The neural architecture of language: Integrative modeling converges on predictive processing

Citation: Schrimpf, Martin, Blank, Idan Asher, Tuckute, Greta, Kauf, Carina, Hosseini, Eghbal A et al. 2021. "The neural architecture of language: Integrative modeling converges on predictive processing." Proceedings of the National Academy of Sciences, 118 (45).

As Published: 10.1073/pnas.2105646118

Publisher: Proceedings of the National Academy of Sciences

Persistent URL: https://hdl.handle.net/1721.1/138214

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

The neural architecture of language: Integrative modeling converges on predictive processing

Martin Schrimpf[a,b,c,1], Idan Asher Blank[a,d,2], Greta Tuckute[a,b,2], Carina Kauf[a,b,2], Eghbal A. Hosseini[a,b], Nancy Kanwisher[a,b,c,1], Joshua B. Tenenbaum[a,c,3], and Evelina Fedorenko[a,b,1,3]

[a]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; [b]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139; [c]Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139; and [d]Department of Psychology, University of California, Los Angeles, CA 90095

Contributed by Nancy Kanwisher, September 3, 2021 (sent for review April 27, 2021; reviewed by Matthew M. Botvinick and Adele E. Goldberg)

The neuroscience of perception has recently been revolutionized with an integrative modeling approach in which computation, brain function, and behavior are linked across many datasets and many computational models. By revealing trends across models, this approach yields novel insights into cognitive and neural mechanisms in the target domain. We here present a systematic study taking this approach to higher-level cognition: human language processing, our species' signature cognitive skill. We find that the most powerful "transformer" models predict nearly 100% of explainable variance in neural responses to sentences and generalize across different datasets and imaging modalities (functional MRI and electrocorticography). Models' neural fits ("brain score") and fits to behavioral responses are both strongly correlated with model accuracy on the next-word prediction task (but not other language tasks). Model architecture appears to substantially contribute to neural fit. These results provide computationally explicit evidence that predictive processing fundamentally shapes the language comprehension mechanisms in the human brain.

computational neuroscience | language comprehension | neural recordings (fMRI and ECoG) | artificial neural networks | deep learning

A core goal of neuroscience is to decipher from patterns of neural activity the algorithms underlying our abilities to perceive, think, and act. Recently, a new "reverse engineering" approach to computational modeling in systems neuroscience has transformed our algorithmic understanding of the primate ventral visual stream (1–8) and holds great promise for 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 (1, 2, 4, 9, 10), 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, and inferior temporal [IT] cortex), in both macaque and human brains, approaching the noise ceiling of the data. Thus, despite abstracting away aspects of biology, DCNNs provide the basis for a first complete hypothesis of how the brain extracts object percepts from visual input.

Inspired by this success story, analogous ANN models have now been applied to other domains of perception (11, 12). Could these models also let us reverse-engineer the brain mechanisms of higher-level human cognition? Here we show how the modeling approach pioneered in the ventral stream can be applied to a higher-level cognitive domain that plays an essential role in human life: language comprehension, or the extraction of meaning from spoken, written, or signed words and sentences. Cognitive scientists have long treated neural network models of language processing with skepticism (13, 14), given that these systems lack (and often deliberately attempt to do without) explicit symbolic representation—traditionally seen as a core feature of linguistic meaning. Recent ANN models of language, 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 integrative modeling approach has proven so powerful 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 (4) that establishes consistent patterns of performance across many different ANNs applied to multiple neural and behavioral datasets, together with their

Significance

Language is a quintessentially human ability. Research has long probed the functional architecture of language in the mind and brain using diverse neuroimaging, behavioral, and computational modeling approaches. However, adequate neurally-mechanistic accounts of how meaning might be extracted from language are sorely lacking. Here, we report a first step toward addressing this gap by connecting recent artificial neural networks from machine learning to human recordings during language processing. We find that the most powerful models predict neural and behavioral responses across different datasets up to noise levels. Models that perform better at predicting the next word in a sequence also better predict brain measurements—providing computationally explicit evidence that predictive processing fundamentally shapes the language comprehension mechanisms in the brain.

Author contributions: M.S., I.A.B., G.T., C.K., N.K., J.B.T., and E.F. designed research; M.S., I.A.B., G.T., and C.K. performed research; M.S. contributed new reagents/analytic tools; M.S., I.A.B., G.T., C.K., and E.A.H. analyzed data; and M.S., I.A.B., G.T., C.K., N.K., J.B.T., and E.F. wrote the paper.

Reviewers: M.M.B., DeepMind; and A.E.G., Princeton University.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: msch@mit.edu, ngk@mit.edu, or evelina9@mit.edu.

[2]I.A.B., G.T., and C.K. contributed equally to this work.

[3]J.B.T. and E.F. contributed equally to this work.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2105646118/-/DCSupplemental.

Published November 4, 2021.

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, any one of these models is surely oversimplified and ultimately wrong—at best, an approximation of some aspects of what the brain does. However, some models are less wrong than others, and consistent trends in performance across models can reveal not just which model best fits the brain but also which properties of a model underlie its fit to the brain, thus yielding critical insights that transcend what any 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 nonhuman primates (1, 2, 4, 15). This integrative benchmarking reveals a rich pattern of correlations among three classes of performance measures—1) neural variance explained in IT neurophysiology or functional MRI (fMRI) responses (brain scores), 2) accuracy in predicting hits and misses in human object recognition behavior or human object similarity judgments (behavioral scores), and 3) accuracy on the core object recognition task (computational task score)—such that for any individual DCNN model we can predict how well it would score 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 laboratories and several years. 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 with ever-increasing precision.

Here we describe an analogous pattern of results for ANN models of human language, establishing a link between language models, including transformer-based ANN architectures that have revolutionized NLP in AI systems over the last 3 y, and fundamental computations of human language processing as reflected in both neural and behavioral measures. Language processing is known to depend causally on a left-lateralized frontotemporal brain network (16–22) (Fig. 1) that responds robustly and selectively to linguistic input (23, 24), whether auditory or visual (25, 26). 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 (27–41) and neural responses to linguistic input (42–64). However, none of the prior studies have attempted large-scale integrative benchmarking that has proven so valuable in understanding key brain–behavior–computation relationships in the ventral stream; instead, they have typically tested one or a small number of models against a single dataset, and the same models have not been evaluated on all three metrics of neural, behavioral, and objective task performance. Previously tested models have also left much of the variance in human neural/behavioral data unexplained. Finally, until the rise of recent ANNs (e.g., transformer architectures), language models did not have sufficient capacity to solve the full linguistic problem that the brain solves—to form a representation of sentence meaning capable of performing a broad range of real-world language tasks on diverse natural linguistic input. We are thus left with a collection of suggestive results but no clear sense of how close ANN models are to fully explaining language processing in the brain, or what model features are key in enabling models to explain neural and behavioral data.

Our goal here is to present a systematic integrative modeling study of language in the brain, at the scale necessary to discover robust relationships between neural and behavioral measurements from humans, and performance of models on language tasks. We seek to determine not just which model fits empirical data best but also what dimensions of variation across models are correlated with fit to human data. This approach has not been applied in the study of language or any other higher cognitive system, and even in perception has not been attempted within a single integrated study. Thus, we view our work more generally as a template for how to apply the integrative benchmarking approach to any perceptual or cognitive system.

Specifically, we examined the relationships between 43 diverse state-of-the-art ANN language models (henceforth "models") across three neural language comprehension datasets (two fMRI, one electrocorticography [ECoG]), as well as behavioral signatures of human language processing in the form of self-paced reading times, and a range of linguistic functions assessed via standard engineering tasks from NLP. The models spanned all major classes of existing ANN language approaches and included simple embedding models [e.g., GloVe (65)], more complex recurrent neural networks [e.g., LM1B (66)], and many variants of transformers or attention-based architectures—including both "unidirectional-attention" models [trained to predict the next word given the previous words, e.g., GPT (67)] and "bidirectional-attention" models [trained to predict a missing word given the surrounding context, e.g., BERT (68)].

Our integrative approach yielded four major findings. First, models' relative fit to neural data (neural predictivity or "brain score")—estimated on held-out test data—generalizes across different datasets and imaging modality (fMRI and ECoG), and certain architectural features consistently lead to more brain-like models: Transformer-based models perform better than recurrent networks or word-level embedding models, and larger-capacity models perform better than smaller models. Second, the best models explain nearly 100% of the explainable variance (up to the noise ceiling) in neural responses to sentences. This result stands in stark contrast to earlier generations of models that have typically accounted for at most 30 to 50% of the predictable neural signal. Third, across models, significant correlations hold among all three metrics of model performance: brain scores (fit to fMRI and ECoG data), behavioral scores (fit to reading time), and model accuracy on the next-word prediction task. Importantly, no other linguistic task was predictive of models' fit to neural or behavioral data. These findings provide strong evidence for a classic hypothesis about the computations underlying human language understanding, that the brain's language system is optimized for predictive processing in the service of meaning extraction. Fourth, intriguingly, the scores of models initialized with random weights (prior to training, but with a trained linear readout) are well above chance and correlate with trained model scores, which suggests that network architecture is an important contributor to a model's brain score. In particular, one architecture introduced just in 2019, the generative pretrained transformer (GPT-2), consistently outperforms all other models and explains almost all variance in both fMRI and ECoG data from sentence-processing tasks. GPT-2 is also arguably the most cognitively plausible of the transformer models (because it uses unidirectional, forward attention) and performs best overall as an AI system when considering both natural language understanding and natural language generation tasks. Thus, even though the goal of contemporary AI is to improve model performance and not necessarily to build models of brain processing, this endeavor appears to be rapidly converging on architectures that might capture key aspects of language processing in the human mind and brain.

Results

We evaluated a broad range of state-of-the-art ANN language models on the match of their internal representations to three

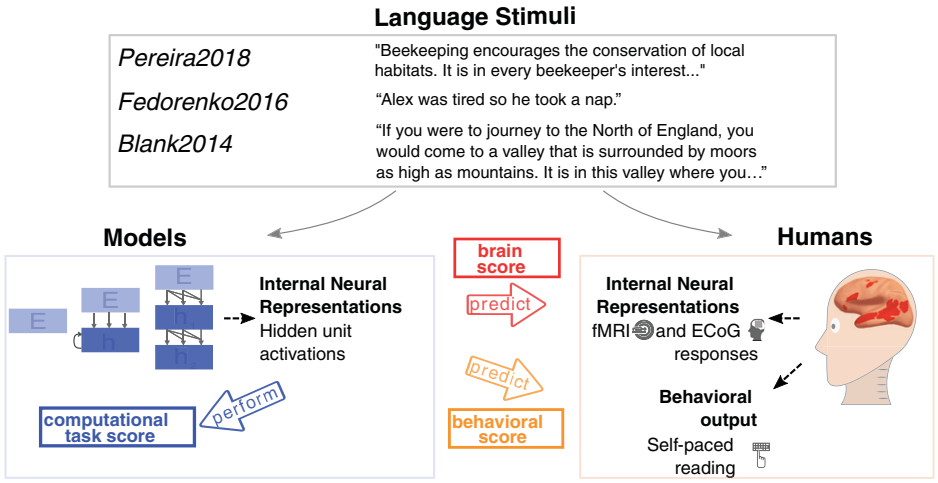

Fig. 1. Comparing ANN models of language processing to human language processing. We tested how well different models predict measurements of human neural activity (fMRI and ECoG) and behavior (reading times) during language comprehension. The candidate models ranged from simple embedding models to more 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 (brain score, extracted from within the frontotemporal language network [e.g., Fedorenko et al. (71)]; the network topography is schematically illustrated in red on the template brain above); their ability to predict human behavior in the form of reading times (behavioral score); and their ability to perform computational tasks such as next-word prediction (computational task score). Consistent relationships between these measures across many different models reveal insights beyond what a single model can tell us.

human neural datasets. The models spanned all major classes of existing language models (*Methods*, section 5 and *SI Appendix*, Table S1). The neural datasets consisted of 1) fMRI activations while participants read short passages, presented one sentence at a time (across two experiments) that spanned diverse topics [*Pereira2018* dataset (45)]; 2) ECoG recordings while participants read semantically and syntactically diverse sentences, presented one word at a time [*Fedorenko2016* dataset (69)]; and 3) fMRI blood oxygen level–dependent (BOLD) signal time series elicited while participants listened to ∼5-min-long naturalistic stories [*Blank2014* dataset (70)] (*Methods*, sections 1 through 3). Thus, the datasets varied in the imaging modality (fMRI/ECoG), the nature of the materials (unconnected sentences/passages/stories), the grain of linguistic units to which responses were recorded (sentences/words/2-s-long story fragments), and presentation modality (reading/listening). In most analyses, we consider the overall results across the three neural datasets; when considering the results for the individual neural datasets, we give the most weight to *Pereira2018* because it includes multiple repetitions per stimulus (sentence) within each participant and quantitatively exhibits the highest internal reliability (*SI Appendix*, Fig. S1). Because our research questions concern language processing, we extracted neural responses from language-selective voxels or electrodes that were functionally identified by an extensively validated independent "localizer" task that contrasts reading sentences versus nonword sequences (69). This localizer robustly identifies the frontotemporal language-selective network (*Methods*, sections 1 through 3).

To compare a given model to a given dataset, we presented the same stimuli to the model that were presented to humans in neural recording experiments and "recorded" the model's internal activations (*Methods*, sections 5 and 6 and Fig. 1). We then tested how well the model recordings could predict the neural recordings for the same stimuli, using a method originally developed for studying visual object recognition (1, 2). Specifically, using a subset of the stimuli, we fit a linear regression from the model activations to the corresponding human measurements, modeling the response of each voxel (*Pereira2018*)/electrode (*Fedorenko2016*)/brain region (*Blank2014*) as a linear weighted sum of responses of different units from the model. We then computed model predictions by applying the learned regression weights to model activations for the held-out stimuli and evaluated how well those predictions matched the corresponding held-out human measurements by computing Pearson's correlation coefficient. We further normalized these correlations by the extrapolated reliability of the particular dataset, which places an upper bound ("ceiling") on the correlation between the neural measurements and any external predictor (*Methods*, section 7 and *SI Appendix*, Fig. S1). The final measure of a model's performance ("score") on a dataset is thus Pearson's correlation between model predictions and neural recordings divided by the estimated ceiling and averaged across voxels/electrodes/regions and participants. We report the score for the best-performing layer of each model (*Methods*, section 6 and *SI Appendix*, Fig. S10) but control for the generality of the layer choice in a train/test split (*SI Appendix*, Fig. S2 *B* and *C*).

Specific Models Accurately Predict Human Brain Activity. We found (Fig. 2 *A* and *B*) that specific models predict *Pereira2018* and *Fedorenko2016* datasets with up to 100% predictivity relative to the noise ceiling (*Methods*, section 7 and *SI Appendix*, Fig. S1). These scores generalize to another metric, based on representational dissimilarity matrices (RDM), without any fitting (*SI Appendix*, Fig. S2*A*). The *Blank2014* dataset is also reliably predicted, but with lower predictivity. Models vary substantially in their ability to predict neural data. Generally, embedding models such as GloVe do not perform well on any dataset. In contrast, recurrent networks such as skip-thoughts, as well as transformers such as BERT, predict large portions of the data. The model that predicts the human data best across datasets is GPT2-xl, a unidirectional-attention transformer model, which predicts *Pereira2018* and *Fedorenko2016* at close to 100% of the noise ceiling and is among the highest-performing models on *Blank2014* with 32% normalized predictivity. These scores are higher in the language network than other parts of the brain (*SI Appendix*, section SI-1). Intermediate layer representations in the models are most predictive, significantly outperforming

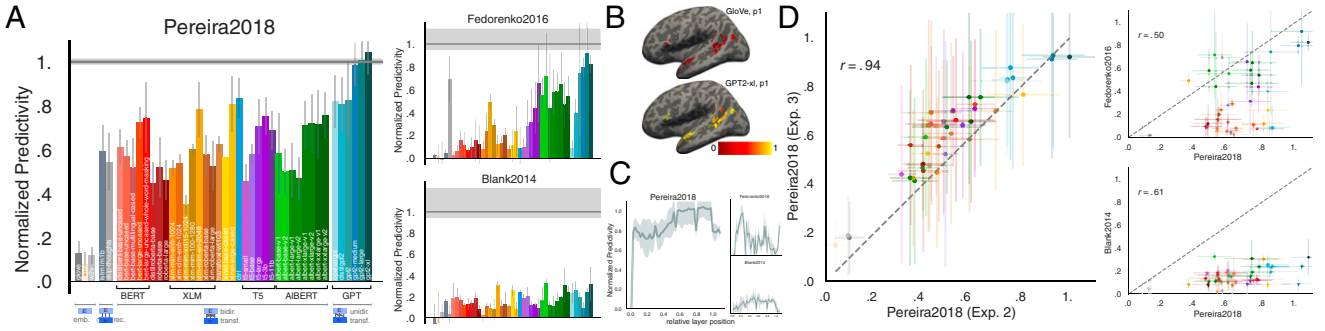

Fig. 2. Specific models accurately predict neural responses consistently across datasets. (*A*) We compared 43 computational models of language processing (ranging from embedding to recurrent and bi- and unidirectional transformer models) in their ability to predict human brain data. The neural datasets include fMRI voxel responses to visually presented (sentence-by-sentence) passages (*Pereira2018*), ECoG electrode responses to visually presented (word-by-word) sentences (*Fedorenko2016*), and fMRI ROI responses to auditorily presented ~5-min-long stories (*Blank2014*). For each model, we plot the normalized predictivity ("brain score"), i.e., the fraction of ceiling (gray line; *Methods*, section 7 and *SI Appendix*, Fig. S1) that the model can predict. Ceiling levels are 0.32 (*Pereira2018*), 0.17 (*Fedorenko2016*), and 0.20 (*Blank2014*). Model classes are grouped by color (*Methods*, section 5 and *SI Appendix*, Table S1). Error bars (here and elsewhere) represent median absolute deviation 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 a representative participant from *Pereira2018* (also *SI Appendix*, Fig. S3). (*C*) Brain score per layer in GPT2-xl. Middle-to-late layers generally yield the highest scores for *Pereira2018* and *Blank2014* whereas earlier layers better predict *Fedorenko2016*. This difference might be due to predicting individual word representations (within a sentence) in *Fedorenko2016*, as opposed to whole-sentence representations in *Pereira2018*. (*D*) To test how well model brain scores generalize across datasets, we correlated 1) two experiments with different stimuli (and some participant overlap) in *Pereira2018* (obtaining a very strong correlation) and 2) *Pereira2018* brain scores with the scores for each of *Fedorenko2016* and *Blank2014* (obtaining lower but still highly significant correlations). Brain scores thus tend to generalize across datasets, although differences between datasets exist which warrant the full suite of datasets.

representations at the first and output layers (Fig. 2*C* and *SI Appendix*, Fig. S10).

Model scores are consistent across experiments/datasets. To test the generality of the model representations, we examined the consistency of model brain scores across datasets. Indeed, if a model achieves a high brain score on one dataset it tends to also do well on other datasets (Fig. 2*D*), ruling out the possibility that we are picking up on spurious, dataset-idiosyncratic predictivity and suggesting that the models' internal representations are general enough to capture brain responses to diverse linguistic materials presented visually or auditorily, and across three independent sets of participants. Specifically, model brain scores across the two experiments in *Pereira2018* (overlapping sets of participants) correlate at $r = 0.94$ (Pearson here and elsewhere, $P << 0.00001$), scores from *Pereira2018* and *Fedorenko2016* correlate at $r = 0.50$ ($P < 0.001$), and from *Pereira2018* and *Blank2014* at $r = 0.63$ ($P < 0.0001$).

Next-Word-Prediction Task Performance Selectively Predicts Brain Scores. In the critical test of which computations might underlie human language understanding, we examined the relationship between the models' ability to predict an upcoming word and their brain scores. Words from the WikiText-2 dataset (72) were sequentially fed into the candidate models. We then fit a linear classifier (over words in the vocabulary; $n = 50{,}000$) from the last layer's feature representation (frozen, i.e., no fine-tuning) on the training set to predict the next word and evaluated performance on the held-out test set (*Methods*, section 8). Indeed, next-word-prediction task performance robustly predicts brain scores (Fig. 3*A*; $r = 0.44$, $P < 0.01$, averaged across datasets). The best language model, GPT2-xl, also achieves the highest brain score (see previous section). This relationship holds for model variants within each model class—embedding models, recurrent networks, and transformers—ruling out the possibility that this correlation is due to between-class differences in next-word-prediction performance.

To test whether next-word prediction is special in this respect, we asked whether model performance on any language task correlates with brain scores. As with next-word prediction, we kept the model weights fixed and only trained a linear readout. We found that performance on tasks from the General Language Understanding Evaluation (GLUE) benchmark collection (73–80)—including grammaticality judgments, sentence similarity judgments, and entailment—does not predict brain scores (Fig. 3 *B* and *C*). The difference in the strength of correlation between brain scores and the next-word-prediction task performance vs. the GLUE tasks performance is highly reliable ($P << 0.00001$, t test over 1,000 bootstraps of scores and corresponding correlations; *Methods*, section 9). This result suggests that optimizing for predictive representations may be a critical shared objective of biological and artificial neural networks for language, and perhaps more generally (81, 82)

Brain Scores and Next-Word-Prediction Task Performance Correlate with Behavioral Scores. Beyond internal neural representations, we tested the models' ability to predict external behavioral outputs because, ultimately, in integrative benchmarking we strive for a computationally precise account of language processing that can explain both neural response patterns and observable linguistic behaviors. We chose a large corpus ($n = 180$ participants) of self-paced reading times for naturalistic story materials [*Futrell2018* dataset (83)]. Per-word reading times provide a theory-neutral measure of incremental comprehension difficulty, which has long been a cornerstone of psycholinguistic research in testing theories of sentence comprehension (28, 33, 83–87) and which were recently shown to robustly predict neural activity in the language network (88).

Specific models accurately predict reading times. We regressed each model's last layer's feature representation (i.e., closest to the output) against reading times and evaluated predictivity on held-out words. As with the neural datasets, we observed a spread of model ability to capture human behavioral data, with models such as GPT2-xl and AlBERT-xxlarge predicting these data close to the noise ceiling (Fig. 4*A* and refs. 89 and 90).

Brain scores correlate with behavioral scores. To test whether models with the highest brain scores also predict reading times best, we compared models' neural predictivity (across datasets) with those same models' behavioral predictivity. Indeed, we observed a strong correlation (Fig. 4*B*; $r = 0.65$, $P << 0.0001$), which also holds for the individual neural datasets (Fig. 4*B*,

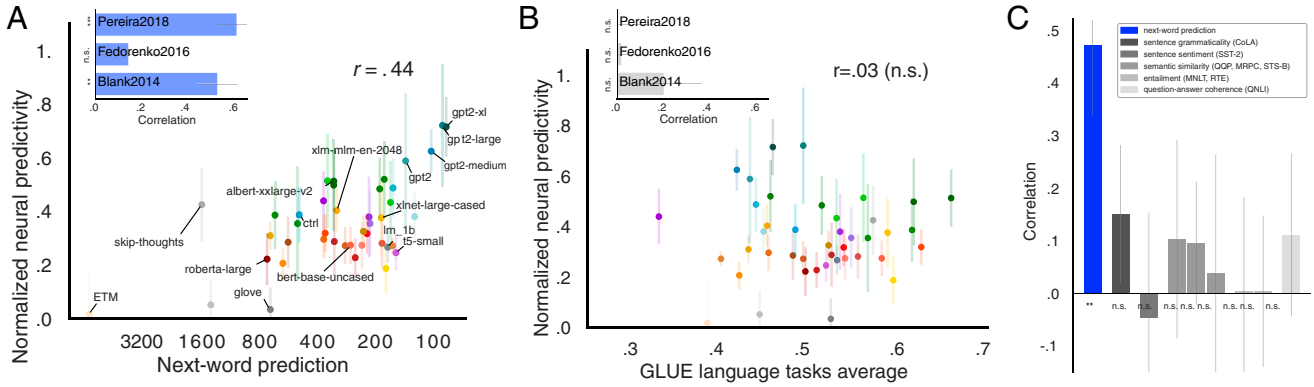

Fig. 3. Model performance on a next-word-prediction task selectively predicts brain scores. (*A*) Next-word-prediction task performance was evaluated as the surprisal between the predicted and true next word in the WikiText-2 dataset of 720 Wikipedia articles, or perplexity (*x* axis, lower values are better; training only a linear readout leading to worse perplexity values than canonical fine-tuning, see *Methods*, section 8). Next-word-prediction task scores strongly predict brain scores across datasets (*Inset*: this correlation is significant for two individual datasets: *Pereira2018* and *Blank2014*; the correlation for *Fedorenko2016* is positive but not significant). (*B*) Performance on diverse language tasks from the GLUE benchmark collection does not correlate with overall or individual-dataset brain scores (*Inset*; *SI Appendix*, SI-2; training only a linear readout). (*C*) Correlations of individual tasks with brain scores. Only improvements on next-word prediction lead to improved neural predictivity.

Inset and *SI Appendix*, Fig. S5). These results suggest that further improving models' neural predictivity will simultaneously improve their behavioral predictivity.

Next-word-prediction task performance correlates with behavioral scores. Next-word-prediction task performance is predictive of reading times (Fig. 4*C*; $r = 0.67$, $P << 0.0001$), in line with earlier studies (91, 92) and thus connecting all three measures of performance: brain scores, behavioral scores, and task performance on next-word prediction.

Model Architecture Contributes to Model-to-Brain Relationship. The brain's language network plausibly arises through a combination of evolutionary and learning-based optimization. In a first attempt to test the relative importance of the models' intrinsic architectural properties vs. training-related features, we performed two analyses. First, we found that architectural features (e.g., number of layers) but neither of the features related to training (dataset and vocabulary size) significantly predicted improvements in model performance on the neural data (*SI Appendix*, Section SI-3, Fig. S8, and Table S1). These results align with prior studies that had reported that architectural differences affect model performance on normative tasks like next-word prediction after training and define the representational space that the model can learn (93–95). Second, we computed brain scores for the 43 models without training, i.e., with initial (random) weights. Note that the predictivity metric still trains a linear readout on top of the model representations. Surprisingly, even with no training, several models achieved reasonable scores (Fig. 5), consistent with recent results of models in high-level visual cortex (95) as well as findings on the power of random initializations in NLP (96–98). For example, across the three datasets, untrained GPT2-xl achieves an average predictivity of ∼51%, only ∼20% lower than the trained network. A similar trend is observed across models: Training generally improves brain scores, on average by 53%. Across models, the untrained scores are strongly predictive of the trained scores ($r = 0.74$, $P << 0.00001$), indicating that models that already perform well with random weights improve further with training.

To ensure the robustness and generalizability of the results for untrained models, and to gain further insights into these results, we performed four additional analyses (*SI Appendix*, Fig. S8). First, we tested a random context-independent embedding with equal dimensionality to the GPT2-xl model but no architectural priors and found that it predicts only a small fraction of the neural data, on average below 15%, suggesting that a large feature space alone is not sufficient (*SI Appendix*, Fig. S8*A*). Second, to ensure that the overlap between the linguistic materials (words, bigrams, etc.) used in the train and test splits is not driving the results, we quantified the overlap and found it to be low,

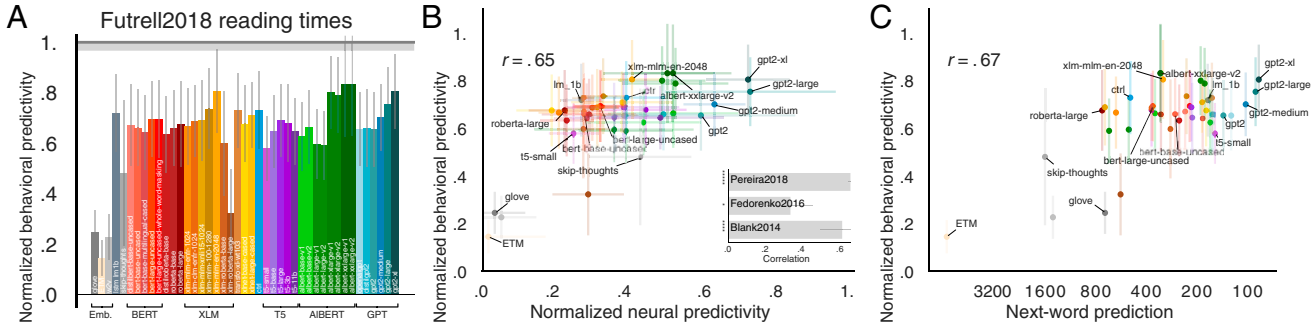

Fig. 4. Behavioral scores, brain scores, and next-word-prediction task performance are pairwise correlated. (*A*) Behavioral predictivity of each model on *Futrell2018* reading times (notation similar to Fig. 2). Ceiling level is 0.76. (*B*) Models' neural scores aggregated across the three neural datasets (or for each dataset individually; *Inset* and *SI Appendix*, Fig. S5) correlate with behavioral scores. (*C*) Next-word-prediction task performance (Fig. 3) correlates with behavioral scores. Performance on other language tasks (from the GLUE benchmark collection) does not correlate with behavioral scores (*SI Appendix*, Fig. S6).

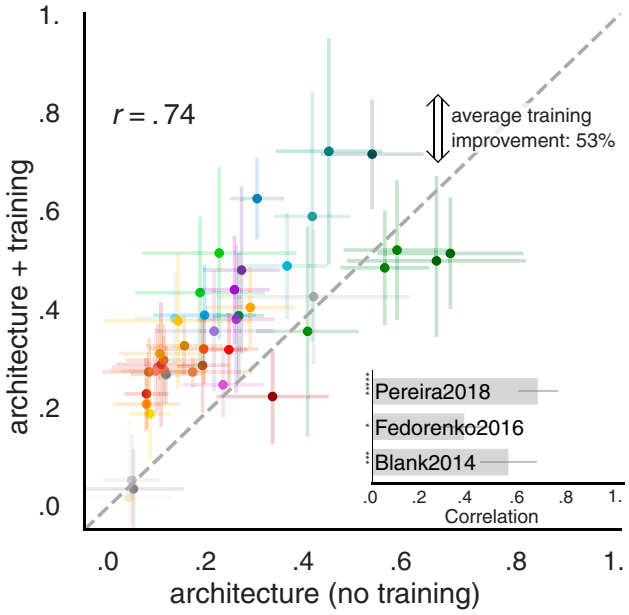

Fig. 5. Model architecture contributes to the model–brain relationship. We evaluate untrained models by keeping weights at their initial random values. The remaining representations are driven by architecture alone and are tested on the neural datasets (Fig. 2). Across the three datasets, architecture alone yields representations that predict human brain activity considerably well. On average, training improves model scores by 53%. For *Pereira2018*, training improves predictivity the most whereas for *Fedorenko2016* and *Blank2014*, training does not always change—and for some models even decreases—neural scores (*SI Appendix*, Fig. S7). The untrained model performance is consistently predictive of its performance after training across and within (*Inset*) datasets.

especially for bi- and trigrams (*SI Appendix*, Fig. S8*B*). Third, to ensure that the linear regression used in the predictivity metric did not artificially inflate the scores of untrained models, we used an alternative metric—RDM—that does not involve any fitting. Scores of untrained models on the predictivity metric generalized to scores on the RDM metric (*SI Appendix*, Fig. S8*D*). Finally, we examined the performance of untrained models with a trained linear readout on the next-word-prediction task and found performance trends similar to those we observed for the neural scores (*SI Appendix*, Fig. S8*C*), confirming the representational power of untrained representations.

Discussion

Summary of Key Results and Their Implications. Our results, summarized in Fig. 6, show that specific ANN language models can predict human neural and behavioral responses to linguistic input with high accuracy: The best models achieve, on some datasets, perfect predictivity relative to the noise ceiling. Model scores correlate across neural and behavioral datasets spanning recording modalities (fMRI, ECoG, and reading times) and diverse materials presented visually and auditorily across four sets of participants, establishing the robustness and generality of these findings. Critically, both neural and behavioral scores correlate with model performance on the normative next-word-prediction task—but not other language tasks. Finally, untrained models with random weights (and a trained linear readout) produce representations beginning to approximate those in the brain's language network.

Predictive language processing. Underlying the integrative modeling framework, implemented here in the cognitive domain of language, is the idea that large-scale neural networks can serve as hypotheses of the computations conducted in the brain. We here identified some models—unidirectional-attention transformer architectures—that accurately capture brain activity during language processing. We then began dissecting variations across the range of model candidates to explain why they achieve high brain scores. Two core findings emerged, both supporting the idea that the human language system is optimized for predictive processing. First, we found that the models' performance on the next-word-prediction task, but not other language tasks, is correlated with neural predictivity (see ref. 51 for related evidence of fine-tuning of one model on tasks other than next-word prediction leading to worse model-to-brain fit). Recent preprints conceptually replicate and extend this basic finding (88, 90, 99, 100). Language modeling (predicting the next word) is the task of choice in the

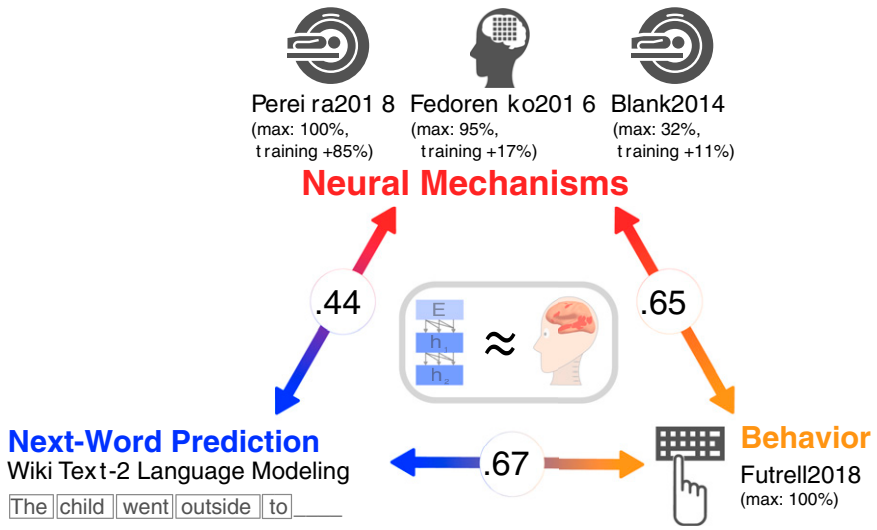

Fig. 6. Overview of results connecting neural mechanisms, behavior, and computational task (next-word prediction). Specific ANN language models are beginning to approximate the brain's mechanisms for processing language (*Middle*, gray box). For the neural datasets (fMRI and ECoG recordings; *Top*, red) and for the behavioral dataset (self-paced reading times; *Bottom Right*, orange) we report 1) the value for the model achieving the highest predictivity ("max") and 2) the average improvement on brain scores across models after training. Model performances on the next-word-prediction task (WikiText-2 language modeling perplexity; *Bottom Left*, blue) predict brain and behavioral scores and brain scores predict behavioral scores (circled numbers).

NLP community: It is simple, unsupervised, scalable, and appears to produce the most generally useful, successful language representations. This is likely because language modeling encourages a neural network to build a joint probability model of the linguistic signal, which implicitly requires sensitivity to diverse kinds of regularities in the signal.

Second, we found that the models that best match human language processing are precisely those that are trained to predict the next word. Predictive processing has advanced to the forefront of theorizing in cognitive science (101–109) and neuroscience (81, 110–113), including in the domain of language (39, 114). The rich sources of information that comprehenders combine to interpret language—including lexical and syntactic information, world knowledge, and information about others' mental states (115–119)—can be used to make informed guesses about how the linguistic signal may unfold, and much behavioral and neural evidence now suggests that readers and listeners indeed engage in such predictive behavior (33, 47, 114, 120, 121). 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.

Going beyond the broad idea of prediction in the neuroscience of language, the work presented here validates, refines, and computationally implements an explicit account of predictive processing: We were able to accurately predict (relative to the noise ceiling) activity across voxels as well as neuronal populations in human cortex during the processing of sentences. We quantitatively test the predictive processing hypothesis at the level of voxel/electrode/fROI (functional region of interest) responses and, through the use of end-to-end models, related neural mechanisms to performance of models on computational tasks. Moreover, we were able to reject multiple alternative hypotheses about the objective of the language system: Model performance on diverse benchmarks from the GLUE suite of benchmarks (73), including judgments about syntactic and semantic properties of sentences, was not predictive of brain or behavioral scores. The best-performing computational models identified in this work serve as computational explanations for the entire language processing pipeline from word inputs to neural mechanisms to behavioral outputs. These best-performing models can now be further dissected, as well as tested on new diverse, linguistic inputs in future experiments, as discussed below.

Importance of architecture. We also found that architecture is an important contributor to the models' match to human brain data: Untrained models with a trained linear readout performed well above chance in predicting neural activity, and this finding held under a series of controls to alleviate concerns that it could be an artifact of our training or testing methodologies. This result is consistent with findings in models of early (5, 7, 95) and high-level visual processing (95) and speech perception (122), as well as recent results in NLP (96–98), but it raises important questions of interpretation in the context of human language. 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 relatively rapid acquisition of language (123–125). In that analogy, the human research community's development of new architectures such as the transformer networks that perform well in both NLP tasks and neural language modeling could be akin to recapitulating evolution (126), 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 still optimizes for language modeling, only implicitly and on a different timescale from biological and cultural evolutionary mechanisms conventionally studied in brain and language.

More explicitly, but speculatively, it is possible that transformer networks can work as brain models of language even without extensive training because the hierarchies of local spatial filtering and pooling as found in convolutional as well as attention-based networks are a generally applicable brain-like mechanism to extract abstract features from natural signals. Regardless of the exact filter weights, transformer architectures build on word embeddings that capture both semantic and syntactic features of words and integrate contextually weighted predictions across scales such that contextual dependencies are captured at different scales in different kernels. The representations in such randomized architectures could thus reflect a kind of multiscale, spatially smoothed average (over consecutive inputs) of word embeddings, which might capture the statistical gist-like processing of language observed in both behavioral studies (34, 38, 127) and human neuroimaging (128). The weight sharing within architectural sublayers ("multihead attention") introduced by combinations of query-key-value pairs in transformers might provide additional consistency and coverage of representations. Relatedly, an idea during early work on perceptrons was to have random projections of input data into high-dimensional spaces and to then only train thin readouts on top of these projections. This was motivated by Cover's theorem, which states that nonlinearly separable data can likely be linearly separated after projection into a high-dimensional space (129). These ideas have successfully been applied to kernel machines (130) and are more recently explored again with deep neural networks (131); in short, it is possible that even random features with the right multiscale structure in time and space could be more powerful for representing human language than is currently understood. Finally, it is worth noting that the initial weights in the networks we study stem from weight initializer distributions that were chosen to provide solid starting points for contemporary architectures and lead to reasonable initial representations that model training further refines. These initial representations could thus include some important aspects of language structure already. A concrete test for these ideas would be the following: Construct model variants that average over word embeddings at different scales and compare these models' representations with those of different layers in untrained transformer architectures as well as the neural datasets. More detailed analyses, including minimal-pair model variant comparisons, will be needed to fully separate the representational contributions of architecture and training.

Limitations and Future Directions. 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 of their inner workings. Such research is much easier to perform on models than on biological systems, given that all their structure and weights are easily accessible and manipulable (132, 133). For example, controlled comparisons of architectural variants and training objectives could define the necessary and sufficient conditions for human-like language processing (134), synergizing with parallel ongoing efforts in NLP to probe ANNs' linguistic representations (135–137). 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. Retraining 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 vast space of hyperparameters. 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. Similarly, here, we

leveraged existing neural datasets. This work can be expanded in many new directions, including 1) assembling a wider range of publicly available language datasets for model testing [cf. vision (2, 4)]; 2) collecting data on new language stimuli for which different models make maximally different predictions [cf. vision (138)], including sampling a wider range of language stimuli (e.g., naturalistic dialogs/conversations); 3) modeling the fine-grained temporal trajectories of neural responses to language in data with high temporal resolution (which requires computational accounts that make predictions about representational dynamics); and 4) querying models on the sentence stimuli that elicit the strongest responses in the language network to generate hypotheses about the critical response-driving feature/feature spaces, and perhaps to discover new organizing principles of the language system [cf. vision (139, 140)].

One of the major limiting factors in modeling the brain's language network is the availability of adequate recordings. Although an increasing number of language fMRI, magnetoencephalography (MEG), electroencephalography, and intracranial datasets are becoming publicly available, they often lack key properties for testing computational language models. In particular, what is needed are data with high signal-to-noise ratio, where neural responses to a particular stimulus (e.g., sentence) can be reliably estimated. However, most past language neuroscience research has focused on coarse distinctions (e.g., sentences with vs. without semantic violations, or sentences with different syntactic structures); as a result, any single sentence is generally only presented once, and neural responses are averaged across all the sentences within a "condition" (in contrast, monkey physiology studies of vision typically present each stimulus dozens of times to each animal, e.g. ref. 141). (Studies that use "naturalistic" language stimuli like stories or movies also typically present the stimuli once, although naturally occurring repetitions of words/n-grams can be useful.) One of the neural datasets in the current study (*Pereira2018*) presented each sentence thrice to each subject and exhibited the highest ceiling (0.32; cf. *Fedorenko2016*: 0.17, *Blank2014*: 0.20). However, even this ceiling is low relative to single cell recordings in the primate ventral stream [e.g., 0.82 for IT recordings (2)]. Such high reliability may not be attainable for higher-level cognitive domains like language, where processing is unlikely to be strictly bottom-up/stimulus-driven. However, this is an empirical question that past work has not attempted to answer and that will be important in the future for building models that can accurately capture the neural mechanisms of language.

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 problem 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., refs. 142–146). 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 (142). The fact that language models lack meaning and focus on local linguistic coherence (90, 147) may explain why their representations fall short of ceiling on *Blank2014*, which uses story materials and may therefore require long-range contexts.

Another 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 frontotemporal 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 (3) and subregions (148). However, whereas function and its mapping onto anatomy is at least coarsely defined in the case of vision (149), a similar mapping is not yet established in language beyond the broad distinction between perceptual processing and higher-level linguistic interpretation (e.g., ref. 21). The ANN models of human language processing identified in this work might also serve to uncover these kinds of anatomical distinctions for the brain's language network—perhaps, akin to vision, groups of layers relate to different cortical regions and uncovering increased similarity to neural activity of one group over others could help establish a cortical hierarchy. The brain 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 (150, 151), 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.

Conclusions

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

Methods

More detailed information can be found in *SI Appendix*, *SI-Methods*.

1) **Neural dataset 1: fMRI (*Pereira2018*).** We used the data from Pereira et al.'s (45) experiments 2 ($n = 9$) and 3 ($n = 6$) (10 unique participants). Stimuli for experiment 2 consisted of 384 sentences (96 text passages, four sentences each), and stimuli for experiment 3 consisted of 243 sentences (72 text passages, three or four sentences each). Sentences were presented on the screen one at a time for 4 s.
2) **Neural dataset 2: ECoG (*Fedorenko2016*).** We used the data from Fedorenko et al. (69) ($n = 5$). Stimuli consisted of 80 hand-constructed eight-word-long semantically and syntactically diverse sentences and 80 lists of nonwords; we selected the 52 sentences that were presented to all participants. Materials were presented visually one word at a time for 450 or 700 ms.
3) **Neural dataset 3: fMRI (*Blank2014*).** We used the data from Blank et al. (70) ($n = 5$, 5 of the 10 participants that have been exposed to the same materials). Stimuli consisted of stories from the publicly available Natural Stories Corpus (83), which are stories adapted from existing texts (fairy tales and short stories). Stories were presented auditorily (~5 min in duration each).
4) **Behavioral dataset: Self-paced reading (*Futrell2018*).** We used the data from Futrell et al. (83) ($n = 179$, excluding one participant with too few data points). Stimuli consisted of 10 stories from the Natural Stories Corpus (83) (same as in *Blank2014*, plus two additional stories). Stories were presented online (on Amazon Mechanical Turk) visually in a dashed moving window display (152), with any given participant reading between 5 and all 10 stories.

5) **Computational models.** We tested 43 language models that were selected to sample a broad range of computational designs across three major types of architecture: embeddings [GloVe (65), ETM (153), word2vec (154)], recurrent architectures [lm_1b (66), skip-thoughts (155)], and attention-based "transformers" [from the HuggingFace library (156), with variants of BERT (68), RoBERTa (157), XLM (158), XLM-RoBERTa (159), XLNet (160), CTRL (161), T5 (162), AlBERT (163), and GPT (67, 164)].
6) **Comparison of models to brain measurements.** We treated the model representation at each layer separately and tested how well it could predict human recordings. To generate predictions, we used 80% of the stimuli to fit a linear regression from the corresponding 80% of model representations to the corresponding 80% of human recordings. We applied the regression on model representations of the held-out 20% of stimuli to generate model predictions, which we then compared against the held-out 20% of human recordings with a Pearson correlation. This process was repeated five times, leaving out different 20% of stimuli each time. We aggregated voxel/electrode/ROI predictivity scores by taking the median of scores for each participant's voxels/electrodes/ROIs and then computing the median across participants. Finally, this score was divided by the estimated ceiling value (see below) to yield a final score in the range [0, 1].
7) **Estimation of ceiling.** Due to intrinsic noise in biological measurements, we estimated a ceiling value to reflect how well the best possible model of an average human could perform. In brief, we subsampled the data, designating one subject as the prediction target and treating the recordings of the remaining subject pool as the source representations to predict from (see above). To obtain the most conservative ceiling estimate, we extrapolated the size of the subject pool and used the final ceiling value when extrapolating to infinitely many humans.
8) **Language modeling.** To assess models' performance on the normative next-word-prediction task, we used WikiText-2 (165), a dataset of 720 Wikipedia articles. We sequentially fed tokens into models (rather than chunks of tokens) and captured representations at each step from each model's penultimate layer. To predict the next word, we fit a linear decoder—trained with a cross-entropy loss—from those representations to the next token. We kept weights in the network frozen rather than fine-tuning the entire model in order to maintain the same model representations that were used in model-to-brain and model-to-behavior comparisons. These choices led to worse model performances than state-of-the-art, but we ensured that our pipeline could reproduce published results when fine-tuning the entire model and increasing the batch size. The final language modeling score is reported as the perplexity on the held-out test set.
9) **Statistical tests.** Model-to-brain predictivity scores are reported as the Pearson correlation coefficient (r). Error estimates are computed with a bootstrapped correlation coefficient (1,000 iterations), leaving out 10% of scores and computing r on the remaining 90% held-out scores. All *P* values less than 0.05 are summarized with an asterisk, less than 0.005 with two asterisks, less than 0.0005 with three asterisks, and less than 0.00005 with four asterisks. For interaction tests, we used two-sided *t* tests with 1,000 bootstraps and 90% of samples per bootstrap.

Data Availability. Code, data, and models are available at GitHub (https://github.com/mschrimpf/neural-nlp). All other study data are included in the article and/or *SI Appendix*. Previously published data were used for this work [Pereira et al. (45), Blank et al. (70), Fedorenko et al. (69), and Futrell et al. (83)].

ACKNOWLEDGMENTS. We thank Roger Levy, Steve Piantadosi, Cory Shain, Noga Zaslavsky, Antoine Bosselut, and Jacob Andreas for comments on the manuscript; Tiago Marques for comments on ceiling estimates and feature analysis; Jon Gauthier for comments on language modeling; and Bruce Fischl and Ruopeng Wang for adding a Freeview functionality. M.S. was supported by a Takeda Fellowship, the Massachusetts Institute of Technology Shoemaker Fellowship, and the SRC Semiconductor Research Corporation. G.T. was supported by the Massachusetts Institute of Technology Media Lab Consortia and the Massachusetts Institute of Technology Singleton Fellowship. C.K. was funded by the Massachusetts Institute of Technology Presidential Graduate Fellowship. E.A.H. was supported by the Friends of the McGovern Institute Fellowship. N.K. and J.B.T. were supported by the Center for Brains, Minds, and Machines, funded by NSF STC CCF-1231216. E.F. was supported by NIH awards R01-DC016607, R01-DC016950, and U01-NS121471, and by funds from the Brain and Cognitive Sciences Department and the McGovern Institute for Brain Research at the Massachusetts Institute of Technology.

1. D. L. K. Yamins *et al.*, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
2. M. Schrimpf *et al.*, Brain-score: Which artificial neural network for object recognition is most brain-like? bioRxiv [Preprint] (2018). https://www.biorxiv.org/content/10.1101/407007v1 (Accessed 8 October 2021).
3. J. Kubilius *et al.*, "Brain-like object recognition with high-performing shallow recurrent ANNs" in *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems* (Neural Information Processing Systems Foundation, Inc., 2019), pp. 12785–12796.
4. M. Schrimpf *et al.*, Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* **108**, 413–423 (2020).
5. R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva, Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* **6**, 27755 (2016).
6. T. C. Kietzmann *et al.*, Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 21854–21863 (2019).
7. S. A. Cadena *et al.*, Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLOS Comput. Biol.* **15**, e1006897 (2019).
8. P. Bao, L. She, M. McGill, D. Y. Tsao, A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
9. D. Cireşan, U. Meier, J. Schmidhuber, "Multi-column deep neural networks for image classification" in *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE, 2012), pp. 3642–3649.
10. A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet classification with deep convolutional neural networks" in *Advances in Neural Information Processing Systems (NIPS 2012)* (Neural Information Processing Systems Foundation, Inc., 2012), pp. 1097–1105.
11. A. J. E. Kell, D. L. K. Yamins, E. N. Shook, S. V. Norman-Haignere, J. H. McDermott, A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* **98**, 630–644.e16 (2018).
12. C. Zhuang, J. Kubilius, M. J. Hartmann, D. L. Yamins, "Toward goal-driven neural network models for the rodent whisker-trigeminal system" in *Advances in Neural Information Processing Systems (NIPS 2017)* (Neural Information Processing Systems Foundation, Inc., 2017), pp. 2555–2565.
13. S. Pinker, A. Prince, On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* **28**, 73–193 (1988).
14. G. Marcus, Deep learning: A critical appraisal. arXiv [Preprint] (2018). https://arxiv.org/abs/1801.00631 (Accessed 5 October 2020).
15. R. Rajalingham *et al.*, Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* **38**, 7255–7269 (2018).
16. J. R. Binder *et al.*, Human brain language areas identified by functional magnetic resonance imaging. *J. Neurosci.* **17**, 353–362 (1997).
17. E. Bates *et al.*, Voxel-based lesion-symptom mapping. *Nat. Neurosci.* **6**, 448–450 (2003).
18. M. L. Gorno-Tempini *et al.*, Cognition and anatomy in three variants of primary progressive aphasia. *Ann. Neurol.* **55**, 335–346 (2004).
19. A. D. Friederici, The cortical language circuit: From auditory perception to sentence comprehension. *Trends Cogn. Sci.* **16**, 262–268 (2012).
20. C. J. Price, The anatomy of language: A review of 100 fMRI studies published in 2009. *Ann. N. Y. Acad. Sci.* **1191**, 62–88 (2010).
21. E. Fedorenko, S. L. Thompson-Schill, Reworking the language network. *Trends Cogn. Sci.* **18**, 120–126 (2014).
22. P. Hagoort, The neurobiology of language beyond single-word processing. *Science* **366**, 55–58 (2019).
23. E. Fedorenko, M. K. Behr, N. Kanwisher, Functional specificity for high-level linguistic processing in the human brain. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 16428–16433 (2011).
24. M. M. Monti, L. M. Parsons, D. N. Osherson, Thought beyond language: Neural dissociation of algebra and natural language. *Psychol. Sci.* **23**, 914–922 (2012).
25. M. Regev, C. J. Honey, E. Simony, U. Hasson, Selective and invariant neural responses to spoken and written narratives. *J. Neurosci.* **33**, 15978–15988 (2013).
26. F. Deniz, A. O. Nunez-Elizalde, A. G. Huth, J. L. Gallant, The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *J. Neurosci.* **39**, 7722–7736 (2019).
27. R. Futrell, E. Gibson, R. P. Levy, Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cogn. Sci. (Hauppauge)* **44**, e12814 (2020).
28. E. Gibson, Linguistic complexity: Locality of syntactic dependencies. *Cognition* **68**, 1–76 (1998).
29. M. J. Spivey-Knowlton, "Integration of visual and linguistic information: Human data and model simulations," PhD thesis, University of Rochester, Rochester, NY (1996).
30. M. Steedman, *The Syntactic Process* (MIT Press, 2000).
31. M. van Schijndel, A. Exley, W. Schuler, A model of language processing as hierarchic sequential prediction. *Top. Cogn. Sci.* **5**, 522–540 (2013).
32. J. Dotlačil, Building an ACT-R reader for eye-tracking corpus data. *Top. Cogn. Sci.* **10**, 144–160 (2018).
33. N. J. Smith, R. Levy, The effect of word predictability on reading time is logarithmic. *Cognition* **128**, 302–319 (2013).
34. E. Gibson, L. Bergen, S. T. Piantadosi, Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 8051–8056 (2013).
35. J. Hale, "A probabilistic Earley parser as a psycholinguistic model" in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (NAACL, 2001), pp. 1–8.

36. D. Jurafsky, A probabilistic model of lexical and syntactic access and disambiguation. *Cogn. Sci.* **20**, 137–194 (1996).
37. Y. Lakretz, S. Dehaene, J. R. King, What limits our capacity to process nested long-range dependencies in sentence comprehension? *Entropy (Basel)* **22**, 446 (2020).
38. R. Levy, "A noisy-channel model of rational human sentence comprehension under uncertain input" in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2008), pp. 234–243.
39. R. Levy, Expectation-based syntactic comprehension. *Cognition* **106**, 1126–1177 (2008).
40. R. L. Lewis, S. Vasishth, J. A. Van Dyke, Computational principles of working memory in sentence comprehension. *Trends Cogn. Sci.* **10**, 447–454 (2006).
41. J. McDonald, B. Macwhinney, "Maximum likelihood models for sentence processing" in *The Crosslinguistic Study of Sentence Processing*, B. M. Whinney and E. Bates, Eds. (Cambridge University Press, 1998), pp. 397–421.
42. J. R. Brennan, E. P. Stabler, S. E. Van Wagenen, W. M. Luh, J. T. Hale, Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain Lang.* **157-158**, 81–94 (2016).
43. J. R. Brennan, L. Pylkkänen, MEG evidence for incremental sentence composition in the anterior temporal lobe. *Cogn. Sci. (Hauppauge)* **41** (suppl. 6), 1515–1531 (2017).
44. C. Pallier, A. D. Devauchelle, S. Dehaene, Cortical representation of the constituent structure of sentences. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 2522–2527 (2011).
45. F. Pereira *et al.*, Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* **9**, 963 (2018).
46. M. Rabovsky, S. S. Hansen, J. L. McClelland, Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nat. Hum. Behav.* **2**, 693–705 (2018).
47. C. Shain, I. A. Blank, M. van Schijndel, W. Schuler, E. Fedorenko, fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia* **138**, 107307 (2020).
48. L. Wehbe, A. Vaswani, K. Knight, T. Mitchell, "Aligning context-based statistical models of language with brain activity during reading" in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, 2014), pp. 233–243.
49. R. M. Willems, S. L. Frank, A. D. Nijhof, P. Hagoort, A. van den Bosch, Prediction during natural language comprehension. *Cereb. Cortex* **26**, 2506–2516 (2016).
50. J. Gauthier, A. Ivanova, Does the brain represent words? An evaluation of brain decoding studies of language understanding. arXiv [Preprint] (2018). http://arxiv.org/abs/1806.00591 (Accessed 7 July 2019).
51. J. Gauthier, R. Levy, "Linking artificial and human neural representations of language" in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Association for Computational Linguistics, 2019), pp. 529–539.
52. S. Jain, A. Huth, Incorporating context into language encoding models for fMRI. https://proceedings.neurips.cc/paper/2018/hash/f471223d1a1614b58a7dc45c9d01df19-Abstract.html (Accessed 8 October 2021).
53. S. Wang, J. Zhang, H. Wang, N. Lin, C. Zong, Fine-grained neural decoding with distributed word representations. *Inf. Sci.* **507**, 256–272 (2020).
54. N. Ding, L. Melloni, H. Zhang, X. Tian, D. Poeppel, Cortical tracking of hierarchical linguistic structures in connected speech. *Nat. Neurosci.* **19**, 158–164 (2016).
55. D. Schwartz, M. Toneva, L. Wehbe, "Inducing brain-relevant bias in natural language processing models" in *Advances in Neural Information Processing Systems (NeurIPS 2019)*. https://github.com/danrsc/bert_brain_neurips_2019. Accessed 6 January 2020.
56. M. Toneva, L. Wehbe, "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)" in *Advances in Neural Information Processing Systems (NeurIPS 2019)* (Neural Information Processing Systems Foundation, Inc., 2019), vol. 32, pp. 14954–14964.
57. J. Hu, J. Gauthier, P. Qian, E. Wilcox, R. P. Levy, A systematic assessment of syntactic generalization in neural language models. arXiv [Preprint] (2020). http://arxiv.org/abs/2005.03692 (Accessed 12 May 2020).
58. S. L. Frank, L. J. Otten, G. Galli, G. Vigliocco, The ERP response to the amount of information conveyed by words in sentences. *Brain Lang.* **140**, 1–11 (2015).
59. J. M. Henderson, W. Choi, M. W. Lowder, F. Ferreira, Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *Neuroimage* **132**, 293–300 (2016).
60. A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, J. L. Gallant, Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
61. A. Lopopolo, S. L. Frank, A. van den Bosch, R. M. Willems, Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PLoS One* **12**, e0177794 (2017).
62. B. Lyu *et al.*, Neural dynamics of semantic composition. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 21318–21327 (2019).
63. T. M. Mitchell *et al.*, Predicting human brain activity associated with the meanings of nouns. *Science* **320**, 1191–1195 (2008).
64. M. J. Nelson *et al.*, Neurophysiological dynamics of phrase-structure building during sentence processing. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E3669–E3678 (2017).
65. J. Pennington, R. Socher, C. D. Manning, "Glove: Global vectors for word representation" in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, 2014), pp. 1532–1543.
66. R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, Y. Wu, Exploring the limits of language modeling. arXiv [Preprint] (2016). http://arxiv.org/abs/1602.02410 (Accessed 15 November 2018).
67. A. Radford *et al.*, Language models are unsupervised multitask learners. arXiv [Preprint] (2019). https://github.com/codelucas/newspaper. (Accessed 8 October 2021).
68. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv [Preprint] (2018). https://arxiv.org/abs/1810.04805 (Accessed 11 October 2018).
69. E. Fedorenko *et al.*, Neural correlate of the construction of sentence meaning. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E6256–E6262 (2016).
70. I. Blank, N. Kanwisher, E. Fedorenko, A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *J. Neurophysiol.* **112**, 1105–1118 (2014).
71. E. Fedorenko, P. J. Hsieh, A. Nieto-Castañón, S. Whitfield-Gabrieli, N. Kanwisher, New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *J. Neurophysiol.* **104**, 1177–1194 (2010).
72. S. Merity, C. Xiong, J. Bradbury, R. Socher, Pointer sentinel mixture model. arXiv [Preprint] (2016). http://arxiv.org/abs/1609.07843 (Accessed 22 May 2017).
73. A. Wang *et al.*, Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv [Preprint] (2019). http://arxiv.org/abs/1804.07461 (Accessed 21 May 2020).
74. A. Warstadt, A. Singh, S. R. Bowman, Neural network acceptability judgments. *Trans. Assoc. Comput. Linguist.* **7**, 625–641 (2019).
75. R. Socher *et al.*, "Recursive deep models for semantic compositionality over a sentiment treebank" in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2013), pp. 1631–1642.
76. W. B. Dolan, C. Brockett, "Automatically constructing a corpus of sentential paraphrases" in *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)* (Asian Federation of Natural Language Processing, 2005), pp. 9–16.
77. D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, L. Specia, "SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation" in *Proceedings of the 11th International Workshop on Semantic Evaluation* (Association for Computational Linguistics, 2018), pp. 1–14.
78. A. Williams, N. Nangia, S. R. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference" in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (Association for Computational Linguistics, 2018), pp. 1112–1122.
79. P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, "SQuad: 100,000+ questions for machine comprehension of text" in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, 2016), pp. 2383–2392.
80. H. J. Levesque, E. Davis, L. Morgenstern, "The winograd schema challenge" in *International Workshop on Temporal Representation and Reasoning* (AAAI, 2012), pp. 552–561.
81. G. B. Keller, T. D. Mrsic-Flogel, Predictive processing: A canonical cortical computation. *Neuron* **100**, 424–435 (2018).
82. Y. Singer *et al.*, Sensory cortex is optimized for prediction of future input. *eLife* **7**, e31557 (2018).
83. R. Futrell *et al.*, "The natural stories corpus" in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)* (European Language Resources Association, 2018), pp. 76–82.
84. K. Rayner, Eye movements in reading and information processing. *Psychol. Bull.* **85**, 618–660 (1978).
85. M. A. Just, P. A. Carpenter, A theory of reading: From eye fixations to comprehension. *Psychol. Rev.* **87**, 329–354 (1980).
86. D. C. Mitchell, "An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading" in *New Methods in Reading Comprehension Research*, K. Keras, M. A. Just, Eds. (Erlbaum, Hillsdale, NJ, 1984), pp. 69–89.
87. V. Demberg, F. Keller, Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* **109**, 193–210 (2008).
88. L. Wehbe *et al.*, Incremental Language Comprehension Difficulty Predicts Activity in the Language Network but Not the Multiple Demand Network. *Cereb. Cortex* **31**, 4006–4023 (2021).
89. D. Merkx, S. L. Frank, Comparing transformers and RNNs on predicting human sentence processing data. arXiv [Preprint] (2020). http://arxiv.org/abs/2005.09471 (Accessed 25 June 2020).
90. E. G. Wilcox, J. Gauthier, J. Hu, P. Qian, R. Levy, On the predictive power of neural language models for human real-time comprehension behavior. arXiv [Preprint] (2020). http://arxiv.org/abs/2006.01912 (Accessed 23 June 2020).
91. A. Goodkind, K. Bicknell, "Predictive power of word surprisal for reading times is a linear function of language model quality" in *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)* (Association for Computational Linguistics, 2018), pp. 10–18.
92. M. van Schijndel, T. Linzen, "A neural model of adaptation in reading" in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2018), pp. 4704–4710.
93. K. Fukushima, Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Netw.* **1**, 119–130 (1988).

94. S. Arora, N. Cohen, E. Hazan, "On the optimization of deep networks: Implicit acceleration by overparameterization" in *Proceedings of the 35th International Conference on Machine Learning* (International Conference on Machine Learning, 2018), pp. 372–389.
95. F. Geiger, M. Schrimpf, T. Marques, J. J. Dicarlo, Wiring up vision: Minimizing supervised synaptic updates needed to produce a primate ventral stream. bioRxiv [Preprint] (2020). https://doi.org/10.1101/2020.06.08.140111 (Accessed 8 October 2021).
96. A. Merchant, E. Rahimtoroghi, E. Pavlick, I. Tenney, What happens to BERT embeddings during fine-tuning? https://aclanthology.org/2020.blackboxnlp-1.4/ (Accessed 8 October 2021).
97. I. Tenney *et al.*, What do you learn from context? Probing for sentence structure in contextualized word representations. arXiv [Preprint] (2019). http://arxiv.org/abs/1905.06316 (Accessed 21 May 2021).
98. K. W. Zhang, S. R. Bowman, "Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis" in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (Association for Computational Linguistics, 2018), pp. 359–361.
99. C. Caucheteux, J.-R. King, Language processing in brains and deep neural networks: computational convergence and its limit. bioRxiv [Preprint] (2020). https://doi.org/10.1101/2020.07.03.186288 (Accessed 8 October 2021).
100. A. Goldstein *et al.*, Thinking ahead: Prediction in context as a keystone of language in humans and machines. bioRxiv [Preprint] (2020). https://doi.org/10.1101/2020.12.02.403477 (Accessed 8 October 2021).
101. J. B. Tenenbaum, C. Kemp, T. L. Griffiths, N. D. Goodman, How to grow a mind: Statistics, structure, and abstraction. *Science* **331**, 1279–1285 (2011).
102. A. Clark, Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).
103. M. H. Christiansen, N. Chater, Toward a connectionist model of recursion in human linguistic performance. *Cogn. Sci.* **23**, 157–205 (1999).
104. M. J. Spivey, M. K. Tanenhaus, Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *J. Exp. Psychol. Learn. Mem. Cogn.* **24**, 1521–1543 (1998).
105. K. McRae, M. J. Spivey-Knowlton, M. K. Tanenhaus, Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *J. Mem. Lang.* **38**, 283–312 (1998).
106. D. L. T. Rohde, D. C. Plaut, Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition* **72**, 67–109 (1999).
107. J. L. Elman, Distributed representations, simple recurrent networks, and grammatical structure. *Mach. Learn.* **7**, 195–225 (1991).
108. J. L. Elman, Learning and development in neural networks: The importance of starting small. *Cognition* **48**, 71–99 (1993).
109. J. L. Elman, Finding structure in time. *Cogn. Sci.* **14**, 179–211 (1990).
110. A. M. Bastos *et al.*, Canonical microcircuits for predictive coding. *Neuron* **76**, 695–711 (2012).
111. R. P. N. Rao, D. H. Ballard, Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
112. D. Mumford, On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol. Cybern.* **66**, 241–251 (1992).
113. M. V. Srinivasan, S. B. Laughlin, A. Dubs, Predictive coding: A fresh view of inhibition in the retina. *Roy. Soc. London. Biol. Sci.* **216**, 427–459 1982
114. G. R. Kuperberg, T. F. Jaeger, What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* **31**, 32–59 (2016).
115. S. M. Garnsey, N. J. Pearlmutter, E. Myers, M. A. Lotocky, The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *J. Mem. Lang.* **37**, 58–93 (1997).
116. M. C. MacDonald, N. J. Pearlmutter, M. S. Seidenberg, The lexical nature of syntactic ambiguity resolution [corrected]. *Psychol. Rev.* **101**, 676–703 (1994).
117. J. C. Trueswell, M. K. Tanenhaus, S. M. Garnsey, Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *J. Mem. Lang.* **33**, 285–318 (1994).
118. J. C. Trueswell, M. K. Tanenhaus, C. Kello, Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *J. Exp. Psychol. Learn. Mem. Cogn.* **19**, 528–553 (1993).
119. M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, J. C. Sedivy, Integration of visual and linguistic information in spoken language comprehension. *Science* **268**, 1632–1634 (1995).
120. G. T. M. Altmann, Y. Kamide, Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* **73**, 247–264 (1999).
121. S. L. Frank, R. Bod, Insensitivity of the human sentence-processing system to hierarchical structure. *Psychol. Sci.* **22**, 829–834 (2011).
122. J. Millet, J.-R. King, Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech. arXiv [Preprint] (2021). http://arxiv.org/abs/2103.01032 (Accessed 11 March 2021).
123. T. H. Heibeck, E. M. Markman, Word learning in children: An examination of fast mapping. *Child Dev.* **58**, 1021–1034 (1987).
124. S. Carey, E. Bartlett, Acquiring a single new word. *Pap. Reports Child Lang. Dev.* **15**, 17–29 (1978).
125. D. K. Dickinson, First impressions: Children's knowledge of words gained from a single exposure. *Appl. Psycholinguist.* **5**, 359–373 (1984).
126. U. Hasson, S. A. Nastase, A. Goldstein, Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron* **105**, 416–434 (2020).
127. F. Ferreira, K. G. D. Bailey, V. Ferraro, Good-enough representations in language comprehension. *Curr. Dir. Psychol. Sci.* **11**, 11–15 (2002).
128. F. Mollica *et al.*, Composition is the core driver of the language-selective network. *Neurobiol. Lang.* **1**, 104–134 (2020).
129. T. M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electron. Comput.* **EC-14**, 326–334 (1965).
130. A. Rahimi and B. Recht, "Random features for large-scale kernel machines" in *Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems* (Neural Information Processing Systems Foundation, Inc., 2007).
131. J. Frankle, G. K. Dziugaite, D. M. Roy, M. Carbin, The lottery ticket hypothesis at scale. arXiv [Preprint] (2019). http://arxiv.org/abs/1903.01611 (Accessed 9 July 2019).
132. N. Cheney, M. Schrimpf, G. Kreiman, On the robustness of convolutional neural networks to internal architecture and weight perturbations. arXiv [Preprint] (2017). http://arxiv.org/abs/1703.08245 (Accessed 27 March 2017).
133. J. Lindsey, S. A. Ocko, S. Ganguli, S. Deny, A unified theory of early visual representations from retina to cortex through anatomically constrained deep cnNs. arXiv [Preprint] (2019). http://arxiv.org/abs/1901.00945 (Accessed 9 January 2019).
134. W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv [Preprint] (2017). http://arxiv.org/abs/1708.08296 (Accessed 25 June 2020).
135. T. Linzen, E. Dupoux, Y. Goldberg, Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Trans. Assoc. Comput. Linguist.* **4**, 521–535 (2016).
136. J. Hewitt, C. D. Manning, "A structural probe for finding syntax in word representations" in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Association for Computational Linguistics, 2019), pp. 4129–4138.
137. I. Tenney, D. Das, E. Pavlick, "BERT rediscovers the classical NLP pipeline" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* Association for Computational Linguistics, 2020), pp. 4593–4601.
138. T. Golan, P. C. Raju, N. Kriegeskorte, Controversial stimuli: Pitting neural networks against each other as models of human recognition. arXiv [Preprint] (2019). http://arxiv.org/abs/1911.09288 (Accessed 10 January 2020).
139. P. Bashivan, K. Kar, J. J. DiCarlo, Neural population control via deep image synthesis. *Science* **364**, 6439 (2019).
140. C. R. Ponce *et al.*, Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell* **177**, 999–1009.e10 (2019).
141. N. J. Majaj, H. Hong, E. A. Solomon, J. J. DiCarlo, Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *J. Neurosci.* **35**, 13402–13418 (2015).
142. Y. Bisk *et al.*, Experience grounds language. arXiv [Preprint] (2020). http://arxiv.org/abs/2004.10151 (Accessed 23 June 2020).
143. A. Bosselut *et al.*, "CoMET: Commonsense transformers for automatic knowledge graph construction" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2020), pp. 4762–4779.
144. M. Sap *et al.*, "ATOMIC: An atlas of machine commonsense for if-then reasoning" in *AAAI Conference on Artificial Intelligence* (Association for the Advancement of Artificial Intelligence, 2019), vol. 33, pp. 3027–3035.
145. M. Sap *et al.*, "Commonsense reasoning about social interactions" in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Association for Computational Linguistics, 2020), pp. 4463–4473.
146. K. Yi *et al.*, "Disentangling reasoning from vision and language understanding" in *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Neural Information Processing Systems Foundation, Inc., 2018), pp. 1039–1050.
147. K. Mahowald, G. Kachergis, M. C. Frank, What counts as an exemplar model, anyway? A commentary on Ambridge (2020). *First Lang.* **40**, 5–6 (2020).
148. H. Lee, J. DiCarlo, Topographic deep artificial neural networks (TDANNs) predict face selectivity topography in primate inferior temporal (IT) cortex (2018). https://doi.org/10.32470/ccn.2018.1085-0. Accessed 8 October 2021.
149. D. J. Felleman, D. C. Van Essen, Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
150. R. Mukamel, I. Fried, Human intracranial recordings and cognitive neuroscience. *Annu. Rev. Psychol.* **63**, 511–537 (2012).
151. J. Parvizi, S. Kastner, Promises and limitations of human intracranial electroencephalography. *Nat. Neurosci.* **21**, 474–483 (2018).
152. M. A. Just, P. A. Carpenter, J. D. Woolley, Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* **111**, 228–238 (1982).
153. A. B. Dieng, F. J. R. Ruiz, D. M. Blei, Topic modeling in embedding spaces. arXiv [Preprint] (2019). http://arxiv.org/abs/1907.04907 (Accessed 15 June 2020).
154. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality. arXiv [Preprint] (2013). http://arxiv.org/abs/1310.4546 (Accessed 15 June 2020).
155. R. Kiros *et al.*, "Skip-thought vectors" in *Advances in Neural Information Processing Systems 28* (Neural Information Processing Systems Foundation, Inc., 2015), pp. 3294–3302.

156. T. Wolf *et al.*, HuggingFace's transformers: State-of-the-art natural language processing. arXiv [Preprint] (2019). http://arxiv.org/abs/1910.03771 (Accessed 15 June 2020).
157. Y. Liu *et al.*, RoBERTa: A robustly optimized BERT pretraining approach. arXiv [Preprint] (2019). http://arxiv.org/abs/1907.11692 (Accessed 15 June 2020).
158. G. Lample, A. Conneau, "Cross-lingual language model pretraining" in *Advances in Neural Information Processing Systems 32* (Neural Information Processing Systems Foundation, Inc., 2019), pp. 7059–7069.
159. A. Conneau *et al.*, Unsupervised cross-lingual representation learning at scale. arXiv [Preprint] (2019). http://arxiv.org/abs/1911.02116 (Accessed 15 June 2020).
160. Z. Yang *et al.*, XLNet: Generalized autoregressive pretraining for language understanding. arXiv [Preprint] (2019). http://arxiv.org/abs/1906.08237 (Accessed 9 July 2019).
161. N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, R. Socher, CTRL: A conditional transformer language model for controllable generation. arXiv [Preprint] (2019). http://arxiv.org/abs/1909.05858 (Accessed 15 June 2020).
162. C. Raffel *et al.*, Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv [Preprint] (2019). http://arxiv.org/abs/1910.10683 (Accessed 7 May 2020).
163. Z. Lan *et al.*, ALBERT: A lite BERT for self-supervised learning of language representations. arXiv [Preprint] (2019). http://arxiv.org/abs/1909.11942 (Accessed 15 June 2020).
164. A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training (2018). https://gluebenchmark.com/leaderboard. Accessed 8 October 2021.
165. J. Bradbury, S. Merity, C. Xiong, R. Socher, Quasi-recurrent neural networks. arXiv [Preprint] (2016). http://arxiv.org/abs/1611.01576 (Accessed 17 November 2016).