The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing — **Source link** ↗

Martin Schrimpf, Martin Schrimpf, Idan Blank, Idan Blank ...+8 more authors

Institutions: McGovern Institute for Brain Research, Massachusetts Institute of Technology, University of California, Los Angeles

Related papers:

- Natural speech reveals the semantic maps that tile human cerebral cortex

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

- Learning Neural Representations of Human Cognition across Many fMRI Studies

- Integrated accounts of behavioral and neuroimaging data using flexible recurrent neural network models

- From fMRI Data To Cognitive Models: Testing the ACT-R Brain Mapping Hypothesis with an Ex-Post Model

Share this paper:

View more about this paper here: https://typeset.io/papers/the-neural-architecture-of-language-integrative-reverse-1xm7merzzp

The neural architecture of language:
Integrative reverse-engineering converges on
a model for predictive processing

Martin Schrimpf[1,2,3], **Idan Blank**[*,1,4], **Greta Tuckute**[*,1,5], **Carina Kauf**[*,1], **Eghbal A. Hosseini**[1],

Nancy Kanwisher[1,2,3], **Joshua Tenenbaum**[†,1,3], **Evelina Fedorenko**[†,1,2]

1 Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA
2 McGovern Institute for Brain Research, MIT, Cambridge, MA, USA
3 Center for Brains, Minds and Machines, MIT, Cambridge, MA, USA
4 Psychology Department, UCLA, Los Angeles, CA, USA
5 Media Lab, MIT, Cambridge, MA, USA

1

Abstract

The neuroscience of perception has recently been revolutionized with an integrative reverse-engineering approach in which computation, brain function, and behavior are linked across many different datasets and many computational models. We here present a first systematic study taking this approach into higher-level cognition: human language processing, our species' signature cognitive skill. We find that the most powerful 'transformer' networks predict neural responses at nearly 100% and generalize across different datasets and data types (fMRI, ECoG). Across models, significant correlations are observed among all three metrics of performance: neural fit, fit to behavioral responses, and accuracy on the next-word prediction task (but not other language tasks), consistent with the long-standing hypothesis that the brain's language system is optimized for predictive processing. Model architectures with initial weights further perform surprisingly similar to final trained models, suggesting that inherent structure – and not just experience with language – crucially contributes to a model's match to the brain.

*,† joint second/senior authors

A core goal of neuroscience is to decipher from patterns of neural activity the algorithms underlying our abilities to perceive, think about, and act in the world. Recently, a new "reverse engineering" approach to computational modeling in systems neuroscience has transformed our algorithmic understanding of the ventral stream in primate vision (Bao et al., 2020; Cadena et al., 2019; Cichy et al., 2016; Kietzmann et al., 2019; Kubilius et al., 2019; Schrimpf et al., 2018, 2020; Yamins et al., 2014), and holds great promise for application to other aspects of brain function. This approach has been enabled by a breakthrough in artificial intelligence (AI): the engineering of artificial neural network (ANN) systems that perform core perceptual tasks with unprecedented accuracy, approaching human levels, and that do so using computational machinery that is abstractly similar to biological neurons. In the ventral visual stream, the key AI developments come from deep convolutional neural networks (DCNNs) that perform visual object recognition from natural images (Ciregan et al., 2012; Krizhevsky et al., 2012; Schrimpf et al., 2018, 2020; Yamins et al., 2014), which is widely thought to be the primary function of this pathway. Leading DCNNs for object recognition have now been shown to predict the responses of neural populations in multiple stages of the ventral stream (V1, V2, V4, IT), in both macaque and human brains, approaching the noise ceiling of the data. Thus, although far from perfect models, DCNNs could provide the basis for a first complete account of how the brain computes object percepts from visual images.

Inspired by this success story, analogous ANN models are now regularly applied to other domains of sensation and perception (Kell et al., 2018; Zhuang et al., 2017). Could these models also let us reverse-engineer the brain mechanisms of higher-level human cognition? Here we show for the first time how the reverse-engineering approach pioneered in the ventral stream can be applied to a higher-level cognitive domain that plays an essential role in human mental life: language processing, or the extraction of meaning from spoken or written phrases, sentences, and stories. Cognitive scientists have for decades treated neural network models with skepticism (Marcus, 2018; Pinker & Prince, 1988), as these systems lacked the capacity for explicit symbolic representation, a core feature of language, and thinking and reasoning more generally. Recent ANN models of language in AI, however, have proven capable of at least approximating some aspects of symbolic computation, and have achieved remarkable success on a wide range of applied natural language processing (NLP) tasks. The results presented here, based on this new generation of ANNs, suggest that a computationally adequate model of language processing in the brain may be closer than previously thought.

Because we build on the same logic in our analysis of language in the brain, it is helpful to review why the neural network-based reverse engineering approach has proven so compelling in the study of object recognition in the ventral stream. Crucially, our ability to robustly link computation, brain function, and behavior is supported not by testing a single model on a single dataset or a single kind of data, but by large-scale *integrative benchmarking* (Schrimpf et al., 2020) that establishes consistent patterns of performance across many different ANNs applied to multiple neural and behavioral datasets, together with their performance on the proposed core computational function of the brain system under study. Given the complexities of the brain's structure and the functions it performs, we know that any one of these models is surely oversimplified and ultimately wrong – at best just an approximation of some aspects of what the brain might do. But some models are less wrong, and consistent trends in performance across many models can reveal insights that go substantially beyond what any one model can tell us.
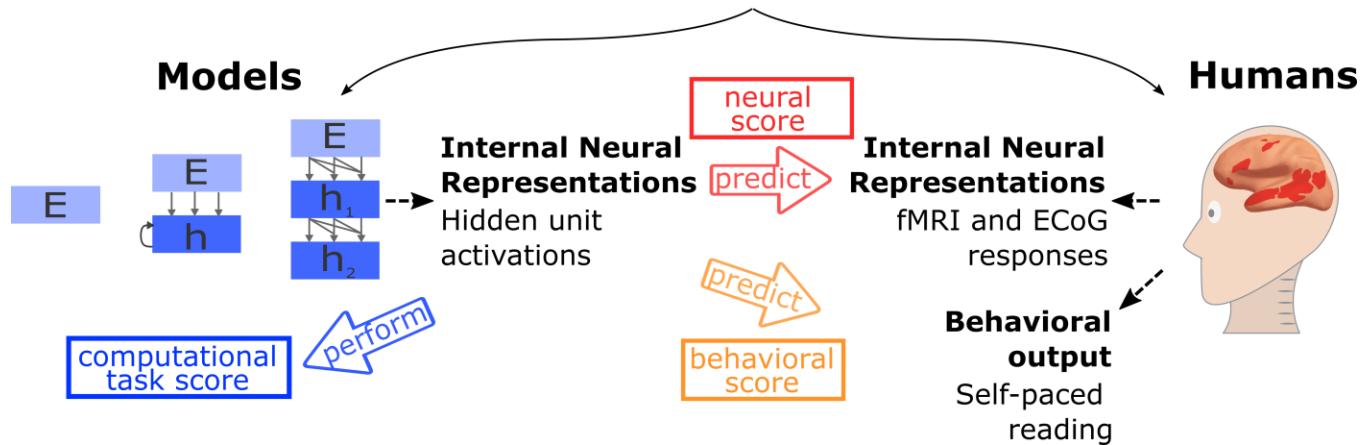
Figure 1: **Comparing Artificial Neural Network models of language processing to human language processing.** We tested how well different models predict measurements of human neural activity (fMRI and ECoG) and behavior during language comprehension. The candidate models ranged from simple embedding models to more complex recurrent and transformer networks. Stimuli ranged from sentences to passages to stories and were 1) fed into the models, and 2) presented to human participants (visually or auditorily). Models' internal representations were evaluated on three major dimensions: their ability to predict human neural representations; their ability to predict human behavior in the form of reading times; their ability to perform computational tasks such as next-word prediction. We establish consistent relationships between these measures across many different models – these trends of performance reveal insights beyond what a single model can tell us.

In the ventral stream specifically, our understanding that computations underlying object recognition are analogous to the structure and function of DCNNs is supported by findings that across hundreds of model variants, DCNNs that perform better on object recognition tasks also better capture human recognition behavior and neural responses in IT cortex of both human and non-human primates (Rajalingham et al., 2018; Schrimpf et al., 2018, 2020; Yamins et al., 2014). This integrative benchmarking reveals a rich pattern of correlations among three classes of performance measures — (i) accuracy on the core object recognition task, (ii) accuracy in predicting hits and misses in human object recognition behavior, or human object similarity judgments, and (iii) neural variance explained, in IT neurophysiology or fMRI responses — such that for any DCNN model we can predict how well it scores on each of these measures from the other measures. This pattern of results was not assembled in a single paper but in multiple papers across several labs and several years of work. Taken together, they provide strong evidence that the ventral stream supports primate object recognition through something like a deep convolutional feature hierarchy, the exact details of which are being modeled more and more precisely.

Here we describe an analogous pattern of results for ANN models of human language, establishing a link between transformer-based ANN architectures that have revolutionized natural language processing in AI systems over the last two years, and fundamental computations of human language processing as revealed through both neural and behavioral measures. Language comprehension is a quintessentially human ability, bridging perception and high-level reasoning, and forming the foundation of human culture. The processing of language is known to depend causally on a left-lateralized fronto-temporal brain network (Bates et al., 2003; Binder et al., 1997; Fedorenko & Thompson-Schill, 2014; Friederici, 2012; Gorno-Tempini et al., 2004; Hagoort, 2019; Price, 2010) (Fig. 1) that responds robustly and selectively to linguistic input (Fedorenko et al., 2011; Monti et al., 2012), whether auditory or visual (Deniz et al., 2019; Regev et al., 2013). Yet the precise computations underlying language processing in the brain remain unknown. Computational models of sentence processing have previously been used to explain both behavioral (Dotlačil, 2018; Futrell, Gibson, & Levy, 2020; Gibson, 1998; Gibson et al., 2013; Hale, 2001; Jurafsky, 1996; Lakretz et al., 2020; Levy, 2008a, 2008b; Lewis et al., 2006; McDonald

76 & Macwhinney, 1998; Smith & Levy, 2013; Spivey-Knowlton, 1996; Steedman, 2000; van Schijndel et al., 2013), and neural
77 responses to linguistic input (Brennan et al., 2016; Brennan & Pylkkänen, 2017; Ding et al., 2015; Frank et al., 2015;
78 Henderson et al., 2016; Huth et al., 2016; Lopopolo et al., 2017; Lyu et al., 2019; T. M. Mitchell et al., 2008; Nelson et al.,
79 2017; Pallier et al., 2011; Pereira et al., 2018; Rabovsky et al., 2018; Shain et al., 2020; Wehbe et al., 2014; Willems et al.,
80 2016; Gauthier & Ivanova, 2018; Gauthier & Levy, 2019; Hu et al., 2020; Jain & Huth, 2018; S. Wang et al., 2020; Schwartz et
81 al., 2019; Toneva & Wehbe, 2019). However, prior studies have not attempted any of the large-scale integrative
82 benchmarking that has proven so valuable in understanding vision in the ventral stream; they typically test just one or a
83 small number of models against a single dataset, and the same models are rarely evaluated on all three metrics of neural,
84 behavioral, and objective task performance. Previous models have also left much of the variance in human neural data
85 unexplained, and most do not have sufficient capacity to solve the full linguistic problem that the brain solves – to form a
86 representation of sentence meaning capable of performing a broad range of real-world language tasks on diverse natural
87 linguistic input. We are thus left with a collection of suggestive results but no clear sense of how close neural models are to
88 fully explaining language processing in the brain, or what features are likely to transcend the substantial inadequacies of any
89 one model.

90 Our goal here is to present a first systematic integrative-benchmarking reverse engineering study of language in the brain,
91 at the scale necessary to discover robust relationships between neural and behavioral measurements from humans, and
92 performance of models on language tasks. We seek to determine not just which model fits empirical data best, but what
93 dimensions of variation across models are correlated with fit to human data. This requires testing a broad suite of ANN
94 architectures with sufficient variance on all three kinds of measures (fit to neural and behavioral data, and model
95 performance). This approach has not been applied in the study of language or any other higher cognitive system, and even
96 in perception has not been attempted within a single integrated study. Thus, we view our work more generally as a
97 template for how to apply the integrative reverse-engineering approach to a novel perceptual or cognitive system.

98 Specifically, we examined the relationships between 43 diverse state-of-the-art ANN language models (henceforth 'models')
99 across three neural language comprehension datasets (two fMRI, one electrocorticography (ECoG)), as well as behavioral
100 signatures of human language processing in the form of self-paced reading times, and a range of linguistic functions
101 assessed via standard engineering tasks from NLP. The models spanned all major classes of existing ANN language
102 approaches and included simple embedding models (e.g., GloVe (Pennington et al., 2014)), more complex recurrent neural
103 networks (e.g., LM1B (Jozefowicz et al., 2016)), and many variants of transformers or attention-based architectures—
104 including both 'unidirectional-attention' models (trained to predict the next word given the previous words; e.g., GPT
105 (Radford et al., 2019)) and 'bidirectional-attention' models (trained to predict a missing word given the surrounding context;
106 e.g., BERT (Devlin et al., 2018)). Our integrative approach yields four major findings. (1) Models' relative fit to neural data
107 ("neural predictivity") generalizes across different datasets and data types (fMRI, ECoG), and certain architectural features
108 consistently lead to more brain-like models: transformer-based models perform better than recurrent networks or word-
109 level embedding models, and larger-capacity models perform better than smaller models. (2) The best models explain
110 nearly 100% of the explainable variance (up to the noise ceiling) in neural data. This result stands in stark contrast to earlier
111 generations of models that have typically accounted for at most 30-50% of the predictable neural signal. (3) Across models,
112 there are significant correlations among all three metrics of model performance: neural fit, fit to reading time in behavior,
113 and model accuracy on the next-word prediction task; no other linguistic task was predictive of models' fit to neural or
114 behavioral data. These findings provide the strongest evidence to date for a classic hypothesis about the computations
115 underlying human language understanding, that the brain's language system is optimized to extract meaning through
116 predictive processing. (4) Models initialized with random weights (prior to training) perform surprisingly similarly in neural
117 predictivity to final trained models, which suggests that network architecture contributes as much or more than experience-
118 dependent learning to a model's match to the brain. In particular, one architecture introduced just in 2019, the generative
119 pre-trained transformer (GPT-2), consistently outperforms all other models and explains almost all variance in both fMRI
120 and ECoG data from sentence processing tasks. GPT-2 is also arguably the most cognitively plausible of the transformer
121 models (because it uses unidirectional, forward attention), and performs best overall as an AI system when considering
122 both natural language understanding and natural language generation tasks. Thus contemporary AI appears to be rapidly
123 converging on architectures that might capture language processing, at least up to the sentence level, in the human mind
124 and brain.

4

125

Results

126

127 We evaluated a broad range of state-of-the-art ANN models on the match of their internal representations to three human
128 neural datasets. The models spanned all major classes of existing language models (Methods 5, Table S10). The neural
129 datasets consisted of i) fMRI activations while participants read short passages, presented one sentence at a time (across
130 two experiments) that spanned diverse topics (*Pereira2018* dataset (Pereira et al., 2018)); ii) ECoG recordings while
131 participants read semantically and syntactically diverse sentences, presented one word at a time (*Fedorenko2016* dataset
132 (Fedorenko et al., 2016)); and iii) fMRI BOLD signal time-series elicited while participants listened to few-minutes-long
133 naturalistic stories (*Blank2014* dataset (I. Blank et al., 2014)) (Methods 1-3). Thus, the datasets varied in the method
134 (fMRI/ECoG), the nature and grain of linguistic units to which responses were recorded (sentences/words/2s-long story
135 fragments), and modality (reading/listening). In most analyses, we consider the overall results across the three neural
136 datasets; when considering the results for the individual neural datasets, we give the most weight to *Pereira2018* because
137 it includes multiple repetitions per stimulus (sentence) within each participant and quantitatively exhibits the highest
138 internal reliability (Fig. S1). Because our research questions concern language processing, we extracted neural responses
139 from language-selective voxels or electrodes that were functionally identified by an extensively validated independent
140 "localizer" task that contrasts reading sentences versus nonword sequences (Fedorenko et al., 2010). This localizer robustly
141 identifies the fronto-temporal language-selective network (Methods 1-3, Fig. 2b, S3).

142 To compare a given model to a given dataset, we presented the same stimuli to the model that were presented to humans
143 in neural recording experiments and 'recorded' the model's internal activations (Methods 5-6, Fig. 1). We then tested how
144 well the model recordings could predict the neural recordings for the same stimuli, using a method originally developed for
145 studying visual object recognition (Schrimpf et al., 2018; Yamins et al., 2014). Specifically, using a subset of the stimuli, we
146 fit a linear regression from the model activations to the corresponding human measurements, modeling the response of
147 each voxel (*Pereira2018*) / electrode (*Fedorenko2016*) / region (*Blank2014*) as a linear weighted sum of responses of
148 different units from the model. We then computed model predictions by applying the learned regression weights to model
149 activations for the held-out stimuli, and evaluated how well those predictions matched the corresponding held-out human
150 measurements by computing Pearson's correlation coefficient. We further normalized these correlations by the
151 extrapolated reliability of the particular dataset, which places an upper bound ("ceiling") on the correlation between the
152 neural measurements and any external predictor (Methods 7, Fig. S1). The final measure of a model's performance
153 ('predictivity' or 'score') on a dataset is thus Pearson's correlation between model predictions and neural recordings
154 divided by the estimated ceiling and averaged across voxels/electrodes/regions and participants. We report the score for
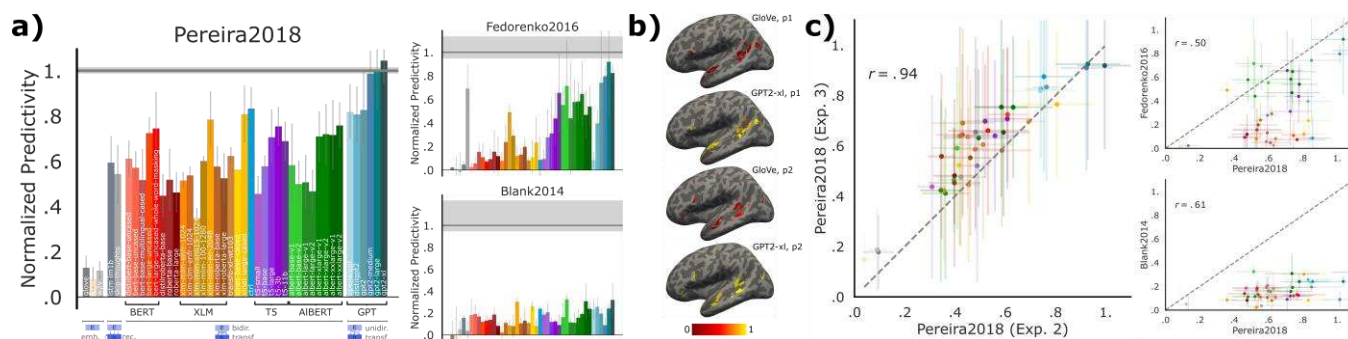155 the best-performing layer of each model (Methods 6, Fig. S10).

Figure 2: **Specific models accurately predict neural responses consistently across datasets. (a)** We compared 43 computational models of language processing (ranging from embedding to recurrent and bi- and uni-directional transformer models) in their ability to predict human brain data. The neural datasets include: fMRI voxel responses to visually presented sentences (*Pereira2018*), ECoG electrode responses to visually presented (word-by-word) sentences (*Fedorenko2016*), fMRI ROI responses to ~5min-long stories (*Blank2014*). For each model, we plot the normalized predictivity, i.e. the fraction of ceiling (gray line; Methods 7, Fig. S1) the model can predict. Ceiling levels are .32 (*Pereira2018*), .17 (*Fedorenko2016*), and .20 (*Blank2014*). Model classes are grouped by color (Methods 5, Table S10). Error bars (here and elsewhere) represent m.a.d. over subject scores. **(b)** Normalized predictivity of GloVe (a low-performing embedding model) and GPT2-xl (a high-performing transformer model) in the language-responsive voxels in the left hemisphere of two representative participants from *Pereira2018* (also Fig. S3). **(c)** To test how well model scores generalize across datasets, we correlated: two experiments with different stimuli (and some participant overlap) in *Pereira2018* (very strong correlation), and *Pereira2018* model scores with the scores for each of *Fedorenko2016* and *Blank2014* (lower but still highly significant correlations). Scores overall thus tend to generalize across datasets, although differences between datasets exist which warrant the full suite of datasets.

156 **Specific models accurately predict human brain activity.** We found (Fig. 2a-b) that specific models predict *Pereira2018* and
157 *Fedorenko2016* datasets with up to 100% predictivity (see Fig. S2 for generalization to another metric) relative to the noise
158 ceiling (Methods 7, Fig. S1). The *Blank2014* dataset is also reliably predicted, but with lower predictivity. Models vary
159 substantially in their ability to predict neural data. Generally, embedding models such as GloVe do not perform well on any
160 dataset. In contrast, recurrent networks such as skip-thoughts, as well as transformers such as BERT, predict large portions
161 of the data. The model that predicts the human data best across datasets is GPT2-xl, a unidirectional-attention transformer
162 model, which predicts *Pereira2018* and *Fedorenko2016* at close to 100% and is among the highest-performing models on
163 *Blank2014* with 32% predictivity. These scores are higher in the language network than other parts of the brain (SI-4).
164 **Model scores are consistent across experiments/datasets.** To test the generality of the model representations, we examined the
165 consistency of model scores across datasets. Indeed, if a model does well on one dataset, it tends to also do well on other
166 datasets (Fig. 2c), ruling out the possibility that we are picking up on spurious, dataset-idiosyncratic predictivity, and
167 suggesting that the models' internal representations are general enough to capture brain responses to diverse linguistic
168 materials presented visually or auditorily, and across three independent sets of participants. Specifically, model scores
169 across the two experiments in *Pereira2018* (overlapping sets of participants) correlate at $r=.94$ (Pearson here and
170 elsewhere, $p<<.00001$), scores from *Pereira2018* and *Fedorenko2016* correlate at $r=.50$ ($p<.001$), and from *Pereira2018* and
171 *Blank2014* at $r=.63$ ($p<.0001$).
172
173 **Next-word-prediction task performance selectively predicts neural scores.** In the critical test of which computations might
174 underlie human language understanding, we examined the relationship between the models' ability to predict an upcoming
175 word and their brain predictivity. Words from the Wikitext-2 dataset (Merity et al., 2016) were sequentially fed into the
176 candidate models. We then fit a linear classifier (over words in the vocabulary; n=50k) from the last layer's feature
177 representation on the training set to predict the next word, and evaluated performance on the held-out test set
178 (Methods 8). Indeed, next-word-prediction task performance robustly predicts neural scores (Fig. 3a; $r=.45$, $p<.01$,
179 averaged across datasets). The best language model, GPT2-xl, also achieves the highest neural predictivity (see previous
180 section). This relationship holds for model variants within each class—embedding models, recurrent networks, and
181 transformers—ruling out the possibility that this correlation is simply due to between-class differences in next-word-
182 prediction performance.
183 To test whether next-word prediction is special in this respect, we asked whether model performance on *any* language task
184 correlates with neural predictivity. Focusing on the high-performing, transformer models, we found that performance on
185 tasks from the GLUE benchmark collection (Cer et al., 2018; Dolan & Brockett, 2005; Levesque et al., 2012; Rajpurkar et al.,
186 2016; Socher et al., 2013; A. Wang, Singh, et al., 2019; Warstadt et al., 2019; Williams et al., 2018)—including
187 grammaticality judgments, sentence similarity judgments, and entailment—do *not* correlate with neural predictivity, in
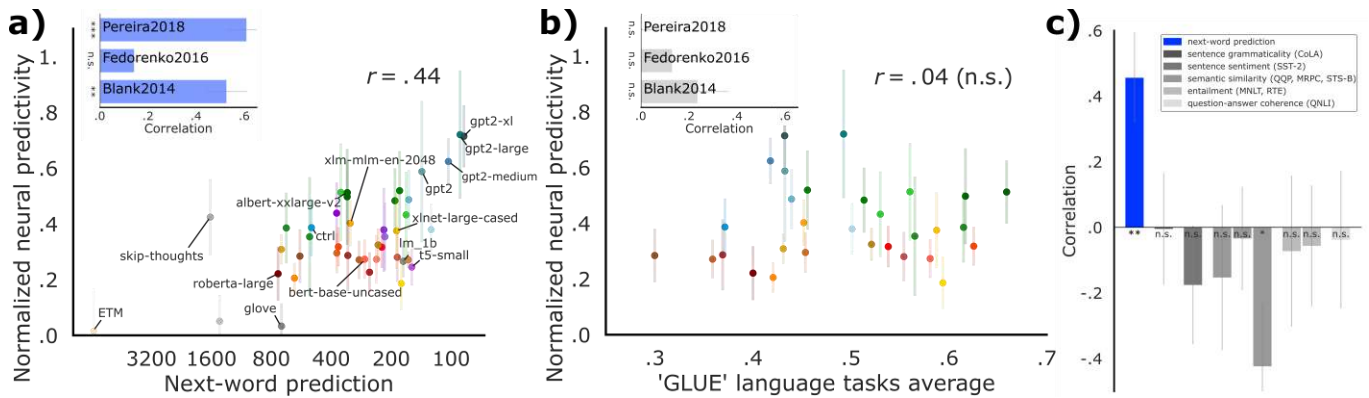
Figure 3: **Model performance on a next-word-prediction task selectively predicts neural scores. a)** Next-word-prediction task performance was evaluated as the surprisal between the predicted and true next word in the WikiText-2 dataset of 720 Wikipedia articles, or *perplexity* (x-axis, lower is better). Next-word-prediction task scores strongly predict neural scores across datasets (inset: this correlation is significant for two individual datasets: *Pereira2018* and *Blank2014*; the correlation for *Fedorenko2016* is also positive but not significant). **b)** Performance on diverse language tasks from the GLUE benchmark collection does *not* correlate with overall or individual (inset; SI-5) neural predictivity. **c)** Correlations of individual tasks with neural predictivity scores. Only improvements on next-word prediction lead to improved neural predictivity.

188 spite of eliciting variable performance across models (Fig. 3b-c). The difference in the strength of correlation between
189 neural data and the next-word prediction task vs. the GLUE tasks is highly reliable (p<<0.00001). This result suggests that
190 optimizing for predictive representations may be a critical shared objective of biological and artificial neural networks for
191 language, and perhaps more generally (Keller and Mrsic-Flogel, 2018; Singer et al., 2018).
192

193 **Neural predictivity and next-word-prediction task performance correlate with behavioral predictivity.** Beyond internal neural
194 representations, we tested the models' ability to predict external behavioral outputs because, ultimately in integrative
195 benchmarking, we strive for a computationally precise account of language processing that can explain both neural
196 response patterns and observable linguistic behaviors. We chose a large corpus (n=180 participants) of self-paced reading
197 times for naturalistic story materials (*Futrell2018* dataset (Futrell, Gibson, Tily, et al., 2020)). Per-word reading times
198 provide a theory-neutral measure of incremental comprehension difficulty, which has long been a cornerstone of
199 psycholinguistic research in testing theories of sentence comprehension (Demberg & Keller, 2008; Gibson, 1998; Just &
200 Carpenter, 1980; D. C. Mitchell, 1984; Rayner, 1978; Smith & Levy, 2013).

201 **Specific models accurately predict human reading times.** We regressed each model's last layer's feature representation against
202 reading times and evaluated predictivity on held-out words. As with the neural datasets, we observed a spread of model
203 ability to capture human behavioral data, with models such as GPT2-xl, skip-thoughts, and AlBERT-xxlarge predicting these
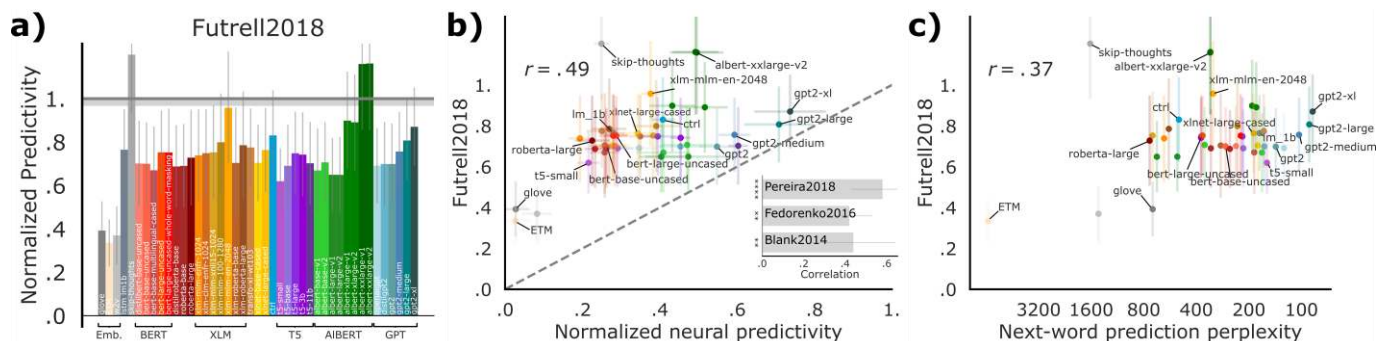
Figure 4: **Behavioral predictivity, neural predictivity, and next-word-prediction task performance are pairwise correlated**. **(a)** Behavioral predictivity of each model on *Futrell2018* human reading times (notation similar to Fig. 2). Ceiling level is .78. **(b)** Models' neural predictivity aggregated across the three neural datasets (or for each dataset individiually; inset and Fig. 6) correlates with behavioral predictivity. **(c)** Next-word-prediction task performance (Fig. 3) correlates with behavioral predictivity.

204   data close to or at ceiling (Fig. 4a; also Merkx & Frank, 2020; Wilcox et al., 2020).

205   **Neural predictivity correlates with behavioral predictivity.** To test whether models with the highest neural scores also predict
206   reading times best, we compared models' neural predictivity (across datasets) with those same models' behavioral
207   predictivity. Indeed, we observed a strong correlation (Fig. 4b; r=.49, *p<.001*), which also holds for the individual neural
208   datasets (Fig. S6). These results suggest that further improving models' neural predictivity will simultaneously improve
209   their behavioral predictivity. An intriguing outlier in this analysis is the skip-thoughts model, which predicts neural activity
210   only moderately, but predicts reading times at ceiling.

211   **Next-word-prediction task performance correlates with behavioral predictivity.** Next-word-prediction task performance is predictive
212   of reading times (Fig. 4c; r=.37, *p<.05)*, in line with earlier studies (Goodkind & Bicknell, 2018; van Schijndel & Linzen, 2018).
213   Note that this relationship, similar to the brain-to-behavior one, is not as strong as the one between next-word-prediction
214   task performance and neural predictivity. This difference could point to additional mechanisms, on top of predictive
215   language processing, that were recruited for the reading task.

216

217   **Model architecture alone yields predictive representations.** The brain's language network plausibly arises through a
218   combination of evolutionary and learning-based optimization. Can we test the relative importance of these two factors
219   using model-to-brain comparisons? All models come with intrinsic architectural properties, like size, the presence of
220   recurrence, and the directionality and length of context used to perform the target task (Methods 5, Table S10). These
221   differences strongly affect model performance on normative tasks like next-word prediction after training, and define the
222   representational space that the model can learn (Arora et al., 2018; Fukushima, 1988). To test whether model architecture
223   alone—without training—already yields representational spaces that are similar to those implemented by the language
224   network in the brain, we evaluated models with their initial (random) weights. Strikingly, even with no training, several
225   model architectures reliably predicted brain activity and behavior (Fig. 5). For example, across the four datasets, untrained
226   GPT2-xl achieves an average predictivity of *~61%*, only *~14%* lower than the trained network. (Importantly, a random
227   context-independent embedding with equal dimensionality but no architectural priors predicts only a small fraction of the
228   datasets, on average below 30% (Fig. S8), suggesting that a large feature space alone, without architectural priors, is not
229   sufficient.) A similar trend is observed across models: training generally improves neural and behavioral predictivity, on
230   average by .1 (26% relative improvement). Across models, the untrained scores are strongly predictive of the trained scores
231   (*r=.82, p<<.00001*), indicating that models that predict human data poorly with random weights also perform poorly after
232   training, but models that already perform well with random weights improve further with training.
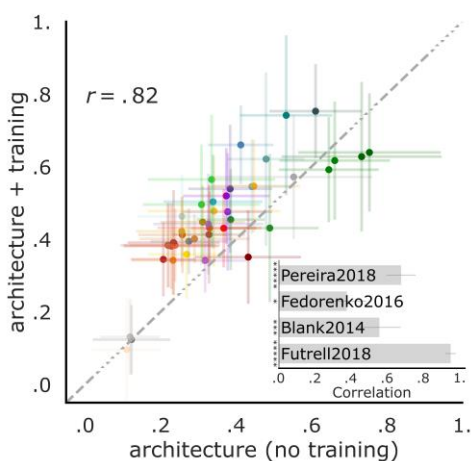

Figure 5: **Model architecture alone already yields predictive representations and untrained performance predicts trained performance**. We evaluate untrained models by keeping weights at their initial random values. The remaining representations are driven by architecture alone and are tested on the three neural (Fig. 2) and the behavioral dataset (Fig. 4). Across all datasets, architecture alone yields representations that predict human brain activity considerably well. On average, training improves model scores by 26%. For *Pereira2018*, training improves predictivity the most whereas for *Fedorenko2016*, *Blank2014* and *Futrell2018* training does not always change—and for some models even decreases—the similarity with human measurements (Fig. S7). The untrained model performance is consistently predictive of its performance after training across and within (inset) datasets.
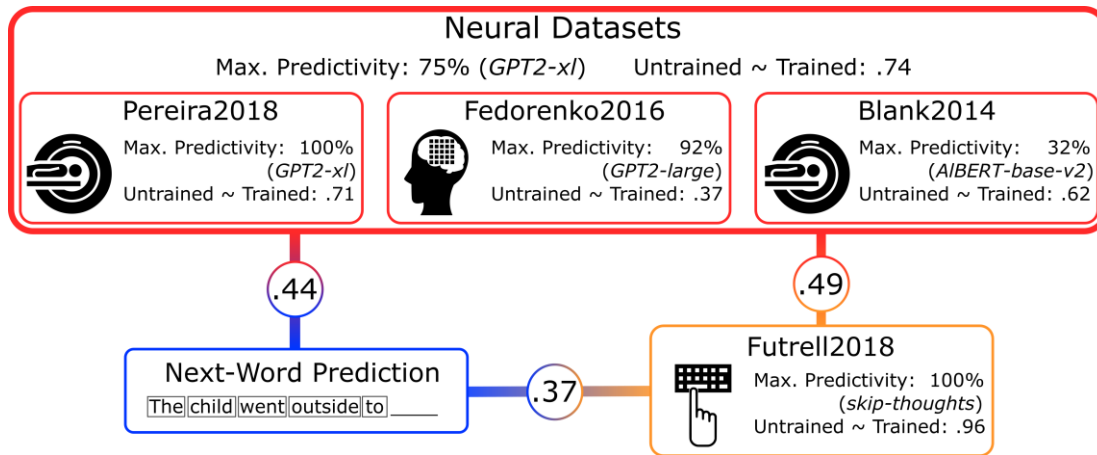
Figure 6: **Summary of the key results**. Normalized neural and behavioral predictivities are shown in the red and orange rectangles. For the neural datasets (averaged and individual, top row), and for the behavioral dataset (bottom right), we report i) the value for the model achieving the highest predictivity, and ii) the correlation between the untrained and trained scores. The next-word-prediction task (bottom left) predicts neural and behavioral scores; and neural scores predict behavioral scores.

233

Discussion

234

235 Our results, summarized in Fig. 6, show that specific ANN language models can predict human neural and behavioral
236 responses to linguistic input with high accuracy: the best models achieve, on some datasets, perfect predictivity relative to
237 the noise ceiling. Neural predictivity correlates across datasets spanning recording modalities (fMRI, ECoG, reading times)
238 and diverse materials presented visually and auditorily, establishing the robustness and generality of these findings.
239 Critically, both neural and behavioral predictivity correlate with model performance on the normative next-word prediction
240 task, but not other language tasks. Finally, model architectures alone, with random weights, produce representations that
241 capture neural and behavioral linguistic responses and closely 'track' the representations with learned weights across
242 datasets.

243

244 Underlying the integrative reverse-engineering framework, as implemented here in the cognitive domain of language, is the
245 idea that large-scale neural networks can serve as possible mechanistic hypotheses of brain processing. We here identified
246 some models—unidirectional-attention transformer architectures—that accurately capture brain activity during language
247 processing, and began dissecting variations across the range of model candidates to explain *why* they achieve high brain
248 predictivity. Two core findings emerged, both supporting the idea that the human language system is optimized for
249 predictive processing. First, we found that the models' performance on the next-word prediction ('language modeling') task,
250 but not other language tasks, relates to neural predictivity (see (Gauthier & Levy, 2019) for related evidence of fine-tuning
251 of one model on tasks other than next-word-prediction leading to worse model-to-brain fit). Language modeling is the task
252 of choice in the natural language processing (NLP) community: it is simple, unsupervised, scalable, and appears to produce
253 the most generally useful, successful language representations. This is likely because language modeling encourages a
254 neural network to build a joint probability model of the linguistic signal, which implicitly requires sensitivity to diverse kinds
255 of regularities in the signal. Second, we found that the models best matching human language processing are precisely
256 those that are trained to predict the next word. Predictive processing has advanced to the forefront of theorizing in
257 cognitive science (Clark, 2013; Tenenbaum et al., 2011) and neuroscience (Keller & Mrsic-Flogel, 2018), including in the
258 domain of language (Kuperberg & Jaeger, 2016; Levy, 2008a). The rich sources of information that comprehenders combine
259 to interpret language—including lexical and syntactic information, pragmatic reasoning, and world knowledge (Garnsey et
260 al., 1997; MacDonald et al., 1994; Tanenhaus et al., 1995; Trueswell et al., 1993, 1994)—can be used to make informed
261 guesses about how the linguistic signal may unfold, and much behavioral and neural evidence now suggests that readers
262 and listeners indeed engage in such predictive behavior (Altmann & Kamide, 1999; Frank & Bod, 2011; Kuperberg & Jaeger,
263 2016; Shain et al., 2020; Smith & Levy, 2013). Some accounts, rooted in the rich tradition of the analysis-by-synthesis
264 approach to cognition (Neisser, 1967), construe prediction as forward-simulation carried out by the language production

system that draws on the generative language model (Dell & Chang, 2014; Pickering & Garrod, 2013). An intriguing possibility is therefore that both the human language system and successful ANN models of language are optimized to predict upcoming words in the service of efficient meaning extraction.

We also demonstrated that architecture alone, with random weights, can yield representations that match human brain data well. If we construe model training as analogous to learning in human development, then human cortex might already provide a sufficiently rich structure that allows for the rapid acquisition of language (Rodriguez & Granger, 2016). Perhaps most of development is then a combination of the system wiring up (Saygin et al., 2016; Zador, 2019) and learning the right decoders on top of largely structurally defined features. In that analogy, community development of new architectures could be akin to evolution (Hasson et al., 2020), or perhaps, more accurately, selective breeding with genetic modification: structural changes are tested and the best-performing ones are incorporated into the next generation of models. Importantly, this process implicitly still optimizes for language modeling, only on a different timescale.

These discoveries pave the way for many exciting future directions. The most brain-like language models can now be investigated in richer detail, ideally leading to intuitive theories around their inner workings. Such research is much easier to perform on models than on biological systems since all their structure and weights are easily accessible and manipulable (Cheney et al., 2017; Lindsey et al., 2019). Controlled comparisons of minimally different architectural variants and training objectives could define the necessary and sufficient conditions for human-like language processing (Samek et al., 2017), synergizing with parallel ongoing efforts in NLP to probe ANNs' linguistic representations (Hewitt & Manning, 2019; Linzen et al., 2016; Tenney et al., 2020). Here, we worked with off-the-shelf models, and compared their match to neural data based on their performance on the next-word-prediction task vs. other tasks. Re-training many models on many tasks from scratch might determine which features are most important for brain predictivity, but is currently prohibitively expensive due to the insurmountable space of hyper-parameters. Further, the fact that language modeling is inherently built into the evolution of language models by the NLP community, as noted above, may make it impossible to fully eliminate its influences on the architecture even for models trained from scratch on other tasks.

How can we develop models that are even more brain-like? Despite impressive performance on the datasets and metrics here, ANN language models are far from human-level performance in the hardest problems of language understanding. An important open direction is to integrate language models like those used here with models and data resources that attempt to capture aspects of meaning important for commonsense world knowledge (e.g., Bisk et al., 2020; Bosselut et al., 2020; Sap et al., 2019, 2020; Yi et al., 2018). Such models might capture not only predictive processing in the brain—what word is likely to come next—but also semantic parsing, mapping language into conceptual representations that support grounded language understanding and reasoning (Bisk et al., 2020). The fact that language models lack meaning and focus on local linguistic coherence (Mahowald et al., 2020; Wilcox et al., 2020) may explain why their representations fall short of ceiling on *Blank2014*, which uses story materials and may therefore require long-range contexts.

One key missing piece in the mechanistic modeling of human language processing is a more detailed mapping from model components onto brain anatomy. In particular, aside from the general targeting of the fronto-temporal language network, it is unclear which parts of a model map onto which components of the brain's language processing mechanisms. In models of vision, for instance, attempts are made to map ANN layers and neurons onto cortical regions (Kubilius et al., 2019) and sub-regions (Lee & DiCarlo, 2018). However, whereas function and its mapping onto anatomy is at least coarsely defined in the case of vision (Felleman & Van Essen, 1991), a similar mapping is not yet established in language beyond the broad distinction between perceptual processing and higher-level linguistic interpretation (Fedorenko & Thompson-Schill, 2014). The network that supports higher-level linguistic interpretation—which we focus on here—is extensive and plausibly contains meaningful functional dissociations, but how the network is precisely subdivided and what respective roles its different components play remains debated. Uncovering the internal structure of the human language network, for which intracranial recording approaches with high spatial and temporal resolution may prove critical (Mukamel & Fried, 2012; Parvizi & Kastner, 2018), would allow us to guide and constrain models of tissue-mapped mechanistic language processing. More precise brain-to-model mappings would also allow us to test the effects of perturbations on models and compare them against perturbation effects in humans, as assessed with lesion studies or reversible stimulation. More broadly, anatomically and functionally precise models are a required software component of any form of brain-machine-interface.

10

316

317 Taken together, our findings suggest that predictive artificial neural networks serve as viable candidate hypotheses for how
318 predictive language processing is implemented in human neural tissue. They lay a critical foundation for a promising
319 research program synergizing high-performing mechanistic models of natural language processing with large-scale neural
320 and behavioral measurements of human language comprehension in a virtuous cycle of integrative reverse-engineering:
321 testing model ability to predict neural and behavioral brain measurements, dissecting the best-performing models to
322 understand which components are critical for high brain predictivity, developing better models leveraging this knowledge,
323 and collecting new data to challenge and constrain the future generations of neurally plausible models of language
324 processing.

325

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Arora, S., Cohen, N., & Hazan, E. (2018). On the optimization of deep networks: Implicit acceleration by overparameterization. *International Conference on Machine Learning (ICML)*, 372–389. http://arxiv.org/abs/1802.06509

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, 1–6. https://doi.org/10.1038/s41586-020-2350-5

Bates, E., Wilson, S. M., Saygin, A. P., Dick, F., Sereno, M. I., Knight, R. T., & Dronker, N. F. (2003). Voxel-based lesion-symptom mapping. *Nature Neuroscience*, *6*(5), 448–450. https://doi.org/10.1038/nn1050

Bautista, A., & Wilson, S. M. (2016). Neural responses to grammatically and lexically degraded speech. *Language, Cognition and Neuroscience*, *31*(4), 567–574. https://doi.org/10.1080/23273798.2015.1123281

Binder, J. R., Frost, J. A., Hammeke, T. A., Cox, R. W., Rao, S. M., & Prieto, T. (1997). Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience*, *17*(1), 353–362. https://doi.org/10.1523/JNEUROSCI.17-01-00353.1997

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., & Turian, J. (2020). Experience Grounds Language. *ArXiv Preprint*. http://arxiv.org/abs/2004.10151

Blank, I. A., & Fedorenko, E. (2017). Domain-general brain regions do not track linguistic input as closely as language-selective regions. *The Journal of Neuroscience*, *37*(41), 9999–10011. https://doi.org/10.1523/JNEUROSCI.3642-16.2017

Blank, I. A., & Fedorenko, E. (2020). No evidence for differences among language regions in their temporal receptive windows. *NeuroImage*, *219*, 116925. https://doi.org/10.1016/j.neuroimage.2020.116925

Blank, I., Balewski, Z., Mahowald, K., & Fedorenko, E. (2016). Syntactic processing is distributed across the language system. *NeuroImage*, *127*, 307–323. https://doi.org/10.1016/j.neuroimage.2015.11.069

Blank, I., Kanwisher, N., & Fedorenko, E. (2014). A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of Neurophysiology*, *112*(5), 1105–1118. https://doi.org/10.1152/jn.00884.2013

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., & Choi, Y. (2020). CoMET: Commonsense transformers for automatic knowledge graph construction. *Association for Computational Linguistics (ACL)*, 4762–4779. https://doi.org/10.18653/v1/p19-1470

Brennan, J. R., & Pylkkänen, L. (2017). MEG Evidence for Incremental Sentence Composition in the Anterior Temporal Lobe. *Cognitive Science*, *41*, 1515–1531. https://doi.org/10.1111/cogs.12445

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W. M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157–158*, 81–94. https://doi.org/10.1016/j.bandl.2016.04.008

Buzsáki, G., Anastassiou, C. A., & Koch, C. (2012). The origin of extracellular fields and currents-EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience*, *13*(6), 407–420. https://doi.org/10.1038/nrn3241

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLOS Computational Biology*, *15*(4), 1–27. https://doi.org/10.1371/journal.pcbi.1006897

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2018). SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. *International Workshop on Semantic Evaluation*, 1–14. https://doi.org/10.18653/v1/s17-2001

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2014). One billion word benchmark for measuring progress in statistical language modeling. *Annual Conference of the International Speech Communication Association*, 2635–2639. http://arxiv.org/abs/1312.3005

Cheney, N., Schrimpf, M., & Kreiman, G. (2017). On the Robustness of Convolutional Neural Networks to Internal Architecture and Weight Perturbations. *ArXiv Preprint*. http://arxiv.org/abs/1703.08245

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*. https://doi.org/10.1038/srep27755

Ciregan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. *Computer

378   *Vision and Pattern Recognition (CVPR)*, 3642–3649. https://doi.org/10.1109/CVPR.2012.6248110

379 Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain*
380   *Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

381 Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., &
382   Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. *ArXiv Preprint*.
383   http://arxiv.org/abs/1911.02116

384 Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2020). Transformer-XL: Attentive language models
385   beyond a fixed-length context. *Association for Computational Linguistics (ACL)*, 2978–2988.
386   https://doi.org/10.18653/v1/p19-1285

387 Dell, G. S., & Chang, F. (2014). The p-chain: Relating sentence production and its disorders to comprehension and
388   acquisition. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 369, Issue 1634, p.
389   20120394). The Royal Society. https://doi.org/10.1098/rstb.2012.0394

390 Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity.
391   *Cognition*, *109*(2), 193–210. https://doi.org/10.1016/j.cognition.2008.07.008

392 Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). The Representation of Semantic Information Across
393   Human Cerebral Cortex During Listening Versus Reading Is Invariant to Stimulus Modality. *Journal of Neuroscience*,
394   *39*(39), 7722–7736. https://doi.org/10.1523/JNEUROSCI.0675-19.2019

395 Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language
396   Understanding. *ArXiv Preprint*. https://arxiv.org/abs/1810.04805

397 Diachek, E., Blank, I., Siegelman, M., Affourtit, J., & Fedorenko, E. (2020). The domain-general multiple demand (MD)
398   network does not support core aspects of language comprehension: A large-scale fMRI investigation. *Journal of*
399   *Neuroscience*, *40*(23), 4536–4550. https://doi.org/10.1523/JNEUROSCI.2036-19.2020

400 Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2019). Topic Modeling in Embedding Spaces. *ArXiv Preprint*.
401   http://arxiv.org/abs/1907.04907

402 Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2015). Cortical tracking of hierarchical linguistic structures in
403   connected speech. *Nature Neuroscience*, *19*(1), 158–164. https://doi.org/10.1038/nn.4186

404 Dolan, W. B., & Brockett, C. (2005). Automatically Constructing a Corpus of Sentential Paraphrases. *International Workshop*
405   *on Paraphrasing (IWP)*, 9–16. https://research.microsoft.com/apps/pubs/default.aspx?id=101076

406 Dotlačil, J. (2018). Building an ACT-R Reader for Eye-Tracking Corpus Data. *Topics in Cognitive Science*, *10*(1), 144–160.
407   https://doi.org/10.1111/tops.12315

408 Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human
409   brain. *Proceedings of the National Academy of Sciences (PNAS)*, *108*(39), 16428–16433.
410   https://doi.org/10.1073/pnas.1112937108

411 Fedorenko, E., Blank, I., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings
412   throughout the language network. *BioRxiv Preprint*. https://doi.org/10.1101/477851

413 Fedorenko, E., Hsieh, P. J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI
414   investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, *104*(2), 1177–
415   1194. https://doi.org/10.1152/jn.00032.2010

416 Fedorenko, E., Nieto-Castañón, A., & Kanwisher, N. (2012). Lexical and syntactic representations in the brain: An fMRI
417   investigation with multi-voxel pattern analyses. *Neuropsychologia*, *50*(4), 499–513.
418   https://doi.org/10.1016/j.neuropsychologia.2011.09.014

419 Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the
420   construction of sentence meaning. *Proceedings of the National Academy of Sciences of the United States of America*
421   *(PNAS)*, *113*(41), E6256–E6262. https://doi.org/10.1073/pnas.1612132113

422 Fedorenko, E., & Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in Cognitive Sciences*, *18*(3), 120–
423   126. https://doi.org/10.1016/j.tics.2013.12.006

424 Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierachical processing in the primate cerebral cortex. *Cerebral Cortex*,
425   *1*(1), 1–47. https://doi.org/10.1093/cercor/1.1.1

426 Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological*
427   *Science*, *22*(6), 829–834. https://doi.org/10.1177/0956797611409589

428 Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words
429   in sentences. *Brain and Language*, *140*. https://doi.org/10.1016/j.bandl.2014.10.006

430 Friederici, A. D. (2012). The cortical language circuit: From auditory perception to sentence comprehension. *Trends in*

Cognitive Sciences, *16*(5), 262–268. https://doi.org/10.1016/j.tics.2012.04.001

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, *1*(2), 119–130. https://doi.org/10.1016/0893-6080(88)90014-7

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, *44*(3). https://doi.org/10.1111/cogs.12814

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2020). The natural stories corpus. *International Conference on Language Resources and Evaluation (LREC)*, 76–82. http://arxiv.org/abs/1708.05763

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, *37*(1), 58–93. https://doi.org/10.1006/jmla.1997.2512

Gauthier, J., & Ivanova, A. (2018). *Does the brain represent words? An evaluation of brain decoding studies of language understanding*. http://arxiv.org/abs/1806.00591

Gauthier, J., & Levy, R. (2019). Linking artificial and human neural representations of language. *Empirical Methods for Natural Language Processing (EMNLP)*, 529–539. https://doi.org/10.18653/v1/d19-1050

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1). https://doi.org/10.1016/S0010-0277(98)00034-1

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, *110*(20), 8051–8056. https://doi.org/10.1073/pnas.1216438110

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *Cognitive Modeling and Computational Linguistics (CMCL)*, 10–18. https://doi.org/10.18653/v1/w18-0102

Gorno-Tempini, M. L., Dronkers, N. F., Rankin, K. P., Ogar, J. M., Phengrasamy, L., Rosen, H. J., Johnson, J. K., Weiner, M. W., & Miller, B. L. (2004). Cognition and Amatomy in Three Variants of Primary Progressive Aphasia. *Annals of Neurology*, *55*(3), 335–346. https://doi.org/10.1002/ana.10825

Hagoort, P. (2019). The neurobiology of language beyond single-word processing. *Science*, *366*(6461), 55–58. https://doi.org/10.1126/science.aax0289

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *North American Chapter of the Association for Computational Linguistics (NAACL)*, 1–8. https://doi.org/10.3115/1073336.1073357

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron*, *105*(3), 416–434. https://doi.org/10.1016/j.neuron.2019.12.002

Henderson, J. M., Choi, W., Lowder, M. W., & Ferreira, F. (2016). Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *NeuroImage*, *132*, 293–300. https://doi.org/10.1016/j.neuroimage.2016.02.050

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. *North American Chapter of the Association for Computational Linguistics (NAACL)*, *1*, 4129–4138.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). *A Systematic Assessment of Syntactic Generalization in Neural Language Models*. http://arxiv.org/abs/2005.03692

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. https://doi.org/10.1038/nature17637

Jain, S., & Huth, A. (2018, May 21). Incorporating Context into Language Encoding Models for fMRI. *Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.1101/327601

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). *Exploring the Limits of Language Modeling*. http://arxiv.org/abs/1602.02410

Jurafsky, D. (1996). A Probabilistic Model of Lexical and Syntactic Access and Disambiguation. *Cognitive Science*, *20*(2), 137–194. https://doi.org/10.1207/s15516709cog2002_1

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. https://doi.org/10.1037/0033-295X.87.4.329

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238. https://doi.org/10.1037/0096-3445.111.2.228

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, *98*(3), 630–644. https://doi.org/10.1016/j.neuron.2018.03.044

484 Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive Processing: A Canonical Cortical Computation. *Neuron*, *100*(2), 424–435.
485   https://doi.org/10.1016/j.neuron.2018.10.003
486 Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). CTRL: A Conditional Transformer Language Model
487   for Controllable Generation. *ArXiv Preprint*. http://arxiv.org/abs/1909.05858
488 Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to
489   capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*
490   *(PNAS)*, *116*(43), 21854–21863. https://doi.org/10.1073/pnas.1905544116
491 Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-Thought Vectors. *Neural*
492   *Information Processing Systems (NIPS)*, 3294–3302. http://papers.nips.cc/paper/5950-skip-thought-vectors
493 Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in*
494   *Systems Neuroscience*, *2*. https://doi.org/10.3389/neuro.06.004.2008
495 Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks.
496   *Neural Information Processing Systems (NIPS)*. https://doi.org/http://dx.doi.org/10.1016/j.protcy.2014.09.007
497 Kubilius, J., Schrimpf, M., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K.,
498   Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2019). Brain-Like Object Recognition with High-Performing
499   Shallow Recurrent ANNs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. D'Alché-Buc, E. Fox, & R. Garnett (Eds.),
500   *Neural Information Processing Systems (NeurIPS)* (pp. 12785--12796). Curran Associates, Inc.
501   http://arxiv.org/abs/1909.06161
502 Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition*
503   *and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299
504 Lakretz, Y., Dehaene, S., & King, J. R. (2020). What limits our capacity to process nested long-range dependencies in
505   sentence comprehension? *Entropy*, *22*(4), 446. https://doi.org/10.3390/E22040446
506 Lample, G., & Conneau, A. (2019). Cross-lingual Language Model Pretraining. *Neural Information Processing Systems*
507   *(NeurIPS)*, 7059–7069. http://arxiv.org/abs/1901.07291
508 Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning
509   of Language Representations. *ArXiv Preprint*. http://arxiv.org/abs/1909.11942
510 Lawrence Marple, S. (1999). Computing the discrete-time analytic signal via fft. *IEEE Transactions on Signal Processing*,
511   *47*(9), 2600–2603. https://doi.org/10.1109/78.782222
512 Lee, H., & DiCarlo, J. (2018, September 21). Topographic Deep Artificial Neural Networks (TDANNs) predict face selectivity
513   topography in primate inferior temporal (IT) cortex. *Cognitive Computational Neuroscience (CCN)*.
514   https://doi.org/10.32470/ccn.2018.1085-0
515 Levesque, H. J., Davis, E., & Morgenstern, L. (2012). The winograd schema challenge. *International Workshop on Temporal*
516   *Representation and Reasoning*, 552–561. www.aaai.org
517 Levy, R. (2008a). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.
518   https://doi.org/10.1016/j.cognition.2007.05.006
519 Levy, R. (2008b). A noisy-channel model of rational human sentence comprehension under uncertain input. *Empirical*
520   *Methods in Natural Language Processing (EMNLP)*, 234–243. https://doi.org/10.3115/1613715.1613749
521 Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension.
522   *Trends in Cognitive Sciences*, *10*(10), 447–454. https://doi.org/10.1016/j.tics.2006.08.007
523 Lindsey, J., Ocko, S. A., Ganguli, S., & Deny, S. (2019, January 3). A unified theory of early visual representations from retina
524   to cortex through anatomically constrained deep cnNs. *International Conference on Learning Representations (ICLR)*.
525   http://arxiv.org/abs/1901.00945
526 Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies.
527   *Transactions of the Association for Computational Linguistics*, *4*, 521–535. https://doi.org/10.1162/tacl_a_00115
528 Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A
529   Robustly Optimized BERT Pretraining Approach. *ArXiv Preprint*. http://arxiv.org/abs/1907.11692
530 Lopopolo, A., Frank, S. L., Van Den Bosch, A., & Willems, R. M. (2017). Using stochastic language models (SLM) to map
531   lexical, syntactic, and phonological information processing in the brain. *PLoS ONE*, *12*(5).
532   https://doi.org/10.1371/journal.pone.0177794
533 Lyu, B., Choi, H. S., Marslen-Wilson, W. D., Clarke, A., Randall, B., & Tyler, L. K. (2019). Neural dynamics of semantic
534   composition. *Proceedings of the National Academy of Sciences (PNAS)*, *116*(42), 21318–21327.
535   https://doi.org/10.1073/pnas.1903402116
536 MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution.

537   *Psychological Review*, *101*(4), 676–703. https://doi.org/10.1037/0033-295x.101.4.676

538   Mahowald, K., Kachergis, G., & Frank, M. C. (2020). What counts as an exemplar model, anyway? A commentary on
539   Ambridge (2020). *First Language*. https://doi.org/10.1177/0142723720905920

540   Marcus, G. (2018). Deep Learning: A Critical Appraisal. *ArXiv Preprint*. http://arxiv.org/abs/1801.00631

541   McDonald, J., & Macwhinney, B. (1998). Maximum Likelihood Models for Sentence Processing. In *The Crosslinguistic Study*
542   *of   Sentence   Processing*.
543   https://www.researchgate.net/publication/230876309_Maximum_Likelihood_Models_for_Sentence_Processing

544   Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). Pointer Sentinel Mixture Models. *ArXiv Preprint*.
545   http://arxiv.org/abs/1609.07843

546   Merkx, D., & Frank, S. L. (2020). Comparing Transformers and RNNs on predicting human sentence processing data. *ArXiv*
547   *Preprint*. http://arxiv.org/abs/2005.09471

548   Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013, October 16). Distributed representations ofwords and
549   phrases and their compositionality. *Neural Information Processing Systems (NIPS)*. http://arxiv.org/abs/1310.4546

550   Mitchell, D. C. (1984). Computational psycholinguistics View project Psycholinguistics View project. *New Methods in*
551   *Reading Comprehension Research*. https://www.researchgate.net/publication/286455549

552   Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting
553   human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191–1195.
554   https://doi.org/10.1126/science.1152876

555   Monti, M. M., Parsons, L. M., & Osherson, D. N. (2012). Thought Beyond Language: Neural Dissociation of Algebra and
556   Natural Language. *Psychological Science*, *23*(8), 914–922. https://doi.org/10.1177/0956797612437427

557   Mukamel, R., & Fried, I. (2012). Human Intracranial Recordings and Cognitive Neuroscience. *Annual Review of Psychology*,
558   *63*(1), 511–537. https://doi.org/10.1146/annurev-psych-120709-145401

559   Neisser, U. (1967). Cognitive psychology. *American Psychological Association*. https://psycnet.apa.org/record/1967-35031-
560   000

561   Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S. S., Naccache, L., Hale, J. T., Pallier, C., &
562   Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing.
563   *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, *114*(18), E3669–E3678.
564   https://doi.org/10.1073/pnas.1701590114

565   Pallier, C., Devauchellea, A. D., & Dehaenea, S. (2011). Cortical representation of the constituent structure of sentences.
566   *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, *108*(6), 2522–2527.
567   https://doi.org/10.1073/pnas.1018711108

568   Parvizi, J., & Kastner, S. (2018). Promises and limitations of human intracranial electroencephalography. *Nature*
569   *Neuroscience*, *21*(4), 474–483. https://doi.org/10.1038/s41593-018-0108-2

570   Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014*
571   *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.
572   https://doi.org/10.3115/v1/D14-1162

573   Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a
574   universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*.
575   https://doi.org/10.1038/s41467-018-03068-4

576   Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain*
577   *Sciences*, *36*(4), 329–347. https://doi.org/10.1017/S0140525X12001495

578   Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of
579   language acquisition. *Cognition*, *28*(1–2), 73–193. https://doi.org/10.1016/0010-0277(88)90032-7

580   Price, C. J. (2010). The anatomy of language: A review of 100 fMRI studies published in 2009. In *Annals of the New York*
581   *Academy of Sciences* (Vol. 1191, pp. 62–88). https://doi.org/10.1111/j.1749-6632.2010.05444.x

582   Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic
583   representation of meaning. *Nature Human Behaviour*, *2*(9), 693–705. https://doi.org/10.1038/s41562-018-0406-4

584   Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-*
585   *Training*. https://gluebenchmark.com/leaderboard

586   Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask
587   Learners. *ArXiv Preprint*. https://github.com/codelucas/newspaper

588   Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the Limits of
589   Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv Preprint*. http://arxiv.org/abs/1910.10683

590 Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-Scale, High-Resolution Comparison
591   of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural
592   Networks. *The Journal of Neuroscience*, *38*(33), 7255–7269. https://doi.org/10.1523/JNEUROSCI.0388-18.2018
593 Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuad: 100,000+ questions for machine comprehension of text.
594   *Empirical Methods in Natural Language Processing (EMNLP)*, 2383–2392. http://arxiv.org/abs/1606.05250
595 Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin*, *85*(3), 618–660.
596   https://doi.org/10.1037/0033-2909.85.3.618
597 Regev, M., Honey, C. J., Simony, E., & Hasson, U. (2013). Selective and invariant neural responses to spoken and written
598   narratives. *Journal of Neuroscience*, *33*(40), 15978–15988. https://doi.org/10.1523/JNEUROSCI.1580-13.2013
599 Rodriguez, A., & Granger, R. (2016). The grammar of mammalian brain capacity. *Theoretical Computer Science*, *633*, 100–
600   111. https://doi.org/10.1016/j.tcs.2016.03.021
601 Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and
602   Interpreting Deep Learning Models. *ArXiv Preprint*. http://arxiv.org/abs/1708.08296
603 Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and
604   lighter. *ArXiv Preprint*. http://arxiv.org/abs/1910.01108
605 Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., & Choi, Y. (2019). ATOMIC: An
606   Atlas of Machine Commonsense for If-Then Reasoning. *AAAI Conference on Artificial Intelligence*, *33*, 3027–3035.
607   https://doi.org/10.1609/aaai.v33i01.33013027
608 Sap, M., Rashkin, H., Chen, D., Le Bras, R., & Choi, Y. (2020). Social IQA: Commonsense reasoning about social interactions.
609   *Empirical Methods in Natural Language Processing (EMNLP)*, 4463–4473. https://doi.org/10.18653/v1/d19-1454
610 Saygin, Z. M., Osher, D. E., Norton, E. S., Youssoufian, D. A., Beach, S. D., Feather, J., Gaab, N., Gabrieli, J. D. E., & Kanwisher,
611   N. (2016). Connectivity precedes function in the development of the visual word form area. *Nature Neuroscience*,
612   *19*(9), 1250–1255. https://doi.org/10.1038/nn.4354
613 Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K.,
614   Yamins, D. L. K., & DiCarlo, J. J. (2018). Brain-Score: Which Artificial Neural Network for Object Recognition is most
615   Brain-Like? *BioRxiv*. https://doi.org/10.1101/407007
616 Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., & Ajemian, R. (2020). Integrative Benchmarking to Advance Neurally
617   Mechanistic Models of Human Intelligence. *Neuron*. https://doi.org/10.1016/j.neuron.2020.07.040
618 Schwartz, D., Toneva, M., & Wehbe, L. (2019). Inducing brain-relevant bias in natural language processing models. *Advances
619   in Neural Information Processing Systems*, *32*, 14123–14133. https://github.com/danrsc/bert_brain_neurips_2019
620 Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive
621   coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*.
622   https://doi.org/10.1016/j.neuropsychologia.2019.107307
623 Singer, Y., Teramoto, Y., Willmore, B. D. B., King, A. J., Schnupp, J. W. H., & Harper, N. S. (2018). Sensory cortex is optimised
624   for prediction of future input. *ELife*, *7*. https://doi.org/10.7554/eLife.31557
625 Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.
626   https://doi.org/10.1016/j.cognition.2013.02.013
627 Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for
628   semantic compositionality over a sentiment treebank. *Empirical Methods in Natural Language Processing (EMNLP)*,
629   1631–1642. http://nlp.stanford.edu/
630 Spivey-Knowlton, M. J. (1996). *Integration of visual and linguistic information: Human data and model simulations*.
631   University of Rochester.
632 Steedman, M. (2000). *The Syntactic Process*. MIT Press. https://mitpress.mit.edu/books/syntactic-process
633 Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic
634   information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.
635   https://doi.org/10.1126/science.7777863
636 Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and
637   Abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788
638 Tenney, I., Das, D., & Pavlick, E. (2020). BERT rediscovers the classical NLP pipeline. *Association for Computational Linguistics
639   (ACL)*, 4593–4601. https://doi.org/10.18653/v1/p19-1452
640 Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural
641   language-processing (in the brain). *Advances in Neural Information Processing Systems*, *32*, 14954–14964.
642   http://arxiv.org/abs/1905.11833

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic Influences On Parsing: Use of Thematic Role Information in Syntactic Ambiguity Resolution. *Journal of Memory and Language*, *33*(3), 285–318. https://doi.org/10.1006/jmla.1994.1014

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-Specific Constraints in Sentence Processing: Separating Effects of Lexical Preference From Garden-Paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 528–553. https://doi.org/10.1037/0278-7393.19.3.528

van Schijndel, M., Exley, A., & Schuler, W. (2013). A Model of Language Processing as Hierarchic Sequential Prediction. *Topics in Cognitive Science*, *5*(3), 522–540. https://doi.org/10.1111/tops.12034

van Schijndel, M., & Linzen, T. (2018). A neural model of adaptation in reading. *Empirical Methods in Natural Language Processing (EMNLP)*, 4704–4710. http://arxiv.org/abs/1808.09930

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Neural Information Processing Systems (NeurIPS)*, 3266–3280. http://arxiv.org/abs/1905.00537

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019, September 27). Glue: A multi-task benchmark and analysis platform for natural language understanding. *International Conference on Learning Representations (ICLR)*.

Wang, S., Zhang, J., Wang, H., Lin, N., & Zong, C. (2020). Fine-grained neural decoding with distributed word representations. *Information Sciences*, *507*, 256–272. https://doi.org/10.1016/j.ins.2019.08.043

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, *7*, 625–641. https://doi.org/10.1162/tacl_a_00290

Wehbe, L., Blank, I. A., Shain, C., Futrell, R., Levy, R., Malsburg, T. von der, Smith, N., Gibson, E., & Fedorenko, E. (2020). Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. *BioRxiv Preprint*. https://doi.org/10.1101/2020.04.15.043844

Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014). Aligning context-based statistical models of language with brain activity during reading. *Empirical Methods in Natural Language Processing (EMNLP)*, 233–243. http://www.aclweb.org/anthology/D14-1030

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. *ArXiv Preprint*. http://arxiv.org/abs/2006.01912

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van Den Bosch, A. (2016). Prediction during Natural Language Comprehension. *Cerebral Cortex*, *26*(6), 2506–2516. https://doi.org/10.1093/cercor/bhv075

Williams, A., Nangia, N., & Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, *1*, 1112–1122. https://doi.org/10.18653/v1/n18-1101

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv Preprint*. http://arxiv.org/abs/1910.03771

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences (PNAS)*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *ArXiv Preprint*. http://arxiv.org/abs/1906.08237

Yi, K., Torralba, A., Wu, J., Kohli, P., Gan, C., & Tenenbaum, J. B. (2018). Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. *Neural Information Processing Systems (NeurIPS)*, *2018-Decem*, 1031–1042. http://nsvqa.csail.mit.edu

Zador, A. (2019). A Critique of Pure Learning: What Artificial Neural Networks can Learn from Animal Brains. *BioRxiv Preprint*. https://doi.org/10.1101/582643

Zhuang, C., Kubilius, J., Hartmann, M. J., & Yamins, D. L. (2017). Toward Goal-Driven Neural Network Models for the Rodent Whisker-Trigeminal System. *Neural Information Processing Systems (NIPS)*, 2555–2565. http://papers.nips.cc/paper/6849-toward-goal-driven-neural-network-models-for-the-rodent-whisker-trigeminal-system

694 ## Methods

695 **1. Neural dataset 1: fMRI (*Pereira2018*).** We used the data from Pereira et al.'s (2018) Experiments 2 (n=9) and 3 (n=6) (10
696 unique participants). (The set of participants is not identical to Pereira et al., 2018: i) one participant (tested at Princeton)
697 was excluded from both experiments here to keep the fMRI scanner the same across participants; and ii) two participants
698 who were excluded from Experiment 2 in Pereira et al., 2018, based on the decoding results in Experiment 1 of that study
699 were included here, to err on the conservative side.) Stimuli for Experiment 2 consisted of 384 sentences (96 text passages,
700 four sentences each), and stimuli for Experiment 3 consisted of 243 sentences (72 text passages, 3 or 4 sentences each). The
701 two sets of materials were constructed independently, and each spanned a broad range of content areas. Sentences were
702 7-18 words long in Experiment 2, and 5-20 words long in Experiment 3. The sentences were presented on the screen one at
703 a time for 4s (followed by 4s of fixation, with additional 4s of fixation at the end of each passage), and each participant read
704 each sentence three times, across independent scanning sessions (see Pereira et al., 2018 for details of experimental
705 procedure and data acquisition).

706 *Preprocessing and response estimation*: Data preprocessing was carried out with SPM5 (using default parameters, unless
707 specified otherwise) and supporting, custom MATLAB scripts. (Note that SPM was only used for preprocessing and basic
708 modeling—aspects that have not changed much in later versions; for several datasets, we have directly compared the
709 outputs of data preprocessed and modeled in SPM5 vs. SPM12, and the outputs were nearly identical.) Preprocessing
710 included motion correction (realignment to the mean image of the first functional run using 2nd-degree b-spline
711 interpolation), normalization (estimated for the mean image using trilinear interpolation), resampling into 2mm isotropic
712 voxels, smoothing with a 4mm FWHM Gaussian filter and high-pass filtering at 200s. A standard mass univariate analysis
713 was performed in SPM5 whereby a general linear model (GLM) estimated the response to each sentence in each run. These
714 effects were modeled with a boxcar function convolved with the canonical Hemodynamic Response Function (HRF). The
715 model also included first-order temporal derivatives of these effects (which were not used in the analyses), as well as
716 nuisance regressors representing entire experimental runs and offline-estimated motion parameters.

717 *Functional localization*: Data analyses were performed on fMRI BOLD signals extracted from the bilateral fronto-temporal
718 language network. This network was defined functionally in each participant using a well-validated language localizer task
719 (Fedorenko et al., 2010), where participants read sentences vs. lists of nonwords. This contrast targets brain areas that
720 support 'high-level' linguistic processing, past the perceptual (auditory/visual) analysis. Brain regions that this localizer
721 identifies are robust to modality of presentation (e.g., Fedorenko et al., 2010; Scott et al., 2017), as well as materials and
722 task (Diachek et al., 2020). Further, these regions have been shown to exhibit strong sensitivity to both lexico-semantic
723 processing (understanding individual word meanings) and combinatorial, syntactic/semantic processing (putting words
724 together into phrases and sentences) (Bautista & Wilson, 2016; I. Blank et al., 2016; I. A. Blank & Fedorenko, 2020;
725 Fedorenko et al., 2010, 2012, 2016, 2020). Following prior work, we used group-constrained, participant-specific functional
726 localization (Fedorenko et al., 2010). Namely, individual activation maps for the target contrast (here, sentences>nonwords)
727 were combined with "constraints" in the form of spatial 'masks'—corresponding to data-driven, large areas within which
728 most participants in a large, independent sample show activation for the same contrast. The masks (available from
729 https://evlab.mit.edu/funcloc/ and used in many prior studies e.g., Jouravlev et al., 2019; Diachek et al., 2020; Shain et al.,
730 2020) included six regions in each hemisphere: three in the frontal cortex (two in the inferior frontal gyrus, including its
731 orbital portion: IFGorb, IFG; and one in the middle frontal gryus: MFG), two in the anterior and posterior temporal cortex
732 (AntTemp and PostTemp), and one in the angular gyrus (AngG). Within each mask, we selected 10% of most localizer-
733 responsive voxels (voxels with the highest t-value for the localizer contrast) following the standard approach in prior work.
734 This approach allows to pool data from the same functional regions across participants even when these regions do not
735 align well spatially. Functional localization has been shown to be more sensitive and to have higher functional resolution
736 (Nieto-Castanon & Fedorenko, 2012) than the traditional group-averaging approach (Holmes & Friston, 1998), which
737 assumes voxel-wise correspondence across participants. This is to be expected given the well-established inter-individual
738 differences in the mapping of function to anatomy, especially pronounced in the association cortex (e.g., Frost & Goebel,
739 2012; Tahmasebi et al., 2012; Vazquez-Rodriguez et al., 2019).

740 We constructed a stimulus-response matrix for each of the two experiments by i) averaging the BOLD responses to each
741 sentence in each experiment across the three repetitions, resulting in 1 data point per sentence per language-responsive
742 voxel of each participant, selected as described above (13,553 voxels total across the 10 participants; 1,355 average, ±6 std.

19

743 dev.), and ii) concatenating all sentences (384 in Experiment 2 and 243 in Experiment 3), yielding a 384x12,195 matrix for
744 Experiment 2, and a 243x8,121 matrix for Experiment 3.

745 To examine differences in neural predictivity between the language network and other parts of the brain, we additionally
746 extracted fMRI BOLD signals from two other networks: the multiple demand (MD) network (Duncan, 2010; Fedorenko et al.,
747 2013) and the default mode network (DMN) (Buckner et al., 2008; Buckner & DiNicola, 2019). These networks were also
748 defined functionally using well-validated localizer contrasts (Fedorenko et al., 2013; Mineroff et al., 2018) using a similar
749 procedure as the one used for defining the language network: combining a set of 'masks' with individual activation maps,
750 and selecting top 10% of most localizer-responsive voxels within each mask. Both networks were defined using a spatial
751 working memory task (Fedorenko et al., 2011, 2013). For the MD network, we used the hard>easy contrast, and for the
752 DMN network, we used the fixation>hard contrast. As for the language network, the MD and DMN masks were derived
753 from large sets of participants for those contrasts, and are also available at https://evlab.mit.edu/funcloc/. The MD network
754 and the DMN included 29,936 (2,994±230) and 10,978 (1,098±7) voxels, respectively.

755

756 **2. Neural dataset 2: ECoG (*Fedorenko2016*).** We used the data from Fedorenko et al.'s (2016) study (n=5). (The set of
757 participants includes one participant, S2, who was excluded from the main analyses in Fedorenko et al., 2016 due to a small
758 number of electrodes of interest; because we here used only language-responsiveness as the criterion for electrode
759 selection, this participant had enough electrodes to be included.) Stimuli consisted of 80 hand-constructed 8-word long
760 semantically and syntactically diverse sentences and 80 lists of nonwords (as well as some other stimuli not used in the
761 current study). For the critical analyses, we selected a set of 52 sentences that were presented to all participants. The
762 materials were presented visually one word at a time (for 450 or 700 ms), and participants performed a memory probe task
763 after each stimulus (see Fedorenko et al., 2016 for details of the experimental procedure and data acquisition).

764 *Preprocessing and response estimation*: We here provide only a brief summary, highlighting points of deviation from
765 Fedorenko et al. (2016). The total numbers of implanted electrodes were 120, 128, 112, 134, and 98 for the five
766 participants, respectively. Signals were digitized at 1200 Hz. Similar to Fedorenko et al. (2016), i) the recordings were high-
767 pass filtered with a cut off frequency of 0.5 Hz; ii) reference, ground, and electrodes with high noise levels were removed,
768 leaving 117, 118, 92, 130, and 88 electrodes (for these analyses, we were more permissive with respect to noise levels
769 compared to Fedorenko et al., 2016, to include as many electrodes in the analyses as possible; hence the numbers of
770 analyzed electrodes are higher here than in the original study for 4 of the 5 participants); iii) spatially distributed noise
771 common to all electrodes was removed using a common average reference spatial filter between electrodes with line noise
772 smaller than a predefined threshold (electrodes connected to the same amplifier); and iv) a set of notch filters were used to
773 remove the 60 Hz line noise and its harmonics. To extract the high gamma band activity—which has been shown to
774 correspond to spiking neural activity in the vicinity of the electrodes (Buzsáki et al., 2012)—we used a gaussian filter bank
775 with centers at 73, 79.5, 87.8, 96.9, 107, 118.1, 130.4, and 144 Hz, and standard deviations of 4.68, 4.92, 5.17, 5.43, 5.7,
776 5.99, 6.3, and 6.62 Hz, respectively. This approach differs from Fedorenko et al. (2016), where an IIR band-pass filter was
777 used to select frequencies in the range of 70-170 Hz, and is likely more sensitive (Dichter et al. 2018). Finally, as in
778 Fedorenko et al. (2016), the Hilbert transform was used to extract the analytic signal (Lawrence Marple, 1999) (except here,
779 the average of the Hilbert signal across the eight filters was used as high-gamma signal), z-scored for each electrode with
780 respect to the activity throughout the experiment, and the signal envelopes were downsampled to 300 Hz for further
781 analysis (we did not additionally low-pass filter at 100 Hz, as in Fedorenko et al., 2016).

782 *Functional localization:* Mirroring the fMRI approach, where we focused on language-responsive voxels, data analyses were
783 performed on signals extracted from language-responsive electrodes. These electrodes were defined in each participant
784 using the same localizer contrast as in the fMRI datasets. In particular, we examined electrodes in which the envelope of the
785 high gamma signal was significantly higher (at p<.01) for trials of the sentence condition than the nonword-list condition
786 (for details, see Fedorenko et al., 2016).

787 We constructed a stimulus-response matrix by i) averaging the *z*-scored high-gamma signal over the full presentation
788 window of each word in each sentence, resulting in 8 data points per sentence per language-responsive electrode (97

20

789   electrodes total across the 5 participants; 47, 8, 9, 15, and 18 for participants S1 through S5, respectively), and ii)
790   concatenating all words in all sentences (416 words across the 52 sentences), yielding a 416x97 matrix.

791   To examine differences in neural predictivity between language-responsive and other electrodes, we additionally extracted
792   high gamma signals from a set of 'stimulus-responsive' electrodes. Stimulus-responsive electrodes were defined as
793   electrodes in which the envelope of the high gamma signal for the sentence condition was significantly different (at p<0.05
794   by a paired-samples *t*-test) from the activity during the inter-trial fixation interval preceding the trial. This selection
795   procedure resulted in 67, 35, 20, 29, and 26 electrodes. As expected, this set of electrodes included many of the language-
796   responsive electrodes; for the analysis in SI-4, we exclude the language-responsive electrodes leaving 105 stimulus- (but not
797   language-) responsive electrodes.

798   **3. Neural dataset 3: fMRI (*Blank2014*).** We used the data from Blank et al. (2014) (n=5). (The set of participants includes 5 of
799   the 10 participants in Blank et al., 2014, because we wanted each participant to have been exposed to the same materials
800   and as many stories as possible; the 5 participants included here all heard eight stories.) Stimuli consisted of stories from
801   the publicly available Natural Stories Corpus (Futrell et al., 2018). These stories, adapted from existing texts (fairy tales and
802   short stories) were designed to be "deceptively naturalistic": they contained an over-representation of rare words and
803   syntactic constructions embedded in otherwise natural linguistic context. The stories were presented auditorily (each was
804   ~5 min in duration), and following each story, participants answered 6 comprehension questions (see Blank et al., 2014 for
805   details of the experimental procedure, data acquisition, and preprocessing).

806   *Functional localization*: As in the Pereira2018 dataset, data analyses were performed on fMRI BOLD signals extracted from
807   the language network. From each language-responsive voxel of each participant, the BOLD time-series for each story was
808   extracted. Across the eight stories, the BOLD time-series included 1,317 time-points (TRs, time of repetition; TR=2s and
809   corresponds to the time it takes to acquire the full set of slices through the brain). To align the neuroimaging data with the
810   story text, we first split the text into consecutive 2-second intervals (corresponding to the fMRI TRs) based on the auditory
811   recording; if a word straddled boundaries of intervals, it was assigned to the 2s interval in which that spoken word ended.
812   Each of the resulting intervals thus included a story "fragment", which could be a full short sentence, part of a longer
813   sentence, or a transition between the end of one sentence and the beginning of another. Due to the temporal resolution of
814   the HRF, whose peak's latency is 4-6 seconds, we assumed that each time-point in the BOLD signal represented activity
815   elicited by the text fragment that occurred 4s (i.e., 2 TRs) earlier.

816   We constructed a stimulus-response matrix by i) averaging the BOLD signals corresponding to each TR in each story across
817   the voxels within each ROI of each participant (averaging across the voxels within ROIs was done to increase the signal-to-
818   noise ratio), resulting in 1 data point per TR per language-responsive ROI of each participant (60 ROIs total across the 5
819   participants), and ii) concatenating all story fragments (1,317 'stimuli'), yielding a 1,317x60 matrix.

820

821   **4. Behavioral dataset: Self-paced reading (*Futrell2018*).** We used the data from Futrell et al. (2018) (n=179). (The set of
822   participants excludes 1 participant for whom data exclusions—see below—left only 6 data points or fewer.) Stimuli
823   consisted of ten stories from the Natural Stories Corpus (same materials as those used in *Blank2014*, plus two additional
824   stories), and any given participant read between 5 and all 10 stories. The stories were presented online (on Amazon's
825   Mechanical Turk platform) visually in a dashed moving window display—a standard approach in behavioral psycholinguistic
826   research (Just et al., 1982). In this approach, participants press a button to reveal each consecutive word of the sentence or
827   story; as they press the button again, the word they just saw gets converted to dashes again, and the next word is
828   uncovered. The time between button presses provides an estimate of overall language comprehension difficulty, and has
829   been shown to be robustly sensitive to both lexical and syntactic features of the stimuli (Grodner & Gibson, 2005; Smith &
830   Levy, 2013, inter alia) (see Futrell et al., 2018 for details of the experimental procedure and data acquisition.) We followed
831   data exclusion criteria in Futrell et al. (2018): for any given participant, we only included data for stories where they
832   answered 5 or all 6 comprehension questions correctly, and we excluded reading times (RTs) that were shorter than 100 ms
833   or longer than 3000 ms.

834

835 We constructed a stimulus-response matrix by i) obtaining the RTs for each word in each story for each participant (848,762
836 RTs total across the 179 participants; 338 average, ±173 std. dev.), and ii) concatenating all words in all sentences (10,256
837 words across 485 sentences), yielding a 10,256x179 matrix.
838
839 **5. Computational models.** We tested 43 language models that were selected to sample a broad range of computational
840 designs across three major types of architecture: embeddings, recurrent architectures, and attention-based 'transformer'
841 architectures. Here we provide a brief overview (see Table SI-10 for a summary of key features varying across the models).
842 **GloVe** (Pennington et al., 2014) is a word embedding model where embeddings are positioned based on co-occurrence in
843 the Common Crawl corpus; **ETM** (Dieng et al., 2019, 20ng dataset) combines word embeddings with an embedding of each
844 word's assigned topic; and **word2vec** (Mikolov et al., 2013)—abbreviated as w2v—provides embeddings which are trained
845 to guess a word based on its context. **lm_1b** (Jozefowicz et al., 2016) is a 2-layer long short-term memory (LSTM) model
846 trained to predict the next word in the One Billion Word Benchmark (Chelba et al., 2014); and the **skip-thoughts** model
847 (Kiros et al., 2015) is trained to reconstruct surrounding sentences in a passage. For all 38 transformer models (pretrained
848 models from the HuggingFace library (Wolf et al., 2019)), we only evaluate the encoder and not the decoder; the encoders
849 process long contexts (100s of words) with a deep neural network stack of multiple attention heads that operate in a feed-
850 forward manner (except the Transformer-XL-wt103 and the two XLNet models, which use recurrent processing), and differ
851 mostly in the choice of directionality, network architecture, and training corpora (Table SI-11). We highlight key features of
852 different classes of transformer models (BERT, RoBERTa, XLM, XLM-RoBERTa, Transformer-XL-wt103, XLNet, CTRL, T5,
853 AlBERT, and GPT) in the order in which they appear in the bar-plots (e.g., Fig. 2a), except for the three 'distilled' models
854 (Sanh et al., 2019), which we mention in the end. **BERT** transformers (Devlin et al., 2018) (n=4; bert-base-uncased, bert-
855 base-multilingual-cased, bert-large-uncased, bert-large-uncased-whole-word-masking) are optimized to train bidirectional
856 representations taking into account context both to the left and right of a masked token. **RoBERTa** transformers (Liu et al.,
857 2019) (n=2; roberta-base, roberta-large) as a variation of BERT improve training hyper-parameters such as masking tokens
858 dynamically instead of always masking the same token. **XLM** models (Lample & Conneau, 2019) (n=7; xlm-mlm-enfr-1024,
859 xlm-clm-enfr-1024, xlm-mlm-xnli15-1024, xlm-mlm-100-1280 , xlm-mlm-en2048) learn cross-lingual models by predicting
860 the next ("clm") or a masked ("mlm") token in a different language. **XLM-RoBERTa** (Conneau et al., 2019) (n=2; xlm-roberta-
861 base, xlm-roberta-large) combines RoBERTa masking with cross-lingual training in XLM. **Transformer-XL-wt103** (Dai et al.,
862 2020) adds a recurrence mechanism to GPT (see below) and trains on the smaller WikiText-103 corpus. **XLNet** transformers
863 (Yang et al., 2019) (n=2; xlnet-base-cased, xlnet-large-cased) permute tokens in a sentence to predict the next token. **CTRL**
864 (Keskar et al., 2019) adds control codes to GPT (see below) which influence text generation in a specific style. **T5**
865 transformers (Raffel et al., 2019) (n=5; t5-small, t5-base, t5-large, t5-3b, t5-11b) train the same model across a range of
866 tasks including the prediction of multiple corrupted tokens, GLUE (A. Wang, Singh, et al., 2019), and SuperGLUE (A. Wang,
867 Pruksachatkun, et al., 2019) in a text-to-text manner where the task is provided as a text prefix. **AlBERT** transformers (Lan et
868 al., 2019) (n=8; albert-base-v1, albert-large-v1, albert-xlarge-v1, albert-xxlarge-v1, albert-base-v2, albert-large-v2, albert-
869 xlarge-v2, albert-xxlarge-v2) use parameter-sharing and model inter-sentence coherence. **GPT** transformers (n=5) are
870 trained to predict the next token in a large dataset emphasizing document quality (openaigpt (Radford et al., 2018) on the
871 Book Corpus dataset, gpt2, gpt2-medium, gpt2-large, and gpt2-xl (Radford et al., 2019) on WebText). Finally, **distilled
872 versions** of models (Sanh et al., 2019) (n=3; distilbert-base-uncased, distilgpt2, distilroberta-base) train compressed models
873 on a larger teacher network.
874
875 To retrieve model representations, we treated each model as an experimental participant (Figure 1) and ran the same
876 experiment on it that was run on humans. Specifically, sentences were fed in sequentially into the model (for Pereira2018,
877 Blank2014, and Futrell2018, sentences were grouped by passage / story to mimic the procedure with human participants).
878 For embedding and recurrent models, sentences were fed in word-by-word; for transformers, the context before (but not
879 after) each word was also fed into the models due to their lack of memory; the length of the context was determined by the
880 models' architectures. For recurrent models, the memory was reset after each paragraph (*Pereira2018*), each sentence
881 (*Fedorenko2016*), or each story (*Blank2014* and *Futrell2018*).
882
883 After the processing of each word, we retrieved ("recorded") model representations at every computational block (e.g., one
884 LSTM cell or one Transformer encoder block). (Word-by-word processing increases computational cost but is necessary to
885 avoid bidirectional models, like the BERT transformers, seeing the future.) When comparing against human recordings

22

886   spanning more than one word such as a sentence (*Pereira2018*) or story fragment (*Blank2014*), we aggregated model
887   representations: for the embedding models, we used the mean of the word representations; for recurrent and transformer
888   models, we used the representation of the last word since these models already aggregate representations of the preceding
889   context, up to a maximum context length of 512 tokens.
890
891   **6. Comparison of models to brain measurements.** We treated the model representation at each layer separately and
892   tested how well it could predict human recordings (for *Pereira2018*, we treated the two experiments separately, but
893   averaged the results across experiments for all plots except Fig. 2c). To generate predictions, we used 80% of the stimuli
894   (sentences in *Pereira2018*, words in *Fedorenko2016* and *Futrell2018*, and story fragments in *Blank2014*; Fig. 1) to fit a linear
895   regression from the corresponding 80% of model representations to the corresponding 80% of human recordings. We
896   applied the regression on model representations of the held-out 20% of stimuli to generate model predictions, which we
897   then compared against the held-out 20% of human recordings with a Pearson correlation. This process was repeated five
898   times, leaving out different 20% of stimuli each time, and we computed the per-voxel/electrode/ROI mean predictivity
899   across those five splits. We aggregated these per-voxel/electrode/ROI scores by taking the median of scores for each
900   participant's voxels/electrodes/ROIs and then computing the median and median absolute deviation (m.a.d.) across
901   participants (over per-participant scores). Finally, this score was divided by the estimated ceiling value (see Estimation of
902   ceiling below) to yield a final score in the range of $[0, 1]$. We report the results for the best-performing layer for each model
903   (SI-12).
904   **7. Estimation of ceiling.** Due to intrinsic noise in biological measurements, we estimated a ceiling value to reflect how well
905   the best possible model of an average human could perform. To do so, we first subsampled—for each dataset separately—
906   the data with n recorded participants into all possible combinations of s participants for all $s \in [2, n]$ (e.g. {2, 3, 4, 5} for
907   *Fedorenko2016* with n=5 participants). For each subsample s, we then designated a random participant as the target that
908   we attempt to predict from the remaining $s - 1$ participants (e.g., predict 1 subject from 1 (other) subject, 1 from 2
909   subjects, …, 1 from 4, to obtain a mean score for each voxel/electrode/ROI in that subsample. To extrapolate to infinitely
910   many humans and thus to obtain the highest possible (most conservative) estimate, we fit the equation $v = v_0 \times \left(1 - \right.$
911   $\left. e^{-\frac{x}{\tau_0}} \right)$ where x is each subsample's number of participants, v is each subsample's correlation score and v_0 and τ_0 are the
912   fitted parameters for asymptote and slope respectively. This fitting was performed for each voxel/electrode/ROI
913   independently with 100 bootstraps each to estimate the variance where each bootstrap draws x and v with replacement.
914   The final ceiling value was the median of the per-voxel/electrode/ROI ceilings v_0.
915   For *Fedorenko2016*, a ceiling was estimated for each electrode in each participant, so each electrode's raw value was
916   divided by its own ceiling value. Similarly, for *Blank2014*, a ceiling was estimated for each ROI in each participant, so each
917   ROI's raw value was divided by its own ceiling value. For *Pereira2018*, we treated the two experiments separately, focusing
918   on the 5 participants that completed both experiments to obtain full overlap in the materials for each participant, and used
919   10 random sub-samples to keep the computational cost manageable. A ceiling was estimated for all voxels in the 5
920   participants who participated in both experiments. Each voxel's raw predictivity value was divided by the average ceiling
921   estimate (across all the voxels for which it was estimated). For *Futrell2018*, given the large number of participants and
922   because most participants only had measurements for a subset of the stimuli, we did not hold out one participant but
923   rather tested how well the mean RTs for one half of the participants predicted the RTs for the other half of participants. We
924   further took 5 random subsamples at every 5 participants, starting from 1, and built 3 random split-halves, again to keep
925   computational cost manageable. A ceiling was estimated for each participant, and each participant's raw values were
926   divided by this ceiling. (Note that this approach is even more conservative than the leave-one-out approach, because split-
927   half correlations tend to be higher than one-vs.-rest, due to a reduction in noise when averaging (for each half).)
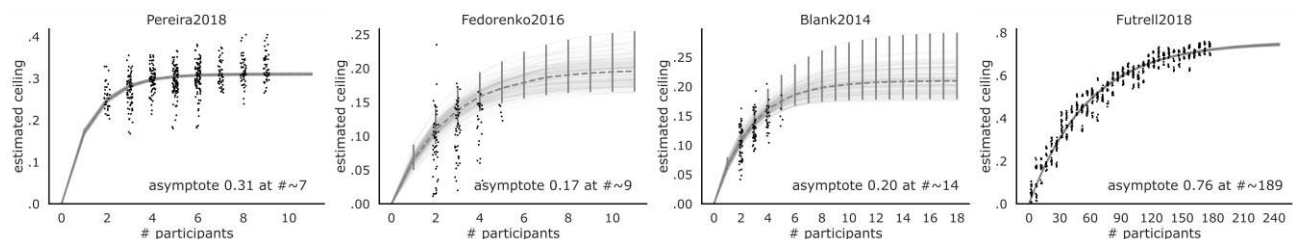928
929   **8. Language Modeling.** To assess the models' performance on the normative next-word-prediction task, we used a dataset
930   of 720 Wikipedia articles, WikiText-2 (Merity et al., 2016), with 2M training, 218k validation, and 246k test tokens (words
931   and word-parts). These tokens were processed by model-specific tokenization with a maximum vocabulary size of 250k,
932   selected based on the tokens' frequency in the model's original training dataset, and split up into blocks of 32 tokens each
933   (both the vocabulary size and the length of blocks were constrained by computational cost limitations). We sequentially

23

934   fed the tokens into models as explained in <u>Computational Models</u> and captured representations at each step from each
935   model's final layer. To predict the next word, we fit a linear decoder from those representations to the next token over
936   words in the vocabulary (n=50k), on the training tokens. This decoder is trained with a cross-entropy-loss $L =$
937    $-\sum_c^C t_c^i \log\left(\frac{e^{s_c^i}}{\sum_d^C e^{s_d^i}}\right)$ where t_c^i is the true label for class c and sample i, and s_c^i is the predicted probability of that class; the
938   linear weights are updated with AdamW and a learning rate of 5e-5 in batches of 4 blocks until convergence as defined on
939   the validation set. Importantly, note that we only trained weights of a readout decoder, *not* the weights of models
940   themselves, in order to maintain the same model representations that we used in model-to-brain and model-to-behavior
941   comparisons. The final language modeling score is reported for each model as the perplexity, i.e. the exponent of the
942   cross-entropy loss, on the held-out test set.
943
944

Supplement

954

955 Figure S1: **Ceiling estimates for neural and behavioral datasets.** Due to intrinsic noise in biological measurements, we
956 estimated a ceiling value to reflect how well the best possible model of an average human could perform, based on sub-
957 samples of the total set of participants (see Methods-7). For each sub-sample, $s - 1$ participants are used to predict a held-
958 out participant (except in *Futrell2018*, where this is done on split-halves, as described in the text). Each dot represents a
959 correlation between the average of the $s - 1$ participants and the left-out participant for a random sub-sample of the
960 number of participants s indicated on the x-axis. We then bootstrapped 100 random combinations of those dots to
961 extrapolate (gray lines) the highest possible ceiling if we had an infinite number of participants at our disposal. The
962 parameters of these bootstraps are then aggregated by taking the median to compute an overall estimated ceiling (dashed
963 gray line with 95% CI in error-bars). We use this estimated ceiling to normalize model scores and here also report the
964 number of participants at which the estimated ceiling would be met (which show that for *Pereira2018* and *Futrell2018,* the
965 number of participants we have is at and close to the asymptote value, respectively).

966
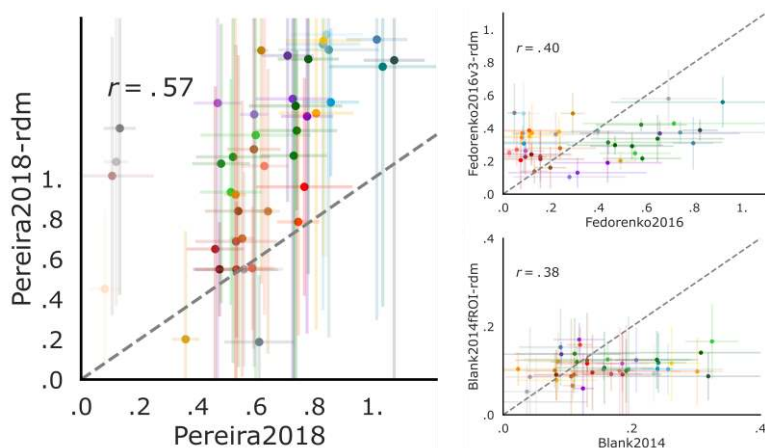
967 Figure S2: **Scores generalize across metrics.** Model scores on each dataset generalize across different choices of a similarity
968 metric; here we plot the predictivity metric used in the manuscript on the x-axis against a model-to-brain similarity metric
969 based on representational dissimilarity matrices (RDMs) between models and neural representations on the y-axis. Like in
970 the predictivity metric, stimuli along with corresponding model activations and brain recordings were split 5-fold but we
971 then only compared the respective test splits given that the RDM metric does not employ fitting. Specifically, we followed
972 (Kriegeskorte, 2008) and computed the RDM for each model's activations, and a separate RDM for each brain recording
973 dataset, based on 1 minus the Pearson correlation coefficient between pairs of stimuli; then, we measured model-brain
974 similarity via Spearman correlation across the two RDMs' upper triangles. The RDM score for one model on one human
975 dataset is then the mean over splits. We ran each model and compared resulting scores with the primarily used scores from
976 the predictivity metric. Correlations for models' scores between the predictivity and the RDM metrics are: Pereira2018
977 r=.57, p<0.0001; Fedorenko2016 r=.40, p<.01; Blank2014 r=.38, p<.05.

978
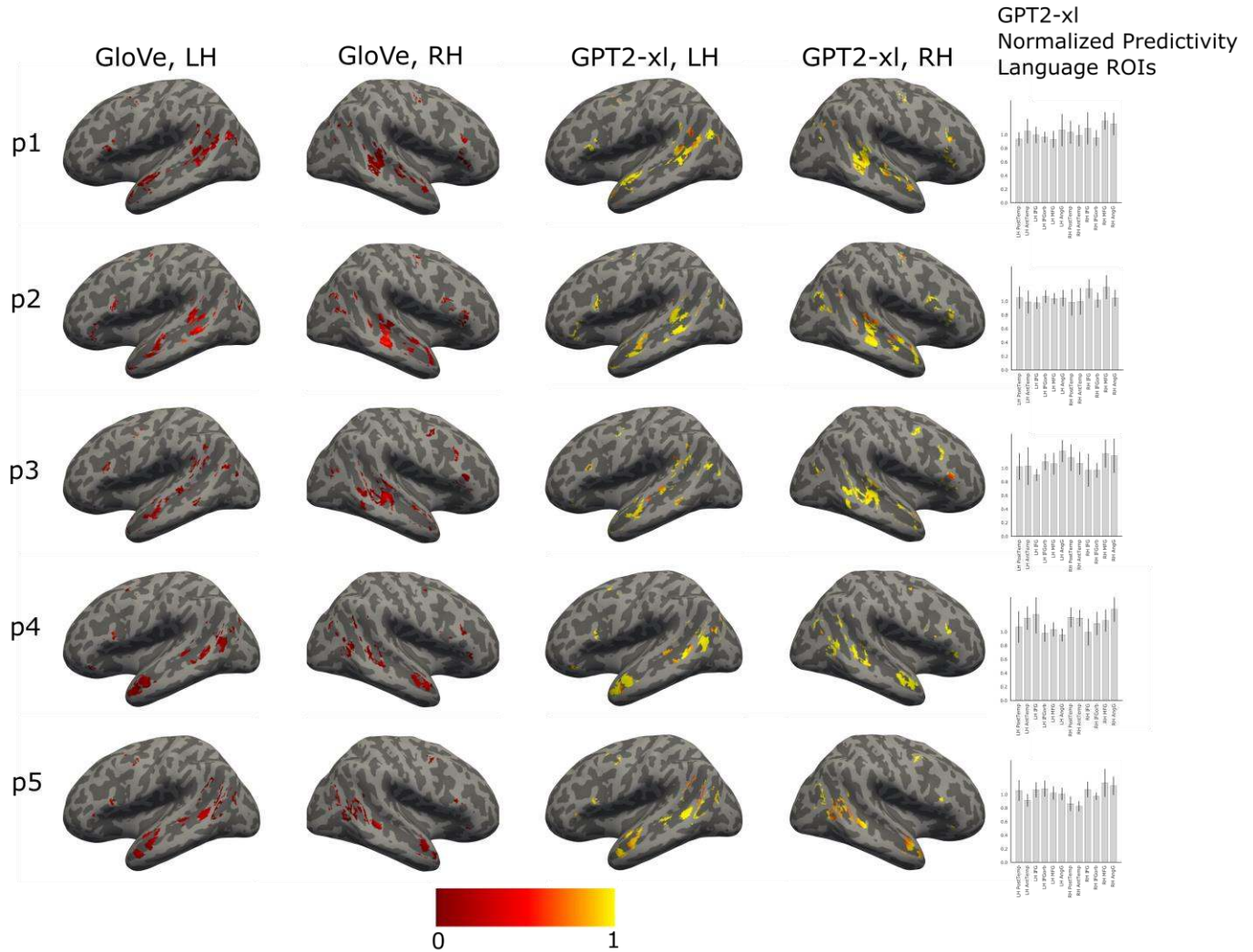
Figure S3: **Brain surface visualization of model predictivity scores.** Plots show surface projections of volumetric individual language-responsive functional ROIs in the left and right hemispheres (LH and RH) for five representative participants from *Pereira2018.* In each voxel of each fROI, we show a normalized predictivity value for two models that differ substantially in their ability to predict human data: GloVe (first two columns) and GPT2-xl (second two columns; for GPT2-xl, we show predictivity values from the overall best-performing layer, in line with how we report the results in the main text). (Note that the voxel locations are identical between GloVe and GPT2-xl, and are determined by an independent functional language localizer as described in the text; we here illustrate the differences in predictivity values, along with showing sample fROIs used in our analyses). Predictivity values were ceiling-normalized for each participant and each of 12 ROIs separately (a slight deviation from the approach in the main analysis, which was designed to control for between-region differences in reliability). The data were analyzed in the volume space and co-registered using SPM12 to Freesurfer's standard brain CVS35 (combined volumetric and surface-based (CVS)) in the MNI152 space using nearest neighbor interpolation and no smoothing. The ceiled predictivity maps for the language localizer contrast (10% of most language-responsive voxels in each 'mask'; Methods-1) were projected onto the cortical surface using mri_vol2surf in Freesurfer v6.0.0 with a projection fraction of 1. The surface projections were visualized on an inflated brain in the MNI152 space using the developer version of Freeview (assembly March 10th, 2020). The bar plots in the rightmost column show the normalized predictivity values per ROI (median across voxels) in the language network for GPT2-xl. Error bars denote m.a.d. across voxels. The distribution of predictivity values across the language-responsive voxels, and the similar predictivity magnitudes across the ROIs in the bar graphs, both suggest that the results (between-model differences in neural scores) are not driven by one particular region of the language network, but are similar across regions, and between the LH and RH components of the network (see also SI-4).

26

999

1000 **SI-4 – Language specificity**

1001 In the analyses reported in the manuscript, we focused on the language-responsive regions / electrodes. Here, for two
1002 datasets, we investigated the model-brain relationship outside the language network in order to assess the spatial
1003 specificity of our results, i.e., to test whether they obtain only, or more strongly, in the language network compared to
1004 other parts of the brain. For both datasets, we report analyses based on *raw predictivity values*, without normalizing by the
1005 estimated noise ceiling because the brain regions of the language network differ from other parts of the brain in how
1006 strongly their activity is tied to stimulus properties during comprehension (e.g., I. A. Blank & Fedorenko, 2017, 2020;
1007 Diachek et al., 2020; Shain et al., 2020; Wehbe et al., 2020). This variability is important to take into account when
1008 comparing between functionally different brain regions/electrodes because we are interested in how well the models
1009 explain linguistic-stimulus-related neural activity. When we normalize the neural responses of a non-language-responsive
1010 region/electrode using a language comprehension task, we're effectively isolating whatever little *stimulus-related activity*
1011 this region/electrode may exhibit, putting them on ~equal or similar footing with the language-responsive
1012 regions/electrodes. (For completeness and ease of comparison with the main analyses, we also report analyses based on
1013 normalized predictivity values.)

1014

1015 *Fedorenko2016:* The scores obtained from language-responsive electrodes were compared to those obtained from
1016 stimulus-responsive electrodes, excluding the language-responsive ones (see Methods-2), for all 43 models. The number of
1017 language-responsive electrodes across five participants was 97, and the number of stimulus-, but not language-, responsive
1018 electrodes across the participants was comparable (n=105). The analysis was identical to the main analysis (see Methods),
1019 besides omitting the ceiling normalization for the raw predictivity analyses. As described in Methods, normalization was
1020 performed for each electrode in each participant separately.
1021 For raw predictivity, neural responses in the language-responsive electrodes were predicted 49.21% better on average
1022 across models than the non-language-responsive electrodes (independent-samples two-tailed t-test: t=3.4, p=0.001). (For
1023 normalized predictivity, neural responses in the language-responsive electrodes were predicted 59.26% better on average
1024 across models than the non-language-responsive electrodes (t=2.24, p=0.03).)

1025

1026 *Pereira2018:* The scores obtained from the language network were compared to those obtained from two control
1027 networks: the multiple demand (MD) network and the default mode network (DMN) (see Methods), for all 43 models. The
1028 number of voxels in the language network across participants was, on average, 1,355 (± 7 SD across participants), and the
1029 average number of voxels in the MD network and the DMN was comparable (MD: 2,994±230); DMN: 1,098±7). The analysis
1030 was identical to the main analysis (see Methods), besides omitting the ceiling normalization for the raw predictivity
1031 analyses. For the normalized predictivity analyses, the network predictivity values were normalized by their respective
1032 network ceiling values.
1033 For raw predictivity, neural responses in the language network ROIs were predicted 16.96% better on average across
1034 models than the MD network ROIs (independent-samples two-tailed t-test: t=2.26, p=0.03) and numerically (14.33%) better
1035 than the DMN ROIs (t=1.78, p=0.08). (For normalized predictivity, neural responses in the language network ROIs were
1036 predicted numerically (6.47%) worse on average than the MD network ROIs (t=-0.92, p=0.36) and also numerically (1.05%)
1037 worse than the DMN ROIs (t=-0.31, p=0.76).)

1038

1039 These results suggest that—when allowing for inter-regional differences in the reliability of language-related responses—
1040 the model-to-brain relationship is stronger in the language-responsive regions/electrodes. However, we leave open the
1041 possibility that language models also explain neural responses outside the boundaries of the language network, perhaps
1042 because these models capture some parts of our general semantic knowledge, which is plausibly stored in a distributed
1043 fashion across the brain. For example, several earlier studies used simple embedding models to decode linguistic meaning
1044 from fMRI data (e.g., Wehbe et al., 2014; Huth et al., 2016; Anderson et al., 2017; Pereira et al., 2018) and reported reliable
1045 decoding not only within the language network, but also across other parts of association cortex. Given that we know that
1046 different large-scale cortical networks differ functionally in important ways (e.g., see Fedorenko & Blank, 2020, for a recent
1047 discussion of the language vs. MD networks), it will be important to investigate in future work the precise mapping between
1048 the language models' representations and neural responses in these different functional networks.

27

1049

1050 **SI-5 – Model performance on diverse language tasks vs. model-to-brain fit**

1051 To test whether the next-word prediction task is special in predicting model-to-brain fit, we used the *Pereira2018* dataset to
1052 examine the relationship between the models' performance on diverse language processing tasks from the General
1053 Language Understanding Evaluation (GLUE) benchmarks (Wang et al., 2018) and neural predictivity. We used a subset of the
1054 high-performing, transformer models (n=30 of the 38 where we could find published commitments of which features to use
1055 for GLUE). The GLUE benchmark encompasses nine tasks that can be classified into three categories: single-sentence
1056 judgment tasks (n=2), sentence-pair semantic similarity judgment tasks (n=3), and sentence-pair inference tasks (n=4). The
1057 two single-sentence tasks are both binary classification tasks: models are asked to determine whether a given sentence is
1058 grammatical or ungrammatical (Corpus of Linguistic Acceptability, *CoLA* (Warstadt et al., 2018)), or whether the sentiment
1059 of a sentence is positive or negative (Stanford Sentiment Treebank, *SST-2* (Socher et al., 2013)). In the semantic similarity
1060 tasks, models are asked to assert or deny the semantic equivalence of question pairs (Quora Question Pairs, *QQP* (Chen et
1061 al., 2018)) or sentence pairs (Microsoft Research Paraphrase Corpus, *MRPC* (Dolan & Brockett, 2005)), or to judge the
1062 degree of semantic similarity between two sentences on a scale of 1-5 (Semantic Textual Similarity Benchmark, *STS-B* (Cer
1063 et al., 2017)). Lastly, the benchmark contains four inference tasks, of which we include three (following Devlin et al., 2018),
1064 we exclude the Winograd Natual Language Inference, *WNLI*, task; see (12) in https://gluebenchmark.com/faq). In two of
1065 these tasks, models are asked to determine the entailment relationship between sentences in a pair using either tertiary
1066 classification: entailment, contradiction, neutral (Multi-Genre Natural Language Inference corpus, *MNLI* (Williams et al.,
1067 2018)), or binary classification: entailment or no entailment (Recognizing Textual Entailment, *RTE* (Dagan et al., 2006, Bar
1068 Haim et al., 2006, Giampiccolo et al., 2007, Bentivogli et al., 2009)). And in the third inference task, the Question Natural
1069 Language Inference, *QNLI,* task (Rajpurkar et al., 2016, White et al., 2017, Demszky et al., 2018), models are presented with
1070 question-answer pairs and asked to decide whether or not the answer-sentence contains the answer to the question.

1071 In order to evaluate model performance on GLUE benchmark tasks, each GLUE dataset was first converted into a format
1072 that is compatible with transformer model input using functionality from the GLUE data processor provided by Huggingface
1073 transformers (https://huggingface.co/transformers/). In particular, each set of materials is represented as a matrix that
1074 includes the following dimensions: item (and sentence for multi-sentence materials) ID, ID for each individual word (with
1075 reference to the vocabulary used by the transformer models), the label (e.g., grammatical vs. ungrammatical), and the
1076 'attention mask' which specifies which part(s) of the sentences the model should pay attention to (e.g., some 'padding' is
1077 commonly used to equalize the lengths of sentences/items to the target length of 128 tokens (again constrained by
1078 computational cost), and the attention mask is set to include only the actual words in the materials, and not the padding,
1079 and in some models to further constain which parts of the input to attend to—e.g., in GPT2 models, the rightward context is
1080 ignored). Next, each GLUE dataset was then fed into each model to obtain a sequence of hidden states at the output of the
1081 last layer of the model. Following default settings from Huggingface transformers, from these hidden states, we then
1082 extracted the token of interest: for bidirectional models such as BERT, this was the first input token—a special token ([cls])
1083 that is appended to each item and designed for sequence classification tasks, and for unidirectional models such as GPT-2,
1084 XLNet or CTRL, this token corresponded to the last attended token (e.g., the last word/word-part in the sentence). In order
1085 to ensure a fair comparison between the models and to avoid the skewing of representations by individual task pre-training,
1086 dense linear pooling projection layers (specific to some transformer) are disregarded. Finally, we fit a linear decoder from
1087 the features of the extracted tokens of interest to the task label(s). For tasks with two or more labels, a cross-entropy loss
1088 function is used; for the task that uses a rating scale, the decoder is trained with a mean-square error (MSE) loss function.
1089 Similar to the next-word prediction task, the linear weights are updated with the AdamW optimizer and a learning rate of
1090 5e-5 in batches of 8 blocks until convergence as defined on the validation set. Importantly, and also similar to the next-
1091 word-prediction task, we only trained weights of a readout decoder, *not* the weights of models themselves, in order to
1092 maintain the same model representations that we used in model-to-brain and model-to-behavior comparisons. To account
1093 for potential bias in the GLUE datasets, multiple metrics within tasks, as well as different metrics across tasks are reported
1094 in the GLUE benchmark. Following standards in the field, we report the final task score as accuracy for *SST-2, QQP, MRPC,*

1095   *MNLI*, *RTE,* and *QNLI*, Matthew's Correlation for *CoLA*, and Pearson correlation for *STS-B*. The results are shown in Fig. S5.
1096   None of the tasks significantly predicted neural scores, suggesting that next-word prediction may be special in its ability to
1097   predict brain-like processing.

1098
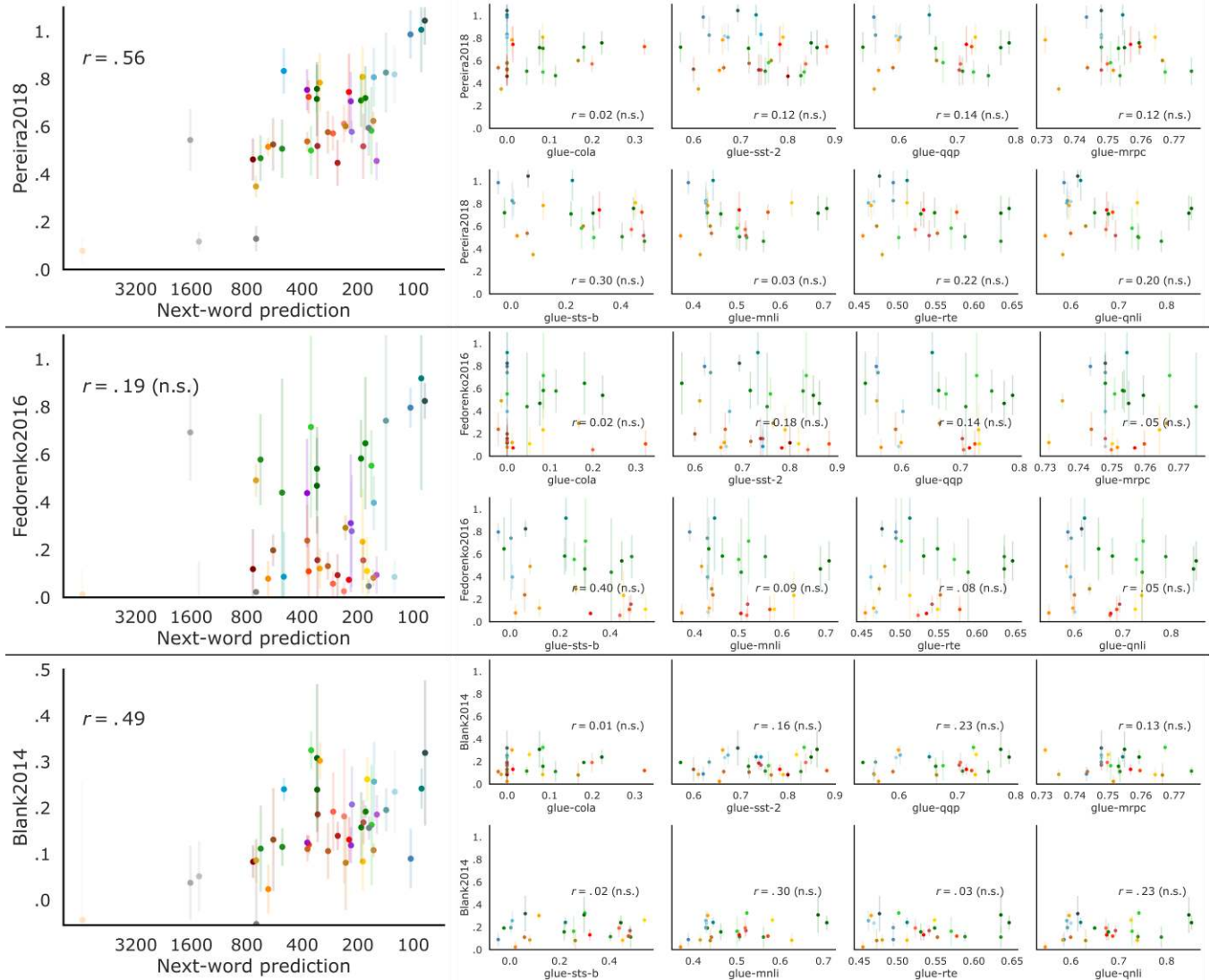
1099   Figure S5: **Performance on next-word prediction selectively predicts model-to-brain fit**. Performance on GLUE tasks was
1100   evaluated as described in SI-5. Only the next-word prediction correlations but none of the GLUE correlations were
1101   significant.

1102
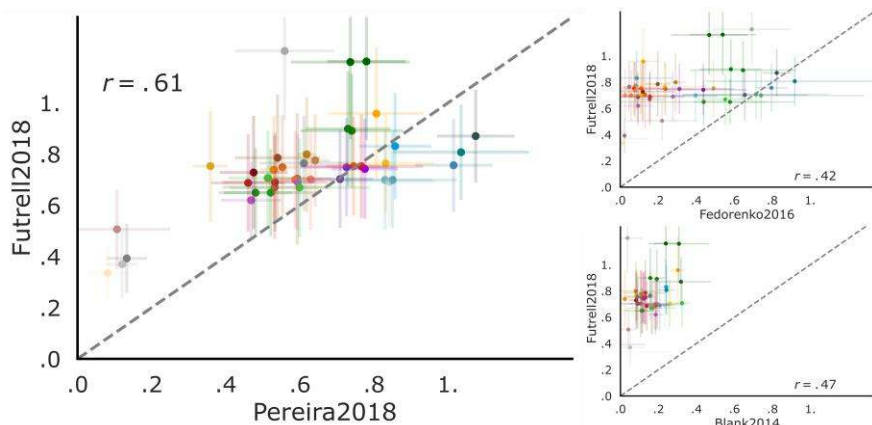
Figure S6: **Models' neural predictivity for each dataset is correlated with behavioral predictivity.** In Fig. 4b, we showed that the models' neural predictivity (averaged across the three neural datasets: Pereira2018, Fedorenko2016, Blank2014) correlates with behavioral predictivity. Here, we show that this relationship also holds for each neural dataset individually: Pereira2018: p<0.0001, Fedorenko2016: p<0.01, Blank2014: p<0.01.
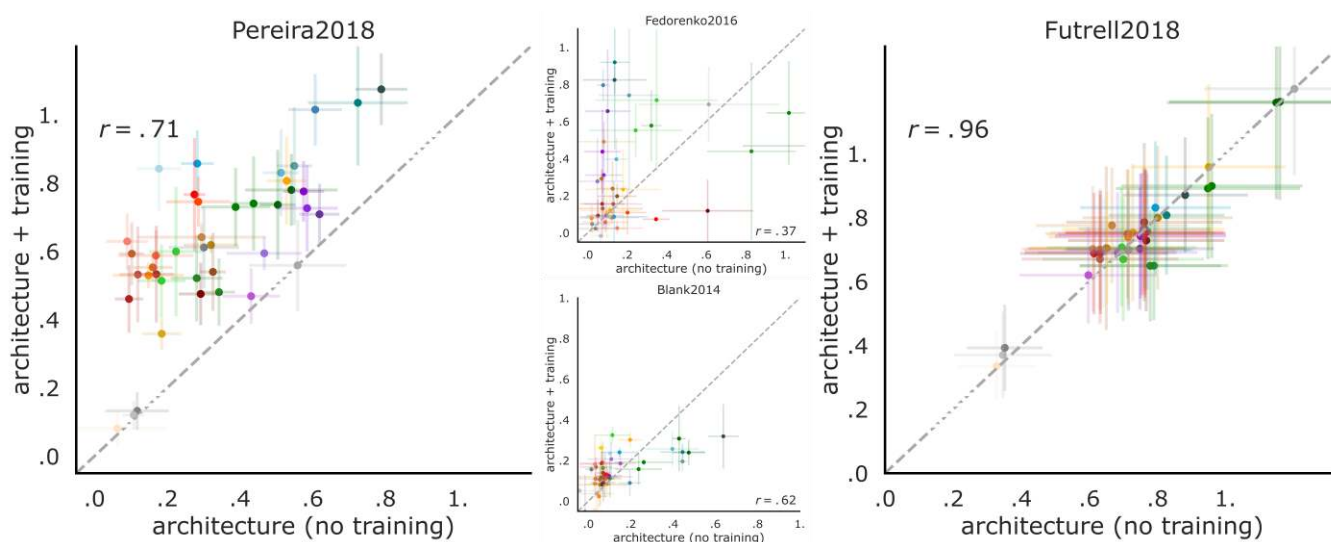
Figure S7: **Model architecture alone already yields predictive representations and untrained performance predicts trained performance.** In Fig. 5, we showed that untrained models already achieve robust brain predictivity (averaged across the three neural and one behavioral datasets). Here, we show that this relationship also holds for each dataset individually: Pereira2018: p<<0.00001, Fedorenko2016: p<0.05, Blank2014: p<0.00001, Futrell2018: p<<0.00001.
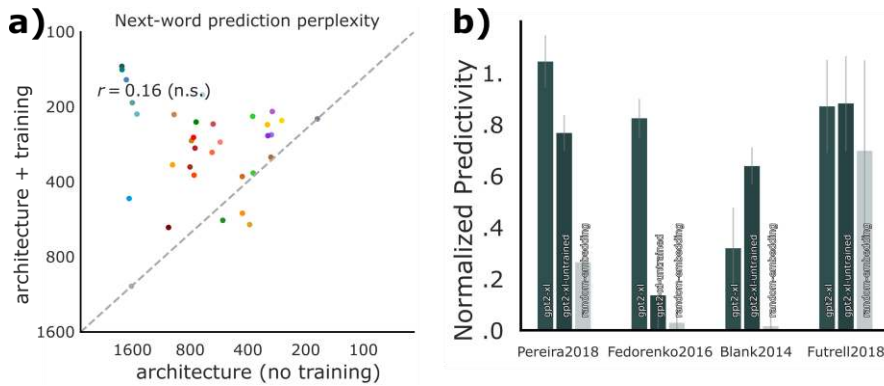
1118 Figure S8: **Performance of models with random weights depends on architecture.** a) The relationship between model
1119 performance with vs. without training on the wikitext-2 next-word-prediction task. Consistent with model performance with
1120 vs. without training on neural and behavioral datasets (Fig. 5), untrained models perform reasonably well. Training improves
1121 scores by 80% on average, and most prominently for GPT models, in teal (where the quality of the training data is
1122 optimized; see Computational models in Methods). b) Neural and behavioral scores of GPT2-xl, the best-performing model,
1123 with vs. without training, and of a random embedding of the same size. Embedding size alone is not sufficient: a random
1124 embedding matched in size to GPT2-xl scores worse than untrained GPT2-xl in all four datasets (3 neural, and 1 behavioral).
1125 These results suggest that model architecture critically contributes to model-to-brain and model-to-behavior fits.

1126

1127 **SI-9 – Effects of model architecture and training on neural and behavioral scores**

1128

1129 The 43 language models included in the current study span three major types of architecture: embedding models, recurrent
1130 models, and attention-based transformer architectures. However, in addition to this coarse distinction, the individual
1131 models vary widely in diverse architectural and training features. A rigorous examination of the effects of different model
1132 features on model-to-brain/behavior fit would require careful pairwise comparisons of minimally different models, which is
1133 not possible for 'off-the-shelf' models without extremely expensive re-training from scratch under many/all possible
1134 combinations of architecture, training diet, optimization objective, and other hyper-parameters. However, we here
1135 undertook a preliminary exploratory investigation. In particular, for a subset of model features (Table SI-9), we computed a
1136 Pearson correlation between the feature values and the averaged model score across all four datasets (3 neural, and 1
1137 behavioral). We included five architectural features. Three features were continuous: i) number of hidden layers, which
1138 varied between 1 and 48 (mean 16.02, std. dev. 11.02); ii) number of features (units across considered layers), which varied
1139 between 300 and 78,400 (mean 20,971.26, std. dev. 18,362.91); and the size of the embedding layer, which varied between
1140 128 and 48,000 (mean 872.28, std. dev. 744.33). And the remaining two features were binary: iv) uni- vs. bi-directionality
1141 (32/43 models were bi-directional), and v) the presence of recurrence (5/43 models had recurrence). And we included two
1142 training-related features: i) training data size (in GB), which varied between 0.2 and 336 (mean 351.06 std. dev. 726.81); and
1143 ii) vocabulary size, which varied between 30,000 and 3,000,000 (mean 223,096.95 std. dev. 561,737.36). All training data
1144 numbers were taken from the original model papers, and if training data was specified in tokens, a conversion rate of 4
1145 bytes per token was used. We further excluded the multilingual XLM and BERT models when examining the effect of
1146 training data size, because those numbers could not be confidently verified. For comparison, we also included performance
1147 on the next-word-prediction task that we examined in the main text.

1148

1149 The results are shown in Fig. S9. As expected—given the results reported in the main text for the individual datasets (Fig. 3,
1150 4c)—next-word prediction performance robustly predicts model-to-brain/behavior fit (r = 0.49, p < 0.01). These results
1151 suggest that optimizing for predictive representations may be a critical shared feature of biological and artificial neural
1152 networks for language. How do architectural and training-related features compare to next-word-prediction task
1153 performance in their effect on neural/behavioral predictivity? Two architectural size features are most correlated with
1154 model performance: number of hidden layers (r = 0.56, p < 0.001), and number of features (r = 0.68, p << 0.0001). This is
1155 expected given that the most recent models with the highest performance on linguistic tasks are also the largest ones that
1156 researchers are able to run on modern hardware. The two training-related features—training data size and vocabulary

31

1157   size—are significantly *negatively* correlated with model performance. To rule out the possibility that the negative effect of
1158   training-related features is driven by models with relatively small training datasets and vocabulary size (e.g., ETM; Table
1159   S10) that have low brain/behavior predictivity, we ran an additional analysis considering only transformer models (n=38):
1160   even in these generally highly predictive models, more training data (r = -0.29, p = 0.11 [not plotted]) or larger vocabulary
1161   size (r = -0.21, p = 0.25 [not plotted]) do not appear to be beneficial, although the negative correlations are non-significant.

1162

1163   Does the collection of model designs investigated in this paper inform the hyperparameters that should be optimized for in
1164   any new model to achieve high predictivity? To provide a preliminary answer to this question, we performed an exploratory
1165   analysis in the form of stepwise forward model selection and examined (a) the most parsimonious model that explains the
1166   data, and (b) how much variance the selected features explain cumulatively (Fig. S9b). High overall explained variance
1167   indicates that the combination of features selected by the model is predictive of model performance, whereas low overall
1168   explained variance indicates that crucial predictive hyperparameters are still being neglected. In the forward regression
1169   analysis, we add predictors based on the highest R^2-adjusted value of the new model, as long as variance increases by
1170   adding a new factor. This analysis revealed that adding training dataset size and recurrence does not lead to variance
1171   increase. Significance markers indicate the p-value for significance of adding each term, and for each regression step we
1172   plot the added explained variance (in R^2-adjusted) of the variable chosen by the model. The overall cumulative R^2-adjusted
1173   value of the selected model is 0.822.

1174
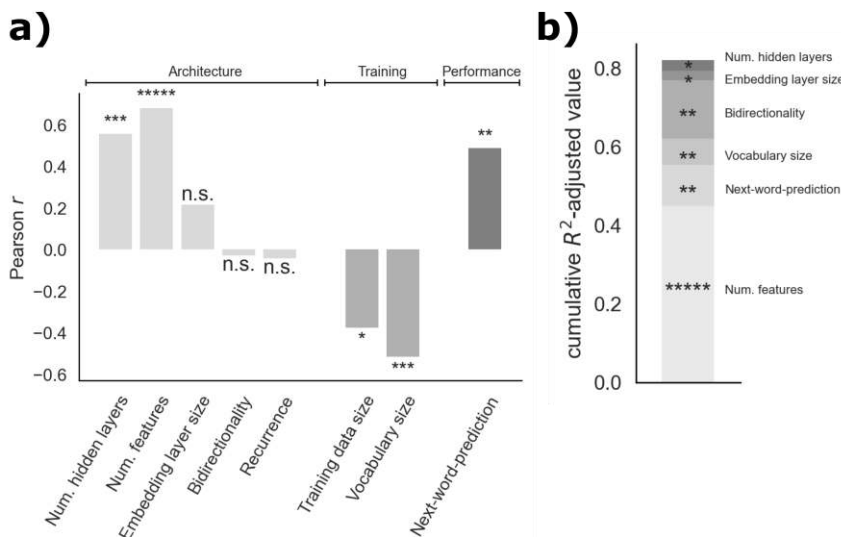
1175   Figure S9: **Effects of model architecture vs. training on neural and behavioral scores. a)** We compared the effects on neural
1176   and behavioral scores (the averaged model score across all four datasets) of three kinds of features: (i) architectural
1177   properties, (ii) training-dependent variables, and, for comparison, iii) performance on the next-word-prediction task
1178   examined in the main text (Fig. 3, 4c). **b**) Alternative combination of predictors with stepwise forward regression model.
1179   New predictors are added based on the highest R^2-adjusted value of the new model, as long as variance increases by adding
1180   a new factor (thus excluding training dataset size and recurrence). Significance markers indicate the p-value for significance
1181   of adding model terms. For each regression step, we plot the added explained variance (in R^2-adjusted) of the variable
1182   chosen by the model. The overall cumulative R^2-adjusted value of the selected model is 0.822. As in a), the preferred
1183   explanatory variable is the number of features. Stepwise forward regression based on significance leads to the same model-
1184   choice. Note that, as above, t5-11b is excluded for regression based on next-word-prediction, and multilingual models are
1185   excluded for regression on training size.

1186

1187
1188

32

	Model identifier	Architecture class	Num. layers	Num. features	Embedding layer size	Bidirec-tional	Recurrent	Training data size	Vocabulary size	Tokenization	Training tasks
1	glove	Embedding	1	300	300	0	0	3360	2200000	Stanford tokenizer	Learning word vectors such that their dot product equals the logarithm of the words' probability of co-occurrence
2	ETM	Embedding	1	300	300	0	0	0.2	52258	RegEx word-level tokenizer	Variational inference topic modeling using embedding representations of both words and topics
3	word2vec	Embedding	1	300	300	0	0	400	3000000	Word-level tokenizer	Predicting a center word from the surrounding context
4	lstm_lm_1b	Recurrent	2	2048	1024	0	1	4	793471	bBPE	Causal Language Modeling
5	skip-thoughts	Recurrent	1	4800	4800	0	1	3	930911	NLTK tokenizer	Predicting words in neighboring sentences
6	distilbert-base-uncased	Bidir. transf.	6	5376	768	1	0	13	30522	WordPiece	
7	bert-base-uncased	Bidir. transf.	12	9984	768	1	0	13	30522	WordPiece	
8	bert-base-multilingual-cased	Bidir. transf.	12	9984	768	1	0	n.a.	119547	WordPiece	Masked Language Modeling Next-Sentence Prediction
9	bert-large-uncased	Bidir. transf.	24	25600	1024	1	0	13	30522	WordPiece	
10	bert-large-uncased-whole-word-masking	Bidir. transf.	24	25600	1024	1	0	13	30522	WordPiece	
11	distilroberta-base	Bidir. transf.	6	5376	768	1	0	161	50265	bBPE	
12	roberta-base	Bidir. transf.	12	9984	768	1	0	161	50265	bBPE	dynamic Masked Language Modeling
13	roberta-large	Bidir. transf.	24	25600	1024	1	0	161	50265	bBPE	
14	xlm-mlm-enfr-1024	Bidir. transf.	6	7168	1024	1	0	n.a.	64139	BPE	multilingual Masked Language Modeling
15	xlm-clm-enfr-1024	Bidir. transf.	6	7168	1024	1	0	n.a.	64139	BPE	multilingual Causal Language Modeling
16	xlm-mlm-xnli15-1024	Bidir. transf.	12	13312	1024	1	0	n.a.	95000	BPE	multilingual Masked Language Modeling
17	xlm-mlm-100-1280	Bidir. transf.	16	21760	1280	1	0	n.a.	200000	BPE	
18	xlm-mlm-en-2048	Bidir. transf.	12	26624	2048	1	0	16	30145	BPE	Masked Language Modeling
19	xlm-roberta-base	Bidir. transf.	12	9984	768	1	0	2500	250002	SentencePiece	multilingual Masked Language Modeling
20	xlm-roberta-large	Bidir. transf.	25	25600	1024	1	0	2500	250002	SentencePiece	
21	transfo-xl-wt103	Bidir. transf.	18	19456	1024	1	1	0.4	267735	Word-level tokenizer	Causal Language Modeling
22	xlnet-base-cased	Bidir. transf.	12	9984	768	1	1	126	32000	SentencePiece	Permutation Language Modeling
23	xlnet-large-cased	Bidir. transf.	24	25600	1024	1	1	126	32000	SentencePiece	
24	ctrl	Bidir. transf.	48	62720	1280	1	0	140	246534	BPE	Causal Language Modeling
25	t5-small	Bidir. transf.	6	3584	512	1	0	862	32128	SentencePiece	Text-to-text training on a variety of tasks (i.a., prediction of multiple corrupted tokens, and tasks from the GLUE and SuperGLUE benchmarks)
26	t5-base	Bidir. transf.	12	9984	768	1	0	862	32128	SentencePiece	
27	t5-large	Bidir. transf.	24	25600	1024	1	0	862	32128	SentencePiece	
28	t5-3b	Bidir. transf.	24	25600	1024	1	0	862	32128	SentencePiece	
29	t5-11b	Bidir. transf.	24	25600	1024	1	0	862	32128	SentencePiece	
30	albert-base-v1	Bidir. transf.	12	9984	128	1	0	16	30000	SentencePiece	
31	albert-base-v2	Bidir. transf.	12	9984	128	1	0	16	30000	SentencePiece	
32	albert-large-v1	Bidir. transf.	24	25600	128	1	0	16	30000	SentencePiece	
33	albert-large-v2	Bidir. transf.	24	25600	128	1	0	16	30000	SentencePiece	Masked Language Modeling Sentence-Order Prediction
34	albert-xlarge-v1	Bidir. transf.	24	51200	128	1	0	16	30000	SentencePiece	
35	albert-xlarge-v2	Bidir. transf.	24	51200	128	1	0	16	30000	SentencePiece	
36	albert-xxlarge-v1	Bidir. transf.	12	53248	128	1	0	16	30000	SentencePiece	
37	albert-xxlarge-v2	Bidir. transf.	12	53248	128	1	0	16	30000	SentencePiece	
38	openaigpt	Unidir. transf.	12	9984	768	0	0	3	40478	BPE	Causal Language Modeling
39	distilgpt2	Unidir. transf.	6	5376	768	0	0	40	50257	bBPE	
40	gpt2	Unidir. transf.	12	9984	768	0	0	40	50257	bBPE	Causal Language Modeling
41	gpt2-medium	Unidir. transf.	24	25600	1024	0	0	40	50257	bBPE	
42	gpt2-large	Unidir. transf.	36	47360	1280	0	0	40	50257	bBPE	
43	gpt2-xl	Unidir. transf.	48	78400	1600	0	0	40	50257	bBPE	

1189   Table S10: Overview of model designs.

1190

1191
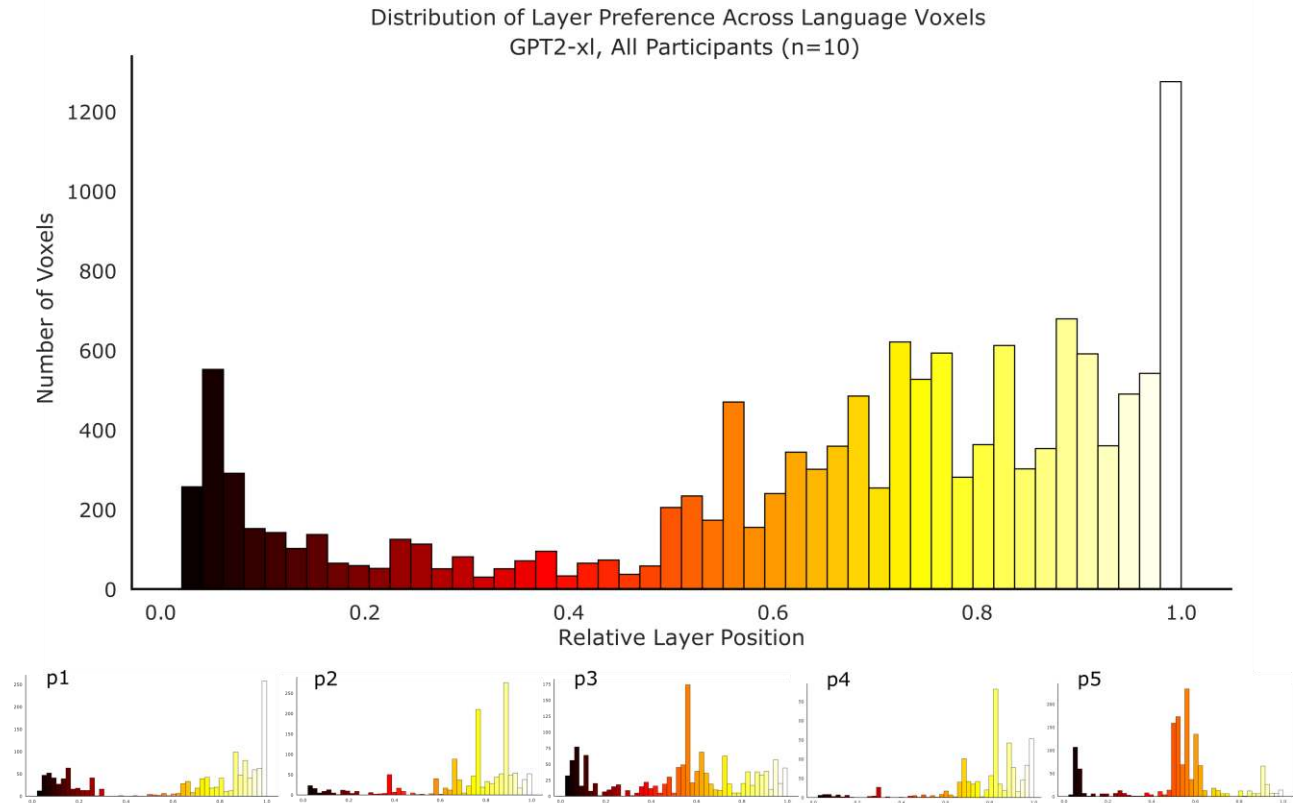
Figure S11: **Distribution of layer preference (best performing layer) per voxel for GPT2-xl for *Pereira2018***. A per-voxel per-participant raw predictivity value was obtained in the language network by computing the mean over cross-validation splits and experiments. For each voxel, the layer with the highest predictivity value was estimated as the "preferred" layer (argmax over layer scores). As in the main analyses, the voxels in the language network were included. Zero on the x-axis corresponds to the embedding layer of the model. The upper plot is averaged across all participants in *Pereira2018* (n=10). The lower panel shows the participant-wise layer preference for five representative participants. Across participants, most voxels show the highest predictivity value for later layers of GPT2-xl. Within participants, the layer preference across voxels varies but is often clustered around particular layers. Investigations of how predictivity fluctuates across model layers, and/or between the language network and other parts of the brain, is left for future work.