific biological neurons or brain regions. By designing stimuli that would push neural activity towards a specified objective (e.g., driving or suppressing), we enter an "out of distribution" regime where the encoding model is tested on stimuli that may be quite different from those that it was trained on. Indeed, we discovered that *drive* sentences were *unusual* on various dimensions (SI 15). Given their properties, these sentences are *a priori* unlikely to be created or selected by experimenters and unlikely to be present in naturalistic stimuli, like stories, and yet these stimuli are able to drive responses in the language network, providing i) new insights into the selectivity of the language network and ii) an opportunity to test encoding models on stimuli that fall outside of the typical distribution of language input. The latter is critical for rigorous evaluation of model-to-brain alignment as it increases the likelihood of identifying discrepancies between models and brain/behavior which is needed for improvement of future models in a virtuous cycle.

And ***finally***, we combined the predictive encoding modeling approach with a detailed examination of stimulus properties in an effort to gain intuitive-level insights about what dimensions of stimulus variation affect neural responses in the language network. Past investigations have uncovered numerous word and sentence features that affect behavioral (e.g., [99–101,86,102–104]) and neural (e.g., [105–112,21]) responses to language. Inspired by these past studies, we examined the effects of 11 sentence properties on neural responses to 2,000 diverse sentences in the language network. In line with much past work (e.g., [90–98]), 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 non-linearity 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 a) the response to sentences in the low range, similar to the previously reported effects of stronger responses to phrases and sentences than lists of unconnected words (e.g., [2,151,125,126]). The response is also higher than b) the response to sentences in the high range—sentences that are highly plausible and use common grammatical structures—which are easy to process. Put differently, in order 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.

A few limitations and future directions are worth noting. First, we here studied the language network—comprised of three frontal and two temporal areas—as a whole. There are good reasons to adopt this approach: the different regions of this network i) have similar functional response profiles, both with respect to their selectivity for language (e.g., [13–15,17,18]) and their responses to linguistic manipulations (e.g., [21,152]), and ii) exhibit highly correlated time courses during naturalistic cognition paradigms (e.g., [80,82,153,127,154,11]). However, some functional heterogeneity has been argued to exist within the language network (e.g., [30,29,32,151,155,156]). Future efforts using an approach like the one adopted here could perhaps discover functional differences within the language network (by searching for stimuli that would selectively drive particular regions within 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. Furthermore, 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, e.g., based on the LLM representational space (e.g., [157,158]), 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 new individuals in the language network in a ‘closed-loop’ manner. This work has far-reaching implications for both neuroscientific research and application. 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. Moreover, accurate encoding models can be used as a ‘virtual language network’ to simulate experimental contrasts in silico (e.g., [159,139,160]). For prospective 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 (e.g., neurosurgical planning and intraoperative testing). Finally, integration of rapid advancements of large language models with larger and/or time-resolved measures of neural activity during language processing opens the door to a new class of non-invasive, therapeutic applications.

Methods

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

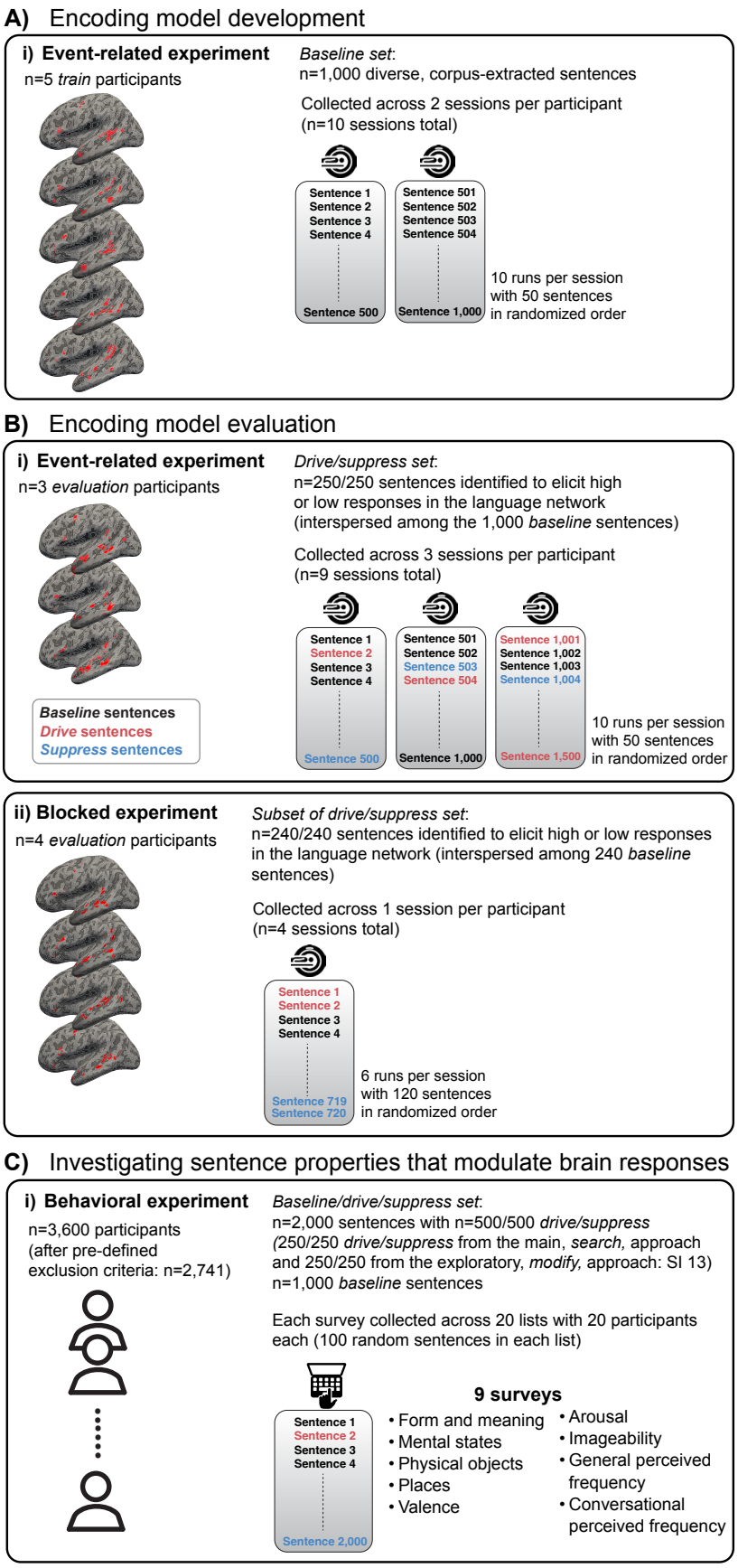

Figure 6. Experimental overview.
A) Encoding model development: We curated a large set (n=1,000) of diverse, corpus-extracted sentences (*baseline set)* and collected brain responses in n=5 *train* participants in an event-related fMRI design across two sessions per participant.
B) Encoding model evaluation: We identified a set of sentences to activate the language network to a maximal extent (250 *drive* sentences) or 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 (**panel 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.
C) Investigating sentence properties that modulate brain responses: In order to understand what sentence properties modulate brain responses in the language network, we collected 10 behavioral rating norms (across 9 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 SI 13) across n=3,600 participants.

Encoding model development

We developed an encoding model of the left hemisphere (LH) language network in the human brain. Developing an encoding model requires brain responses to a broad range of linguistic input. Therefore, we curated a large set of diverse, corpus-extracted 6-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 modeled those responses using a recently developed single-trial modeling framework [149], which we adapted for no-repeats designs (Extended Methods 2.2.1, Extended Methods 2.3.3 and SI 2). 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, etc.) (Extended Methods 1.1). In five *train* participants, we recorded brain responses to the sentences in the *baseline set* across two scanning sessions (**Figure 6A**). Participants were instructed to read attentively and think about the sentence's meaning. To encourage engagement with the stimuli, prior to the session, participants were informed that they would be asked to perform a short memory task after the session (Extended Methods 2.2.3). Sentences were presented one at a time for 2 seconds with a 4 second inter-stimulus interval. Each run contained 50 sentences (5:36 minutes) and sentence order was randomized across participants.

The language network was defined functionally in each participant using an extensively validated localizer task (e.g., [2,3]; Extended Methods 2.4). 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 i) the similarity among the five regions in their functional response profiles across dozens of experiments (e.g., [21,97,152]; see SI 3 for evidence of similar preferences for the *baseline set* in the current data), ii) high inter-regional correlations during naturalistic cognition paradigms (e.g., [80,82,153,127,154,11]). To mitigate the effect of collecting data across multiple scanning sessions and to equalize response units across voxels and participants, the blood-oxygen-level-dependent (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 (Extended Methods 2.6) 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 mapping model from the representations of the unidirectional-attention language model, GPT2-XL ([36]; identified as the most brain-aligned language model in Schrimpf et al. [48]), to the 1,000 fMRI responses (Extended Methods 3). Note that we use the term "encoding model" to describe a model that takes as input linguistic stimuli and predicts brain activity, i.e., the term encompasses both the LLM (GPT2-XL) used to transform a sentence into an embedding as well as the ridge regression mapping model used to map from LLM embedding to brain activity.

Encoding model evaluation

Using our trained encoding model, we identified a set of novel sentences to activate the language network to a maximal extent (*drive* sentences) or 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) (**Figure 6B**; Extended Methods 1.2.1). We collected brain responses to these novel sentences in 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 SI 13 for the details of methods and the results).)

To ensure that the results are 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 in 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*, *suppress* conditions; 240 per condition which were randomly sampled for each participant). Sentences were grouped into blocks of 5 sentences from the same condition and were presented on the screen one at a time for 2s with a 400ms inter-stimulus interval. Each run contained 120 sentences in 24 blocks (5:36 minutes). Condition order was counterbalanced across runs and participants (Extended Methods 2.2.2).

Investigating sentence properties that modulate brain responses

Finally, 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) (**Figure 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 behavioral psycholinguistic work (e.g., [85–89]) and brain imaging investigations (e.g., [90–98]), we computed the average log probability for each sentence using GPT2-XL (surprisal is negative log probability). And second, we collected 10 behavioral rating norms across a total of n=3,600 participants (on average, 15.23 participants per sentence per rating norm, min: 10, max: 19). The norms spanned five broad categories and were all motivated by prior work in linguistics and psycholinguistics. The first two 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 SI 17C) and plausibility (how much sense does the sentence make?). The next three norms probed different aspects of the sentence content: how much does the sentence make you think about i) others' mental states,

ii) physical objects and their interactions, and iii) places and environments. The latter two have to do with the physical world, and the former — with internal representations; the physical vs. social distinction is one plausible organizing dimension of meaning [123,124]. 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 vs. in conversational contexts. See SI 17 for details on experimental procedures.

Statistical analyses

Linear mixed effects (LME) models (implemented using the *lmer* function from the *lme4* R package [161]; version 1.1-31) were used to evaluate the statistical significance i) of the differences in the BOLD response among the sentence conditions (*baseline, drive,* and *suppress*) and ii) of the effect of sentence properties on the BOLD response. The critical variable of interest (either condition or sentence property) was modeled as fixed effect(s). As additional effects, we modeled 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):
BOLD response ~ variable_of_interest + (1 | sentence) + run_within_session + 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 SI 14 and SI 18) a table with model formulae, effect size estimates, standard error estimates, t-statistics, p-values, degrees of freedom, and R2 values. We evaluated the statistical significance of differences between pairs of conditions using estimated marginal means (implemented using the *emmeans* function from the *emmeans* R package [162]; version 1.8.4-1) using Tukey's multiple comparison method. Finally, we evaluated the statistical significance of differences between pairs of LMEs using likelihood ratio tests using the Chi Square value, $\mathcal{X}^2$, as the test statistic (implemented using the *anova* function from the *lme4* R package).

Data and code availability

Data and code will be made publicly available upon publication.

Acknowledgements

GT was supported by the Amazon Fellowship from the Science Hub (administered by the MIT Schwarzman College of Computing) and the International Doctoral Fellowship from American Association of University Women (AAUW). SS was supported by a grant from the MIT-IBM Watson AI Lab. EF was supported by NIH awards R01-DC016607, R01-DC016950, and U01-NS121471, as well as funds from the McGovern Institute for Brain Research, the Simons Center for the Social Brain, and the Brain and Cognitive Sciences Department.

We are grateful to Nancy Kanwisher, N Apurva Ratan Murty, Noga Zaslavsky, Kohitij Kar, Jacob Prince, Josh McDermott, and Una-May O'Reilly for valuable discussions. We thank Colton Casto, Elizabeth Lee, Benjamin Lipkin, Anna Ivanova, and Edward Gibson for their help on the project.

References

1. Bindereliz, J. R. *et al.* Human brain language areas identified by functional magnetic resonance imaging. *J. Neurosci. Off. J. Soc. 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. Lipkin, B. *et al.* Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. *Sci. Data* **9**, 529 (2022).
4. 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).
5. Menenti, L., Gierhan, S. M. E., Segaert, K. & Hagoort, P. Shared language: overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychol. Sci.* **22**, 1173–1182 (2011).
6. Regev, M., Honey, C. J., Simony, E. & Hasson, U. Selective and Invariant Neural Responses to Spoken and Written Narratives. *J. Neurosci.* **33**, 15978–15988 (2013).
7. Scott, T. L., Gallée, J. & Fedorenko, E. A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cogn. Neurosci.* **8**, 167–176 (2017).
8. 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).
9. Trettenbrein, P. C., Papitto, G., Friederici, A. D. & Zaccarella, E. Functional neuroanatomy of language without speech: An ALE meta-analysis of sign language. *Hum. Brain Mapp.* **42**, 699–712 (2021).
10. 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 N. Y. N 1991* bhac350 (2022) doi:10.1093/cercor/bhac350.
11. Malik-Moraleda, S. *et al.* The universal language network: A cross-linguistic investigation spanning 45 languages and 12 language families. 2021.07.28.454040 Preprint at https://doi.org/10.1101/2021.07.28.454040 (2022).
12. Malik-Moraleda, S. *et al.* Functional characterization of the language network of polyglots and hyperpolyglots with precision fMRI. *BioRxiv Prepr. Serv. Biol.* 2023.01.19.524657 (2023) doi:10.1101/2023.01.19.524657.
13. Monti, M. M., Parsons, L. M. & Osherson, D. N. Thought beyond language: neural dissociation of algebra and natural language. *Psychol. Sci.* **23**, 914–922 (2012).
14. Fedorenko, E., Behr, M. K. & Kanwisher, N. Functional specificity for high-level linguistic processing in the human brain. *Proc. Natl. Acad. Sci.* **108**, 16428–16433 (2011).
15. Ivanova, A. A. *et al.* Comprehension of computer code relies primarily on domain-general executive brain regions. *eLife* **9**, e58906 (2020).
16. Ivanova, A. A. *et al.* The Language Network Is Recruited but Not Required for Nonverbal Event Semantics. *Neurobiol. Lang.* **2**, 176–201 (2021).
17. Chen, X. *et al.* The human language system, including its inferior frontal component in “Broca’s area,” does not support music perception. *Cereb. Cortex* bhad087 (2023) doi:10.1093/cercor/bhad087.
18. Fedorenko, E. & Blank, I. A. Broca’s Area Is Not a Natural Kind. *Trends Cogn. Sci.* **24**, 270–284 (2020).
19. Blank, I., Balewski, Z., Mahowald, K. & Fedorenko, E. Syntactic processing is distributed across the language system. *NeuroImage* **127**, 307–323 (2016).
20. Bautista, A. & Wilson, S. M. Neural responses to grammatically and lexically degraded speech. *Lang. Cogn. Neurosci.* **31**, 567–574 (2016).

21. 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).
22. Damasio, H. & Frank, R. Three-dimensional in vivo mapping of brain lesions in humans. *Arch. Neurol.* **49**, 137–143 (1992).
23. Saffran, E. M. Aphasia and the relationship of language and brain. *Semin. Neurol.* **20**, 409–418 (2000).
24. Mesulam, M.-M. Primary progressive aphasia. *Ann. Neurol.* **49**, 425–432 (2001).
25. Bates, E. *et al.* Voxel-based lesion-symptom mapping. *Nat. Neurosci.* **6**, 448–450 (2003).
26. Mesulam, M.-M. *et al.* Primary progressive aphasia and the evolving neurology of the language network. *Nat. Rev. Neurol.* **10**, 554–569 (2014).
27. Wilson, S. M. *et al.* Language Mapping in Aphasia. *J. Speech Lang. Hear. Res. JSLHR* **62**, 3937–3946 (2019).
28. Andrews, J. P. *et al.* 160 The Neurosurgical Anatomy of Deficits in Speech Production. *Neurosurgery* **69**, 43 (2023).
29. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
30. Hagoort, P. On Broca, brain, and binding: a new framework. *Trends Cogn. Sci.* **9**, 416–423 (2005).
31. Price, C. J. The anatomy of language: a review of 100 fMRI studies published in 2009. *Ann. N. Y. Acad. Sci.* **1191**, 62–88 (2010).
32. Friederici, A. D. The Brain Basis of Language Processing: From Structure to Function. *Physiol. Rev.* **91**, 1357–1392 (2011).
33. Hagoort, P. MUC (Memory, Unification, Control) and beyond. *Front. Psychol.* **4**, 416 (2013).
34. Duffau, H., Moritz-Gasser, S. & Mandonnet, E. A re-examination of neural basis of language processing: Proposal of a dynamic hodotopical model from data provided by brain stimulation mapping during picture naming. *Brain Lang.* **131**, 1–10 (2014).
35. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018) doi:10.48550/arXiv.1810.04805.
36. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners. 24 (2019).
37. Liu, Y. *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv* (2019).
38. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *ArXiv200514165 Cs* (2020).
39. Rae, J. W. *et al.* Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *ArXiv* (2021).
40. Chowdhery, A. *et al.* PaLM: Scaling Language Modeling with Pathways. *ArXiv* (2022).
41. OpenAI. GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2023).
42. 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. *ArXiv* (2020).
43. Merkx, D. & Frank, S. L. Human Sentence Processing: Recurrence or Attention? in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* 12–22 (Association for Computational Linguistics, 2021). doi:10.18653/v1/2021.cmcl-1.2.
44. 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 https://doi.org/10.31234/osf.io/4hyna (2022).
45. Oh, B.-D. & Schuler, W. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Trans. Assoc. Comput. Linguist.* **11**, 336–350 (2023).
46. Toneva, M. & Wehbe, L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). 11 (2019).

47. Gauthier, J. & Levy, R. Linking artificial and human neural representations of language. in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* 529–539 (Association for Computational Linguistics, 2019). doi:10.18653/v1/D19-1050.
48. Schrimpf, M. *et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci.* **118**, e2105646118 (2021).
49. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5**, 134 (2022).
50. Goldstein, A. *et al.* Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).
51. Kumar, S. *et al.* Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. 2022.06.08.495348 Preprint at https://doi.org/10.1101/2022.06.08.495348 (2022).
52. Merlin, G. & Toneva, M. Language models and brain alignment: beyond word-level semantics and prediction. Preprint at https://doi.org/10.48550/arXiv.2212.00596 (2022).
53. Oota, S. R. *et al.* Neural Language Taskonomy: Which NLP Tasks are the most Predictive of fMRI Brain Activity? Preprint at https://doi.org/10.48550/arXiv.2205.01404 (2022).
54. Pasquiou, A., Lakretz, Y., Hale, J., Thirion, B. & Pallier, C. Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps. Preprint at https://doi.org/10.48550/arXiv.2207.03380 (2022).
55. Hosseini, E. A. *et al.* Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. 2022.10.04.510681 Preprint at https://doi.org/10.1101/2022.10.04.510681 (2022).
56. Tang, J., LeBel, A., Jain, S. & Huth, A. G. Semantic reconstruction of continuous language from non-invasive brain recordings. 2022.09.29.509744 Preprint at https://doi.org/10.1101/2022.09.29.509744 (2022).
57. Antonello, R. & Huth, A. Predictive Coding or Just Feature Discovery? An Alternative Account of Why Language Models Fit Brain Data. *Neurobiol. Lang.* 1–16 (2023) doi:10.1162/nol_a_00087.
58. 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).
59. 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).
60. Cowley, B. R., Williamson, R. C., Acar, K., Smith, M. A. & Yu, B. M. Adaptive stimulus selection for optimizing neural population responses. 11 (2017).
61. Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).
62. 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.e10 (2019).
63. Kay, K., Rokem, A., Winawer, J., Dougherty, R. & Wandell, B. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. *Front. Neurosci.* **7**, (2013).
64. Lage-Castellanos, A., Valente, G., Formisano, E. & Martino, F. D. Methods for computing the maximum performance of computational models of fMRI responses. *PLOS Comput. Biol.* **15**, e1006397 (2019).
65. Allen, E. J. *et al.* A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* **25**, 116–126 (2022).
66. Duncan, J. The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends Cogn. Sci.* **14**, 172–179 (2010).
67. Duncan, J., Schramm, M., Thompson, R. & Dumontheil, I. Task rules, working memory, and fluid intelligence. *Psychon. Bull. Rev.* **19**, 864–870 (2012).

68. Fedorenko, E., Duncan, J. & Kanwisher, N. Broad domain generality in focal regions of frontal and parietal cortex. *Proc. Natl. Acad. Sci.* **110**, 16616–16621 (2013).
69. Assem, M., Glasser, M. F., Van Essen, D. C. & Duncan, J. A Domain-General Cognitive Core Defined in Multimodally Parcellated Human Cortex. *Cereb. Cortex* **30**, 4361–4380 (2020).
70. Duncan, J., Assem, M. & Shashidhara, S. Integrated Intelligence from Distributed Brain Activity. *Trends Cogn. Sci.* **24**, 838–852 (2020).
71. Raichle, M. E. *et al.* A default mode of brain function. *Proc. Natl. Acad. Sci.* **98**, 676–682 (2001).
72. 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).
73. Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R. & Buckner, R. L. Functional-Anatomic Fractionation of the Brain's Default Network. *Neuron* **65**, 550–562 (2010).
74. Braga, R. M. & Buckner, R. L. Parallel Interdigitated Distributed Networks within the Individual Estimated by Intrinsic Functional Connectivity. *Neuron* **95**, 457-471.e5 (2017).
75. Buckner, R. L. & DiNicola, L. M. The brain's default network: updated anatomy, physiology and evolving insights. *Nat. Rev. Neurosci.* **20**, 593–608 (2019).
76. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
77. Saxe, R., Brett, M. & Kanwisher, N. Divide and conquer: a defense of functional localizers. *NeuroImage* **30**, 1088–1096; discussion 1097-1099 (2006).
78. Nieto-Castañón, A. & Fedorenko, E. Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *NeuroImage* **63**, 1646–1669 (2012).
79. 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).
80. Wilson, S. M., Molnar-Szakacs, I. & Iacoboni, M. Beyond Superior Temporal Cortex: Intersubject Correlations in Narrative Speech Comprehension. *Cereb. Cortex* **18**, 230–242 (2008).
81. 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).
82. 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).
83. Paunov, A. M. *et al.* Differential Tracking of Linguistic vs. Mental State Content in Naturalistic Stimuli by Language and Theory of Mind (ToM) Brain Networks. *Neurobiol. Lang.* **3**, 413–440 (2022).
84. Sueoka, Y., Paunov, A., Ivanova, A., Blank, I. A. & Fedorenko, E. The language network reliably 'tracks' naturalistic meaningful non-verbal stimuli. 2022.04.24.489316 Preprint at https://doi.org/10.1101/2022.04.24.489316 (2022).
85. Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C. & Pollatsek, A. The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychol. Aging* **21**, 448–465 (2006).
86. Demberg, V. & Keller, F. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* **109**, 193–210 (2008).
87. Levy, R. Expectation-based syntactic comprehension. *Cognition* **106**, 1126–1177 (2008).
88. Smith, N. J. & Levy, R. The effect of word predictability on reading time is logarithmic. *Cognition* **128**, 302–319 (2013).
89. 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).

90. Frank, S. L., Otten, L. J., Galli, G. & Vigliocco, G. The ERP response to the amount of information conveyed by words in sentences. *Brain Lang.* **140**, 1–11 (2015).
91. Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P. & van den Bosch, A. Prediction During Natural Language Comprehension. *Cereb. Cortex N. Y. N 1991* **26**, 2506–2516 (2016).
92. 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).
93. Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M. & Hale, J. T. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain Lang.* **157–158**, 81–94 (2016).
94. Lopopolo, A., Frank, S. L., van den Bosch, A. & Willems, R. M. Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PLoS ONE* **12**, e0177794 (2017).
95. Heilbron, M., Ehinger, B., Hagoort, P. & de Lange, F. P. Tracking Naturalistic Linguistic Predictions with Deep Neural Language Models. in *2019 Conference on Cognitive Computational Neuroscience* (2019). doi:10.32470/CCN.2019.1096-0.
96. 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.* **119**, e2201968119 (2022).
97. 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).
98. 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.* 1–71 (2023) doi:10.1162/nol_a_00105.
99. Rayner, K. & Duffy, S. A. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Mem. Cognit.* **14**, 191–201 (1986).
100. Speer, S. R. & Clifton Jr., C. Plausibility and argument structure in sentence comprehension. *Mem. Cognit.* **26**, 965–978 (1998).
101. Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H. & Yap, M. J. Visual Word Recognition of Single-Syllable Words. *J. Exp. Psychol. Gen.* **133**, 283–316 (2004).
102. Brysbaert, M., Warriner, A. B. & Kuperman, V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res. Methods* **46**, 904–911 (2014).
103. Singh, A. D., Mehta, P., Husain, S. & Rajakrishnan, R. Quantifying sentence complexity based on eye-tracking measures. in *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)* 202–212 (The COLING 2016 Organizing Committee, 2016).
104. Arfé, B., Delatorre, P. & Mason, L. Effects of negative emotional valence on readers' text processing and memory for text: an eye-tracking study. *Read. Writ.* (2022) doi:10.1007/s11145-022-10362-7.
105. 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).
106. Kuchinke, L. *et al.* Incidental effects of emotional valence in single word processing: an fMRI study. *NeuroImage* **28**, 1022–1032 (2005).
107. Fiebach, C. J. & Friederici, A. D. Processing concrete words: fMRI evidence against a specific right-hemisphere involvement. *Neuropsychologia* **42**, 62–70 (2004).
108. 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).

109. 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).
110. Yarkoni, T., Speer, N. K., Balota, D. A., McAvoy, M. P. & Zacks, J. M. Pictures of a thousand words: investigating the neural mechanisms of reading with extremely rapid event-related fMRI. *NeuroImage* **42**, 973–987 (2008).
111. Posner, J. *et al.* The neurophysiological bases of emotion: An fMRI study of the affective circumplex using emotion-denoting words. *Hum. Brain Mapp.* **30**, 883–895 (2008).
112. Desai, R. H., Choi, W., Lai, V. T. & Henderson, J. M. Toward Semantics in the Wild: Activation to Manipulable Nouns in Naturalistic Reading. *J. Neurosci.* **36**, 4050–4055 (2016).
113. Lau, J. H., Clark, A. & Lappin, S. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cogn. Sci.* **41**, 1202–1241 (2017).
114. Venhuizen, N. J., Crocker, M. W. & Brouwer, H. Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Process.* **56**, 229–255 (2019).
115. Hu, J., Gauthier, J., Qian, P., Wilcox, E. & Levy, R. P. A Systematic Assessment of Syntactic Generalization in Neural Language Models. Preprint at https://doi.org/10.48550/arXiv.2005.03692 (2020).
116. Warstadt, A. *et al.* BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Trans. Assoc. Comput. Linguist.* **8**, 377–392 (2020).
117. Kauf, C. *et al.* Event knowledge in large language models: the gap between the impossible and the unlikely. Preprint at https://doi.org/10.48550/arXiv.2212.01488 (2022).
118. Dasgupta, I. *et al.* Language models show human-like content effects on reasoning. Preprint at https://doi.org/10.48550/arXiv.2207.07051 (2022).
119. 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).
120. Pereira, F. *et al.* Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* **9**, 963 (2018).
121. Anderson, A. J. *et al.* An Integrated Neural Decoder of Linguistic and Experiential Meaning. *J. Neurosci.* **39**, 8969–8987 (2019).
122. 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 N. Y. N 1991* **29**, 2396–2411 (2019).
123. 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. in (2001).
124. Jack, A. I. *et al.* fMRI reveals reciprocal inhibition between social and physical cognitive domains. *NeuroImage* **66**, 385–401 (2013).
125. 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.* (2020) doi:10.1523/JNEUROSCI.2036-19.2020.
126. Wehbe, L. *et al.* Incremental Language Comprehension Difficulty Predicts Activity in the Language Network but Not the Multiple Demand Network. *Cereb. Cortex N. Y. N 1991* **31**, 4006–4023 (2021).
127. Paunov, 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).

128. Shain, C., Paunov, A., Chen, X., Lipkin, B. & Fedorenko, E. No evidence of theory of mind reasoning in the human language network. *Cereb. Cortex* bhac505 (2022) doi:10.1093/cercor/bhac505.
129. 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).
130. Redcay, E., Velnoskey, K. R. & Rowe, M. L. Perceived communicative intent in gesture and language modulates the superior temporal sulcus. *Hum. Brain Mapp.* **37**, 3444–3461 (2016).
131. Zhang, G. *et al.* A social-semantic-working-memory account for two canonical language area. Preprint at https://doi.org/10.21203/rs.3.rs-2537489/v1 (2023).
132. 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).
133. Hubel, D. H. & Wiesel, T. N. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **148**, 574–591 (1959).
134. Wehbe, L. *et al.* Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. *PLOS ONE* **9**, e112575 (2014).
135. Qian, P., Qiu, X. & Huang, X. Bridging LSTM Architecture and the Neural Dynamics during Reading. *ArXiv* (2016).
136. Jain, S. & Huth, A. G. *Incorporating Context into Language Encoding Models for fMRI*. http://biorxiv.org/lookup/doi/10.1101/327601 (2018) doi:10.1101/327601.
137. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
138. Chang, N. *et al.* BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Sci. Data* **6**, 49 (2019).
139. Ratan Murty, N. A., Bashivan, P., Abate, A., DiCarlo, J. J. & Kanwisher, N. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nat. Commun.* **12**, 5540 (2021).
140. Majaj, N. J., Hong, H., Solomon, E. A. & DiCarlo, J. J. Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *J. Neurosci. Off. J. Soc. Neurosci.* **35**, 13402–13418 (2015).
141. Seeliger, K., Sommers, R. P., Güçlü, U., Bosch, S. E. & Gerven, M. A. J. van. A large single-participant fMRI dataset for probing brain responses to naturalistic stimuli in space and time. (2021) doi:10.34973/j05g-fr58.
142. LeBel, A. *et al.* A natural language fMRI dataset for voxelwise encoding models. 2022.09.22.509104 Preprint at https://doi.org/10.1101/2022.09.22.509104 (2022).
143. Poldrack, R. A. *et al.* Long-term neural and physiological phenotyping of a single human. *Nat. Commun.* **6**, 8885 (2015).
144. Gordon, E. M. *et al.* Precision Functional Mapping of Individual Human Brains. *Neuron* **95**, 791-807.e7 (2017).
145. Naselaris, T., Allen, E. & Kay, K. Extensive sampling for complete models of individual brains. *Curr. Opin. Behav. Sci.* **40**, 45–51 (2021).
146. 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).
147. Poldrack, R. A. Diving into the deep end: a personal reflection on the MyConnectome study. *Curr. Opin. Behav. Sci.* **40**, 1–4 (2021).
148. Smith, D. M., Perez, D. C., Porter, A., Dworetsky, A. & Gratton, C. Light Through the Fog: Using Precision fMRI Data to Disentangle the Neural Substrates of Cognitive Control. *Curr. Opin. Behav. Sci.* **40**, 19–26 (2021).
149. Prince, J. S. *et al.* GLMsingle: a toolbox for improving single-trial fMRI response estimates. 2022.01.31.478431 Preprint at https://doi.org/10.1101/2022.01.31.478431 (2022).

150. DiMattina, C. & Zhang, K. Adaptive stimulus optimization for sensory systems neuroscience. *Front. Neural Circuits* **7**, (2013).
151. Pallier, C. & Devauchelle, A.-D. Cortical representation of the constituent structure of sentences. *Proc. Natl. Acad. Sci.* **108**, (2011).
152. Blank, I. A. & Fedorenko, E. No evidence for differences among language regions in their temporal receptive windows. *NeuroImage* **219**, 116925 (2020).
153. 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).
154. 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).
155. Desbordes, T. *et al.* Dimensionality and ramping: Signatures of sentence integration in the dynamics of brains and deep language models. 2023.02.28.530443 Preprint at https://doi.org/10.1101/2023.02.28.530443 (2023).
156. Regev, T. I. *et al.* Neural populations in the language network differ in the size of their temporal receptive windows. 2022.12.30.522216 Preprint at https://doi.org/10.1101/2022.12.30.522216 (2023).
157. Kim, B. *et al.* Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). Preprint at https://doi.org/10.48550/arXiv.1711.11279 (2018).
158. Li, C. *et al.* Optimus: Organizing Sentences via Pre-trained Modeling of a Latent Space. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 4678–4699 (Association for Computational Linguistics, 2020). doi:10.18653/v1/2020.emnlp-main.378.
159. Wehbe, L. *et al.* BOLD predictions: Automated simulation of fMRI experiments. *NeurIPS Demonstr. Track* (2016).
160. Jain, S., Vo, V. A., Wehbe, L. & Huth, A. G. Computational Language Modeling and the Promise of in Silico Experimentation. *Neurobiol. Lang.* 1–27 (2023) doi:10.1162/nol_a_00101.
161. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
162. Lenth, R. V. emmeans: Estimated Marginal Means, aka Least-Squares Means. (2023).
163. Hasson, U., Yang, E., Vallines, I., Heeger, D. J. & Rubin, N. A Hierarchy of Temporal Receptive Windows in Human Cortex. *J. Neurosci.* **28**, 2539–2550 (2008).
164. Shain, C. *et al.* 'Constituent length'effects in fMRI do not provide evidence for abstract syntactic processing. *bioRxiv* (2021).
165. von Luxburg, U. A Tutorial on Spectral Clustering. Preprint at https://doi.org/10.48550/arXiv.0711.0189 (2007).
166. Pennington, J., Socher, R. & Manning, C. Glove: Global Vectors for Word Representation. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1532–1543 (Association for Computational Linguistics, 2014). doi:10.3115/v1/D14-1162.
167. Zhu, Y. *et al.* Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. Preprint at https://doi.org/10.48550/arXiv.1506.06724 (2015).
168. Bird, S. & Loper, E. NLTK: The Natural Language Toolkit. in *Proceedings of the ACL Interactive Poster and Demonstration Sessions* 214–217 (Association for Computational Linguistics, 2004).
169. Paul, D. B. & Baker, J. M. The design for the Wall Street Journal-based CSR corpus. in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992* (1992).

170. Gozdz-Roszkowski, S. *Patterns of Linguistic Variation in American Legal English*. (2011).
171. Martínez, E., Mollica, F. & Gibson, E. Poor writing, not specialized concepts, drives processing difficulty in legal language. *Cognition* **224**, 105070 (2022).
172. Cohen, K. B. *et al.* Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles. *BMC Bioinformatics* **18**, 372 (2017).
173. Raffel, C. *et al.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Preprint at https://doi.org/10.48550/arXiv.1910.10683 (2020).
174. Danescu-Niculescu-Mizil, C. & Lee, L. Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs. in *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics* 76–87 (Association for Computational Linguistics, 2011).
175. Davies, M. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Lit. Linguist. Comput.* **25**, 447–464 (2010).
176. Oldfield, R. C. The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* **9**, 97–113 (1971).
177. Nieto-Castanon, A. *Handbook of functional connectivity Magnetic Resonance Imaging methods in CONN*. (2020). doi:10.56441/hilbertpress.2207.6598.
178. Ashburner, J. & Friston, K. J. Unified segmentation. *NeuroImage* **26**, 839–851 (2005).
179. Rokem, A. & Kay, K. Fractional ridge regression: a fast, interpretable reparameterization of ridge regression. *GigaScience* **9**, giaa133 (2020).
180. Fedorenko, E. The early origins and the growing popularity of the individual-subject analytic approach in human neuroscience. *Curr. Opin. Behav. Sci.* **40**, 105–112 (2021).
181. Frost, M. A. & Goebel, R. Measuring structural-functional correspondence: spatial variability of specialised brain regions after macro-anatomical alignment. *NeuroImage* **59**, 1369–1381 (2012).
182. Tahmasebi, A. M. *et al.* Creating probabilistic maps of the face network in the adolescent brain: A multicentre functional MRI study. *Hum. Brain Mapp.* **33**, 938–957 (2012).
183. Vázquez-Rodríguez, B. *et al.* Gradients of structure–function tethering across neocortex. *Proc. Natl. Acad. Sci.* **116**, 21219–21227 (2019).
184. Fedorenko, E., Duncan, J. & Kanwisher, N. Language-Selective and Domain-General Regions Lie Side by Side within Broca's Area. *Curr. Biol.* **22**, 2059–2062 (2012).
185. Julian, J. B., Fedorenko, E., Webster, J. & Kanwisher, N. An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage* **60**, 2357–2364 (2012).
186. 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).
187. Fedorenko, E. & Thompson-Schill, S. L. Reworking the language network. *Trends Cogn. Sci.* **18**, 120–126 (2014).
188. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *NeuroImage* **56**, 400–410 (2011).
189. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
190. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
191. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
192. Wolf, T. *et al.* Transformers: state-of-the-art natural language processing. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 38–45 (Association for Computational Linguistics, 2020). doi:10.18653/v1/2020.emnlp-demos.6.

193. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55–67 (1970).
194. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
195. Hale, J. A Probabilistic Earley Parser as a Psycholinguistic Model. in *Second Meeting of the North American Chapter of the Association for Computational Linguistics* (2001).
196. Bergmann, A. Statistical Parametric Mapping The Analysis Of Functional Brain Images. in (2016).
197. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage* **9**, 179–194 (1999).

Extended Methods

Extended Methods Overview

1. Experimental materials 36
1.1. *Baseline* materials 36
1.2. *Drive* and *suppress* materials 38

2. fMRI experiments 40
2.1. Participants 40
2.2. Critical fMRI tasks 40
2.3. fMRI data acquisition, preprocessing and first-level analysis 42
2.4. Definition of language regions of interest (ROIs) 44
2.5. Definition of control regions of interest (ROIs) 45
2.6. Aggregation of voxels within each region of interest (ROI) 46

3. Modeling of brain data using large language models (LLMs) 46
3.1. General modeling approach 46
3.2. Encoding model 46
3.3. Sentence representations from large language models (LLMs) 47

4. Sentence properties 48
4.1. Surprisal features 48
4.2. Behavioral norms 48

5. Surface visualizations 49

1. Experimental materials

1.1. *Baseline* materials

We selected a large sample (n=1,000) of naturally occurring sentences in order to probe brain responses to diverse linguistic stimuli. (Note that although not every stimulus constituted a sentence by some definitions, we use the term 'sentence' throughout for convenience.) As detailed below, this set consists of two subsets: the first subset (n=534 sentences) was selected to maximize semantic diversity, and the second subset (n=466 sentences) was selected via random sampling across diverse genres and styles. Across the two subsets, sentences were drawn from nine corpora. All corpora were filtered to only include 6-word sentences with only printable ASCII characters and had a letter as the first character. We chose to use sentences of fixed length because the number of words in a sentence strongly modulates the response magnitude of the language areas (unpublished data from the Fedorenko lab). And we chose the length of 6 words because a) we wanted to maximize the number of sentences that could be presented, and a larger number is feasible with shorter stimuli; b) the temporal receptive window (e.g., [163,81]) of the language system appears to be ~6 words (e.g., [151,164,156]).
The corpora were preprocessed to remove repeated/leading/trailing whitespace, strip whitespace before common punctuation characters (?.!,:;), append a final period if the last sentence character was not a punctuation character, and uppercase the first letter.

1.1.1. Subset 1: Maximizing semantic diversity

To select a semantically diverse set of sentences, we made use of the 180 semantic clusters identified by Pereira et al. [120] (publicly available at https://osf.io/crwz7/wiki/home/). These clusters were derived by performing spectral clustering [165] on GloVe co-occurrence vectors [166] for n=29,805 English words [102] and spanned diverse semantic categories, including some that corresponded to classic concrete concept categories (like 'body' or 'food') and others that were more abstract (like 'soul' or 'argument').
To select the sentences for this subset, we used a large collection of amateur fiction (The Toronto Book Corpus [167]) from three genres: Adventure (n=62,661 sentences after filtering, as described above), Fantasy (n=488,841 sentences), and Mystery (n=177,523 sentences).

For each of the 180 clusters, we took the cluster target word (a word manually assigned as a descriptive label for each cluster in Pereira et al. [120]) and 20 most frequent cluster member words and computed their average GloVe (840B) embedding (either 20 or 21 unique words per cluster, depending on whether the cluster target word belonged to the 20 most frequent cluster member words). GloVe embeddings were available for all 3,769 unique cluster words (169 clusters x 21 words each + 11 clusters x 20 words). For each sentence from each of the three corpora (Toronto Adventure, Toronto Fantasy, Toronto Mystery), we computed the average GloVe embedding of its content words (nouns, verbs, adjectives and adverbs, as identified using the POS tagger from the NLTK Python library [168]); for the sentences that contained no content words (fewer than 0.02% of all sentences within each corpus), we used all words in the sentence. Next, we computed the pairwise cosine distance between each candidate sentence

from the three corpora and each cluster. For each corpus, we outputted the 20 sentences that were most semantically related to each cluster (i.e., had the smallest Cosine distance to the cluster's average embedding) and manually selected one sentence per corpus per cluster. The manual selection process ensured that the sentence is topically related to the target cluster and is not inappropriate/offensive. To minimize idiosyncratic biases, sentences from each of the three corpora were selected by three different people. For Toronto Adventure and Toronto Mystery, no appropriate sentences could be found for two clusters ('pleasure' and 'stupid'), thus yielding 178 sentences from those two corpora. For Toronto Fantasy, no appropriate sentences could be found for two clusters ('pleasure' and 'sexy'), thus yielding 178 sentences from that corpus. Thus, we obtained 534 sentences (178 sentences * 3 corpora) to maximize semantic diversity.

1.1.2. Subset 2: Random sampling

The remaining 466 sentences were randomly sampled from eight diverse corpora that spanned three main categories:

1) Published written text: Subset of The Wall Street Journal articles published in 1996 (n=2,997 sentences; [169]), The Brown Corpus (n=1,421 sentences; [168]), Subset of The Universal Dependencies Corpus (n=1,063 sentences; https://universaldependencies.org/), The Contract Corpus (n=264 sentences; legal texts from Goźdź-Roszkowski [170] and texts collected by Martinez, Mollica & Gibson [171]), and The Colorado Richly Annotated Full-Text (CRAFT) Corpus (n=34 sentences; biomedical articles from PubMed; [172]).
2) Web media text: Subset of The Common Crawl C4 Corpus (n=921,099 sentences; [173]).
3) Transcribed spoken text: The Cornell Movie-Dialogs Corpus (n=41,058 sentences; [174]), and a subset of the spoken section of The Corpus of Contemporary American English (COCA) (n=101,193 sentences; [175]).

The total number of sentences across these eight corpora was 1,069,129. Given that the corpora varied substantially in size (between 34 and 921,099), we implemented a stratified sampling scheme to ensure that sentences were sampled from all eight corpora. Specifically, for corpora with fewer than 0.025% of all sentences (two corpora: The Contract Corpus and The CRAFT Corpus), we sampled 5 sentences from each. For corpora with 0.1-0.3% of all sentences (three corpora: The Wall Street Journal Corpus, The Brown Corpus, The Universal Dependencies Corpus), we sampled 50 sentences each. For corpora with 3.8-9.5% of all sentences (two corpora: The Cornell Movie-Dialogs Corpus and COCA), we sampled 75 sentences each. This sampling yielded 310 sentences. The remaining 156 sentences were sampled from the remaining (largest) corpus (The Common Crawl C4 Corpus) with 86.2% of all sentences. Thus, we obtained 466 sentences from diverse corpora that spanned different genres (e.g., written vs. spoken language) and styles (e.g., formal vs. conversational language).

After collecting fMRI data from the first participant (unique ID 853; these are numbers in the lab-internal database and can be cross-referenced with the Fedorenko lab's studies on OSF), several minor edits and replacements were made to the original set of 1,000 sentences. Twenty

unique sentences were affected: for 8 sentences, we made minor changes to the punctuation, and 12 sentences were replaced (with sentences from the original sentence sources, as described above, and according to the manual filtering criteria reported in SI 1).

1.2. *Drive* and *suppress* materials

The *drive* and *suppress* materials were designed to elicit high or low fMRI activity in the language network, respectively, based on the predictions from the encoding model (Extended Methods 3.2). The *drive* and *suppress* materials were selected by searching across ~1.8M sentences to identify sentences that are predicted to elicit high or low fMRI responses (the *search* approach).

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 (SI 13).

Using each method (both are described in more detail below), we identified 250 unique *drive* and 250 unique *suppress* sentences. These 500 sentences were interspersed among the 1,000 *baseline* sentences (Extended Methods 2.2.1) in the fMRI scanning sessions (for a total of n=1,500 unique sentences collected across three scanning sessions with 500 unique sentences in each session).

1.2.1. *Search* approach

The *search* approach searches through large-scale corpora of diverse English text to identify candidate *drive* and *suppress* sentences.

1.2.1.1. Identifying candidate materials

To identify candidate *drive/suppress* sentences we searched across 1,871,413 6-word sentences (unique sentences: 1,805,623) from nine diverse large-scale text corpora. The corpora were partly overlapping with the corpora described in section Extended Methods 1.1 and were filtered and preprocessed in the same manner as the *baseline* materials. The corpora spanned three main categories (for more details, see SI 8):

1) Published written text: Subset of The Toronto Book Corpus [167] from three genres: Adventure (n=62,661 sentences after filtering, as described above), Fantasy (n=488,841 sentences), and Mystery (n=177,523 sentences), Subset of The Wall Street Journal articles published in 1996 (n=2,997 sentences; [169]), The Brown Corpus (n=1,421 sentences; [168]), Subset of The Universal Dependencies Corpus (n=1,063 sentences; https://universaldependencies.org/), The Contract Corpus (n=264 sentences; legal texts from Goźdź-Roszkowski [170] and texts collected by Martinez, Mollica & Gibson [171]), and The Colorado Richly Annotated Full-Text (CRAFT) Corpus (n=34 sentences; biomedical articles from PubMed; [172]).

2) Web media text: Subset of The Common Crawl C4 Corpus (n=878,552 sentences; [173]).

3) Transcribed spoken text: The Cornell Movie-Dialogs Corpus (n=41,058 sentences; [174]), and

the spoken section of The Corpus of Contemporary American English (COCA) (n=216,999 sentences; [175]).

We obtained LLM embeddings from GPT2-XL for each sentence in these corpora. Each corpus was partitioned into chunks of 5,000 sentences each (or fewer, if a corpus contained fewer than 5,000 sentences) due to the computational memory load of obtaining LLM embeddings for large numbers of sentences. The corpora were partitioned into a total of 395 chunks. We used our encoding model to predict the left hemisphere language network's response to each sentence in each corpus chunk. For each chunk, we stored the 10 sentences that were predicted to elicit the highest language network response out of all the sentences in that chunk. Similarly, we stored the 10 sentences that were predicted to elicit the lowest language network response. Across the 395 corpus chunks, we hence obtained 10 sentences * 2 * 395 chunks = 7,900 candidate sentences (3,950 *drive* and 3,950 *suppress* sentences).

1.2.1.2. Selecting the final sets from the candidate materials

To identify a set of 250/250 *drive* and *suppress* sentences (500 total), we followed a five-step procedure.
First, we filtered the sentences according to the automatic exclusion criteria reported in SI 1 (to ensure that the candidate sentences did not consist of e.g., exclusively numerical characters or contained excessive punctuation). The automatic exclusion criteria excluded 220 sentences out of 7,900 (213 for *drive*, 7 for *suppress*), leaving us with 7,680 sentences (3,737 for *drive*, 3,943 for *suppress*). Second, we removed duplicate sentences, which excluded 325 additional sentences (55 for *drive*, 270 for *suppress*), leaving us with 7,355 sentences (3,682 for *drive*, 3,673 for *suppress*). Third, out of these candidate 3,682 *drive* sentences and 3,673 *suppress* sentences, we selected the *drive* sentences above the 50th percentile of predicted language network response (out of 3,682 *drive* sentences) and selected the *suppress* sentences below the 50th percentile of predicted language network response (out of 3,673 for *suppress* sentences), effectively filtering out half of our candidate sentences. Hence, for *drive* sentences the percentile exclusion left us with 1,841 sentences (the 50th percentile was 0.497). For *suppress* sentences, the percentile exclusion left us with 1,836 sentences (the 50th percentile was -0.341). Fourth, we manually marked sentences for exclusion according to the manual exclusion criteria reported in SI 1 (to ensure that the candidate sentences were not e.g., inappropriate/offensive or fully in a foreign language). As a result, 838 *drive* sentences and 0 *suppress* sentences were excluded, leaving 1,003 *drive* candidate sentences and 1,836 *suppress* candidate sentences respectively. However, for the *suppress* candidate sentences, 6 sentences occurred in the *baseline set*, and these were excluded, leaving us with 1,830 *suppress* candidate sentences. Fifth and finally, to make our sentence selection independent of human judgment, we randomly sampled 250 sentences from each set (*drive* and *suppress*), leaving us with the final set of 500 *drive/suppress* sentences from the *search* approach.

2. fMRI experiments

2.1. Participants

A total of 14 neurotypical adults (9 female), aged 21 to 31 (mean 25.3; std 3), participated for payment between October 2021 and December 2022. The sample size was based on those used for previous fMRI semantic decoding experiments [119,120]. All participants had normal or corrected-to-normal vision, and no history of neurological, developmental, or language impairments. 12 participants (~86%) were right-handed, as determined by the Edinburgh handedness inventory [176] and 2 (~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 [2]. All participants were native speakers of English. Each scanning session lasted between 1 and 2 hours. All participants gave informed written consent in accordance with the requirements of the MIT's Committee on the Use of Humans as Experimental Subjects (COUHES). To err on the conservative side, no participants were excluded from the study based on data quality considerations.

2.2. Critical fMRI tasks

2.2.1. 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 vision neuroscience, e.g., the Natural Scenes Dataset [65]). 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 4s inter-stimulus interval (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 were instructed to read attentively and think about the sentence's meaning. To encourage engagement with the stimuli, prior to the session, 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 (i.e., 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 the following distribution of *baseline/drive/suppress* sentences in the three scanning sessions: 333/84/83, 333/83/84, and 334/83/83.

2.2.2. 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.

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).

As in the event-related experiment, participants were instructed to read attentively and think about the sentence's meaning. Prior to the session, 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*, *suppress* conditions; 240 per condition) across 6 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 (i.e., participants were exposed to different presentation orders of different subsets of the materials). Condition order was counterbalanced across runs and participants.

2.2.3. Memory task for the sentence-reading task

For both the event-related and blocked critical sentence-reading experiments, 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.

Participants were presented with a set of sentences, one at a time, and asked to decide for whether they had read it during the scanning session. For both the event-related and blocked experiment, 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, see Extended Methods 1.1).

The average accuracy (sum of correct responses divided by total number of responses; chance-level is 50%) was 70.42% (SD across sessions: 11.39%) for the event-related participants (n=24 sessions – responses for one session were not saved due to an error in the script), and 61.67% (SD across sessions: 10%) for the blocked participants (n=4 sessions).

2.3. fMRI data acquisition, preprocessing and first-level analysis

2.3.1. 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 (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: 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. The first 10 s of each run were excluded to allow for steady state magnetization.

2.3.2. fMRI preprocessing

fMRI data were preprocessed using SPM12 (release 7487), 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 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; [177]). Functional and structural data were independently normalized into a common space (the Montreal Neurological Institute [MNI] template; IXI549Space) using SPM12 unified segmentation and normalization procedure [178] with a reference functional image computed as the mean functional data after realignment across all timepoints 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 4th 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 FWHM Gaussian kernel.

2.3.3. First-level modeling

2.3.3.1. Event-related experiments

The critical, event-related experiment was analyzed using GLMsingle [149], a framework for obtaining accurate response estimates in quick event-related single-trial fMRI designs. Modeling 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: 1) Choice of HRF to convolve with the design matrix: an HRF is identified from a library of 20 candidate functions (derived from independent fMRI data; Allen et al., 2021) as the best fitting for each voxel separately, 2) 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 (PCA), and 3) Regularization of voxel responses: instead of an ordinary least squares (OLS) regression, GLMsingle uses fractional ridge regression [179] to model voxel responses in order to dampen the noise inflation in a standard OLS 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 blood oxygenation level dependent (BOLD) response amplitude evoked by each individual sentence trial (fixation was modeled implicitly, such that all timepoints that did not correspond to one of the conditions (sentences) were assumed to correspond to a fixation period). Data from different scanning sessions for a given participant were analyzed together. The 'sessionindicator' option in GLMsingle was used to specify how different input runs were grouped into sessions. For each voxel, the HRF which provided the best fit to the data was identified (based on the amount of variance explained). The data were modeled using a fixed number of noise regressors (5) and a fixed ridge regression fraction (0.05) (these parameters were determined empirically using an extensive joint data modeling and data evaluation framework, see SI 2).

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 betas were z-scored session-wise per voxel (see Extended Methods 6).

2.3.3.2. Blocked experiments

Blocked experiments were analyzed using standard analysis procedures using SPM12 (release 7487), CONN EvLab module (release 19b). Effects were estimated using a GLM in which the beta weights associated with each experimental condition was modeled with a boxcar function

convolved with the canonical HRF) (fixation was modeled implicitly, such that all timepoints that did not correspond to one of the conditions were assumed to correspond to a fixation period). Temporal autocorrelations in the BOLD signal timeseries were accounted for by a combination of high-pass filtering with a 128 seconds cutoff, and whitening using an AR(0.2) model (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 (ReML). In addition to experimental condition effects, the GLM design included first-order temporal derivatives for each condition (included to model variability in the HRF delays), as well as nuisance regressors to control for the effect of slow linear drifts, subject-motion parameters, and potential outlier scans on the BOLD signal.

2.4. Definition of language regions of interest (ROIs)

Language regions of interest (ROIs) were defined in individual participants using functional localization (e.g., [77,2,78,180]). This approach is crucial because many functional regions do not exhibit a consistent mapping onto macro-anatomical landmarks (e.g., [181–183]) and this variability is problematic when functionally distinct regions lie in close proximity to each other, as is the case with both frontal and temporal language areas (e.g., [184,154]; see Fedorenko and Blank[18] for discussion of this issue for 'Broca's area').
For each participant, functional ROIs (fROIs) were defined by combining two sources of information [2,185]: i) the participant's activation map for the localizer contrast of interest (t-map), and ii) 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).

2.4.1. Language network localizer task

The task used to localize the language network was a reading task contrasting sentences (e.g., THE SPEECH THAT THE POLITICIAN PREPARED WAS TOO LONG FOR THE MEETING) and lists of unconnected, pronounceable nonwords (e.g., 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. [2])). The sentences > nonwords 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 (e.g., [14]; see Fedorenko and Blank [18] 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 (e.g., [2,7,11], *inter alia*). Further, a network that corresponds closely to the localizer contrast emerges robustly from whole-brain task-free data—voxel fluctuations during rest [154].

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

2.4.2. Language network fROIs

The language fROIs were defined using the *sentences* > *nonwords* contrast from the language localizer collected in each participant's first scanning session (see e.g., Mahowald and Fedorenko [186], 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 discussion, see Fedorenko and Thompson-Schill [187]).
To define the language fROIs, each individual *sentences* > *nonwords* 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. [2] for the *sentences* > *nonwords* 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, *IFGorb*, 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 (e.g., Blank et al., 2014), 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* > *nonwords* contrast were selected (see SI 12E for number of voxels in each fROI).

2.5. Definition of control regions of interest (ROIs)

In addition to language regions, we examined i) two large-scale brain networks linked to high-level cognitive processing—the multiple demand (MD) network [66–70] and the default mode network (DMN) [71–75] which—similar to the language regions—were functionally defined using independent localizer tasks in each participant, and ii) a set of anatomical parcels [76] in an effort to cover the entire cortex (see SI 12 for details).

2.6. Aggregation of voxels within each region of interest (ROI)

The voxels belonging to each functional ROI (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, in order to minimize potential non-stationarities that exist across different scanning sessions and to equalize response units across voxels. In SI 9 and SI 11, we report fMRI data without any normalization (the key patterns of results are not affected).

On average, we extracted responses from 10 language fROIs (SD=0), 19.43 MD fROIs (SD=1.28), 12 DMN fROIs (SD=0), and 353.71 anatomical Glasser parcels (SD=10.34) across n=14 participants (in a few cases, (f)ROIs could not be extracted due to a negative t-statistic for the contrast of interest of lack of coverage in our functional acquisition sequence).

3. Modeling of brain data using large language models (LLMs)

3.1. General modeling approach

To develop an encoding model of the language network, we fitted a linear model from the representations of a large language model (LLM) to brain responses (an encoding approach; [188]). Details on this model are provided below, in the section Encoding model. The brain data that were used to fit the encoding model were the averaged LH language network's response from the n=5 *train* participants (n=1,000 *baseline* stimuli, Extended Methods 1.1). We obtained an unbiased estimate of model performance using three different approaches: i) Cross-validated predictivity performance on held-out sentences (SI 5), ii) Cross-validated predictivity performance on held-out participants within the *train* participants (SI 6), and iii) held-out prediction performance on new participants (*evaluation* participants) *and* sentences (Results; Section "Model predictions for new sentences capture most of the explainable variance in new participants").

All modeling and analysis code was written in Python (version 3.8.11), making heavy use of the *numpy* ([189]; version 1.21.2), *scipy* ([190]; version 1.7.3), *scikit-learn* ([191]; version 0.24.2), and *transformers* ([192]; version 4.11.3) libraries.

3.2. Encoding model

To map from LLM representations to brain responses, we made use of 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 a L2-regularized ("ridge") regression model which can be seen as placing a zero-mean Gaussian prior on the regression coefficients [193]. 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:

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

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 hence *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. $\mathbf{I}$ 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 (α) 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* ([191]; version 0.24.2). Specifically, for each of 60 logarithmically-spaced α regularization parameter values (1e-30, 1e-29, …, 1e28, 1e29), 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 α regularization parameter values. We then selected the α regularization parameter that minimized this mean squared error (α = 10,000). When cross-validation was performed, the α regularization parameter was always selected using the stimuli in the train split, and with the α parameter selected, the regression model using that parameter was used on the test split.

3.3. Sentence representations from large language models (LLMs)

The primary LLM in this study was the unidirectional-attention Transformer model GPT2-XL [36], which was identified as the most brain-aligned language model in prior work [48] and which was the largest unidirectional OpenAI GPT model available on HuggingFace [192] at the time of the experiments (summer 2021)). (Supplementary analyses were performed using BERT-large (SI 5C).) We used the pretrained model[i] available via the HuggingFace library ([192], *transformers* version 4.11.3). GPT2-XL has 48 layers (i.e., 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 (i.e., 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 (i.e., earlier tokens in the sentence). Further, 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 (SI 5B). The resulting features were used as regressors in the LLM-brain comparisons. Each LLM layer (model stage for which representations were extracted, i.e., Transformer blocks) was treated as a separate set of regressors in LLM-brain comparisons. Layer 22 features were selected as regressors in the encoding model based on cross-validated model performance evaluation (SI 5A).

[i] https://huggingface.co/gpt2-xl

4. Sentence properties

4.1. 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" [194,195]. The log probability of each sentence was computed using the pre-trained unidirectional-attention language model GPT2-XL [36] from the HuggingFace library ([192], *transformers* version 4.11.3). GPT2-XL was trained on 40GB on 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 (SI 16).

4.2. Behavioral norms

To understand which properties of sentences modulate brain responses, we collected 10 behavioral rating norms to characterize the full set of sentences in this study (n=2,000 total sentences across the *baseline*, *drive*, and *suppress* sets) across 9 surveys (two of the norms were grouped into a single survey).

4.2.1. Participants

Participants were recruited using crowd-sourcing platforms: Prolific (n=8 surveys) and Amazon Mechanical Turk (mTurk; 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 mTurk, the study was restricted to "Mechanical Turk Masters" workers. Across the 9 surveys, a total of 3,600 participants took part in the experiment (400 participants for each survey; see SI 17C for details). 2,741 participants remained after pre-defined exclusion criteria (SI 17A). The experiments were conducted with approval from and in accordance with the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology. Participants gave informed consent before starting each experiment and were compensated for their time (minimum $12/hour).

4.2.2. 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. Then they 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 is 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 asked participants to

evaluate the sentence's grammatical well-formedness and plausibility. Three surveys asked participants to rate the sentence for the presence of particular kinds of content (mental states, physical objects and their interactions, and places). Two surveys asked participants to rate the sentence's emotional dimensions: valence (how positive the content is) and arousal. One survey asked for ratings of how easy the sentence is to visualize. Finally, two surveys asked participants to rate the sentences for how common they perceived them: in general, or specifically in conversational settings (see SI 17C for the details of all the instructions). The first survey (with two questions per sentence) took 25.01 minutes on average; the remaining surveys took 14.25 min 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 6 sentence preambles (e.g., "*When I was younger, I would often …*"; see SI 17B for the full set), which were used post-hoc to evaluate English proficiency.

5. Surface visualizations

Data were analyzed in the volumetric MNI IXI549Space (SPM12; [196]). For surface projections, data were registered to FreeSurfer's CVS35 (combined volumetric and surface-based (CVS)) in the MNI152 space using *mri_vol2surf* in FreeSurfer v7.3.2 [197] with a projection distance of 1.5mm and otherwise default parameters. The resulting surfaces were visualized using Freeview.