Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training

Eghbal A. Hosseini[1,2], Martin Schrimpf[1,2,3], Yian Zhang[4], Samuel Bowman[5], Noga Zaslavsky[1,2,6], Evelina Fedorenko[1,2,3,6,7]

[1] Brain and Cognitive Sciences Department, MIT, Cambridge, MA, 02139
[2] McGovern Institute for Brain Research, MIT, Cambridge, MA, 02139
[3] The MIT Quest for Intelligence Initiative, Cambridge, MA, 02139
[4] Stanford University, Stanford, CA 94305
[5] New York University, New York, NY,10012
[6] K. Lisa Young Integrative Computational Neuroscience Center, MIT, Cambridge, MA, 02139
[7] Speech and Hearing Bioscience and Technology Program, Harvard University, Boston, MA 02115

Acknowledgements:
This work was partially supported by NIH award U01-NS121471 to EF. EAH and MS were supported by the Friends of McGovern graduate fellowships. NZ was supported by a K. Lisa Young ICoN Center postdoctoral fellowship. EF was additionally supported by NIH awards R01-DC016607 and R01-DC016950, as well as by research funds from the McGovern Institute for Brain Research, the Brain and Cognitive Sciences department, the Simons Center for the Social Brain, and the Middleton Professorship. EAH is grateful to Josh McDermott and members of the Fedorenko Lab (especially Carina Kauf, Cory Shain, Greta Tuckute, and Chengxu Zhuang) for helpful discussions and comments on the drafts of the manuscript. EAH is also grateful to Stella Biderman, EleutherAI, with setup in experiment 1, Jason Bolton, Laurel Orr, and Siddharth Karamcheti for help with setup in experiment 2.

Contributions:
EAH, NZ and EF designed research, EAH performed research, MS, YZ, and SB contributed analytic tools, EAH analyzed data with supervision from EF, EAH and EF wrote the paper with input from MS, YZ, SB, and NZ.

Corresponding authors: E. Hosseini (ehoseini@mit.edu), E. Fedorenko (evelina9@mit.edu)
Conflict of interest statement: None
Number of figures: 3 main figures, 8 supplementary figures
Number of pages: 27 pages

Abstract

Artificial neural networks have emerged as computationally plausible models of human language processing. A major criticism of these models is that the amount of training data they receive far exceeds that of humans during language learning. Here, we use two complementary approaches to ask how the models' ability to capture human neural and behavioral responses to language is affected by the amount of training data. First, we evaluate GPT-2 models trained on 1 million, 10 million, 100 million, or 1 billion tokens against two fMRI benchmarks and one behavioral (reading times) benchmark. Because children are exposed to approximately 100 million words during the first 10 years of life, we consider the 100-million-token model developmentally plausible. Second, we test the performance of a GPT-2 model that is trained on a 9-billion dataset to reach state-of-the-art next-word prediction performance against the same human benchmarks at different stages during training. Across both approaches, we find that (i) the models trained on a developmentally plausible amount of data already achieve near-maximal performance in capturing neural and behavioral responses to language. Further, (ii) lower perplexity—a measure of next-word prediction performance—is associated with stronger alignment with the human benchmarks, suggesting that models that have received enough training to achieve sufficiently high next-word prediction performance also acquire human-like representations of the linguistic input. In tandem, these findings establish that although *some* training is necessary for the models' ability to predict human responses to language, a developmentally realistic amount of training (~100 million tokens) may suffice.

Summary

Are artificial neural network (ANN) language models useful as models of human language processing? Some of these models have been shown to capture human responses to language with relatively high accuracy. However, these models are trained on vastly more data than what children are exposed to during language acquisition, raising questions about their value for understanding the human language system. Here, we systematically manipulate the amount of training data that ANN models receive and show that models that are trained on developmentally plausible amounts of language data (approximately 100 million words, roughly corresponding to a child's first 10 years of life) achieve near-maximal performance on human neural and behavioral benchmarks. These developmentally plausible models—rather than models that achieve state-of-the-art performance on the next-word prediction task—hold substantial promise in providing mechanistic-level insights into human language processing.

Introduction

A central objective in cognitive neuroscience is to develop models that can accurately predict human brain responses and behavior. In the neuroscience of language, some artificial neural network (ANN) language models were recently shown to be effective for explaining human brain activity and behavior during language processing (Caucheteux & King, 2022; Schrimpf et al., 2021; Toneva & Wehbe, 2019; Gauthier & Levy, 2019; Wilcox et al., 2020). For example, Schrimpf et al. (2021) examined the ability of 40+ language models to capture human responses to language and found that transformer architectures (Radford et al., 2019; Vaswani et al., 2017) fare best in aligning with human data. However, off-the-shelf models vary along many dimensions, making it difficult to unambiguously attribute any given model's success in aligning with human data to particular model properties (architecture, objective function, amount/kind of training data, etc.). Gaining insights into human linguistic mechanisms requires controlled 'experiments' on the models, where different properties are systematically manipulated (Hu et al., 2020; Kumar et al., 2022; Warstadt & Bowman, 2019). This is the approach we adopt here in order to investigate how the amount of training data affects model-to-human alignment.

One common criticism of ANN models as models of human language processing is that their training data size (often, billions of words) far surpasses the amount of language exposure that children get (Chang & Bergen, 2021; Dupoux, 2018; Linzen & Leonard, 2018; van Schijndel et al., 2019); see Warstadt & Bowman, 2022 for discussion). Hart & Risley (1992) estimated that children are exposed to 3-11 million words each year, so by the time they turn 10 and acquire adult-like linguistic competence, they are exposed to 30-110 million words. In contrast to a human child, who can learn a language from only ~100 million words (or less), some current models get orders of magnitude more training data (20,000 human years worth for some models; Warstadt and Bowman 2022). Here, we ask whether this extensive training is necessary for the models to acquire brain-like representations.

Prior studies on the effects of training data on the models' linguistic ability found that even with limited amounts of training data, models achieve considerable proficiency (Warstadt and Bowman, 2022). For example, Hu et al. (2020) and Zhang, Warstadt et al. (2021) report impressive syntactic generalizations in a BERT model (Devlin et al., 2018) trained on only millions of tokens (see also Huebner & Willits, 2021; Pannitto & Herbelot, 2020 for related evidence from a RoBERTa model trained on 5 million words of child-directed speech). And Pérez-Mayos et al. (2021) find that a RoBERTa model (Liu et al., 2019) trained on 100 million tokens performs similarly on several syntactic benchmarks to a model trained on 1 billion words. These findings suggest that massive amounts of training may not be necessary for models to acquire certain aspects of linguistic competence. However, it is not known whether models trained on limited amounts of data can also explain human responses to language.

Here, we evaluate how the amount of training data affects model-to-human alignment. In line with increasing emphasis in the field on robustness and replicability (Button et al., 2013; Ioannidis et al., 2014; Poldrack et al., 2017; Simmons et al., 2011), we adopt two complementary approaches (**Figure 1**). First, we investigate how well GPT-2 models (Radford et al., 2019), trained on different-size datasets (1, 10, 100 million, or 1 billion) to reach their best training task performance, predict human data. Second, we investigate how a GPT-2 model's ability to predict human responses changes over the course of training on a large dataset in an effort to capture the 'developmental trajectory' of model-to-brain alignment. In addition, we also probe the role of model perplexity in the ability of a model to develop human-like representations of the linguistic input. To foreshadow the key result, we find that models reach

high performance in predicting human responses to language even with realistic amounts of training data.

Methods

Experimental Design

Human datasets (benchmarks)

The human datasets used here were identical to those used in Schrimpf et al. (2021). We describe them here briefly.

Neural dataset 1: fMRI (Pereira 2018)

We used data from two experiments in Pereira et al. (2018). Experiment 2 (n=9) consisted of 384 sentences across 96 passages. Experiment 3 (n=9) consisted of 243 sentences across 72 passages. In both experiments, each sentence was presented on the screen for 4 seconds, followed by 4 seconds of fixation, and each participant viewed the set of sentences 3 times across three fMRI scanning sessions. The stimuli for both experiments were designed to span a broad range of contents.

Neural dataset 2: fMRI (Blank 2014)

We used the data from Blank et al. (2014) for 5 out of 10 participants that were exposed to the same materials. Participants listened to stories from the Natural Stories corpus (Futrell et al., 2018), each around 5 minutes long. These stories contain a high number of rare words and syntactic constructions in natural linguistic contexts.

Behavioral dataset (Futrell 2018)

We used the self-paced reading data from (Futrell et al. 2018). 179 participants read stories presented one word at a time (e.g., Just et al., 1982), and with each button press, the current word would disappear in place of the new word. The time between each key press was used as a measure of comprehension difficulty. Each participant read between 5 and 10 stories and answered comprehension questions at the end of each story. We excluded reading times outside of the [100ms, 3000ms] window.

Artificial Neural Network Models

We used two different implementations of a GPT-2-style model. For Experiment 1, where a model was trained on a dataset with a controlled number of tokens, we used the GPT-NEOX library which is a distributed training framework that uses the DeepSpeed library (Black et al. 2022; Aminabadi et al. 2022). We used a unidirectional-attention transformer model (GPT-2; Radford et al. 2019) with 12 layers and an embedding layer which was learned during training. Each layer had a size of 768 units and consisted of 4 main blocks (**Figure 1A**): (i) 1st layer normalization, (ii) self-attention, (iii) 2nd layer normalization, and (iv) the feedforward layer. The final layer consisted of a linear projection with a sigmoid nonlinearity that mapped hidden states into probabilities over the dictionary. The context size was 1,024 tokens. To see if our results would generalize to bidirectional-attention transformer architectures, we additionally used publicly available miniBERTa models[1] trained on the same datasets as the GPT-2 models

[1] https://huggingface.co/nyu-mll

(Zhang et al., 2020). (We did not include the model trained on the smallest (1 million tokens) dataset, for which Zhang et al. (2020) used a smaller-size model, which therefore would not be directly comparable to the other models.) The miniBERTas use the same design as the RoBERTa 'base' model (Liu et al. 2019)—a bidirectional-attention model with 12 layers, each 768 units in size, and a context size of 512 tokens. Importantly, RoBERTa has the same number of parameters as GPT-2 (125 million), allowing for a relatively controlled comparison of uni- and bidirectional architectures.

For Experiment 2, to investigate model training dynamics with a very large dataset, where during the early stages of the training the model continues to see new input (cf. doing multiple passes through a smaller-size training corpus as in Experiment 1), we used GPT-2 model weights from a publicly available model from the HuggingFace Transformers library (https://huggingface.co/stanford-crfm). The model has a similar architecture to the GPT-2 model used in Experiment 1.

Training

Training dataset

For Experiment 1, we combined the BookCorpus (Zhu et al., 2015) and English Wikipedia (Liu et al., 2019; Zhu et al., 2015) with a 1:3 ratio. We then created 4 different datasets with 1 million, 10 million, 100 million, and 1 billion words. These were used for training both the GPT-2 models and the minBERTa models. For Experiment 2, we used a model that was trained on the OpenWebText corpus (Gokaslan & Cohen, 2019) with close to 10 billion words.

Training procedure

For Experiment 1, to train the GPT-2 models, we used standard initialization from the GPT-NEOX library and standard training parameters (Radford et al., 2019; see **Suppl. Figure 1** for details). After training, the model weights with the smallest validation perplexity were selected for evaluation on the fMRI and behavioral benchmarks. To train the miniBERTa models, we used standard initialization and training parameters from the Hugging Face Transformers library (Liu et al. 2019).

Additionally, we created another untrained version of the GPT-2 model in order to investigate the effects of different initializations on the model-to-human alignment and thus to isolate the effects of model architecture alone (i.e., the units and the patterns of connections among them) on brain predictivity. This version implemented the same unidirectional mask as the trained models, but all the weights were set to a gaussian distribution with a fixed mean and standard deviation (mean: 0, standard deviation: 0.02 for the layer normalization, self-attention, and feedforward layer weights; see **Supp. Figure 5** for a detailed comparison with the Hugging Face initialization parameters).

For Experiment 2, the GPT-2 model was trained with standard initialization and training parameters until it reached state-of-the art perplexity values. We selected several checkpoints at which we extracted model representations for evaluation on the human benchmarks. The checkpoints were selected in a logarithmic manner (0, 0.01, 0.1, 1.0, 10.0, and 100% of training steps. Based on Radford et al. (2019), the size of the dataset used in training GPT-2 is estimated at 40 billion tokens; given the batch size used for training (512 tokens), the context size (1,024 tokens), and the total number of training steps (400,000), 100% of training represents about 5 complete passes over the training data (and only in the 100% condition

does the model see any training sequence more than once), and 10% (or fewer) of training steps does not go fully over the training data, so the model continues to receive new input. It is important to note that unlike Experiment 1, in Experiment 2 models are continuously exposed to new samples over batches (rather than doing multiple passes through a smaller-size training corpus as in Experiment 1), making it not straightforward to draw a parallel between the amounts of data used for the two experiments.

Analyses

Model comparison to the fMRI benchmarks

Following Schrimpf et al. (2021), from each layer of each model, we first extracted the representation for all the stimuli that were used in the human fMRI experiments. We then split the data into 5 equal-size batches and used 80% of the data to build a regression model between model activations and responses in the language network (individual voxel responses for Pereira et al., 2018)or responses in the brain regions of interest (ROIs) for Blank et al., 2014), as defined by an extensively validated language 'localizer' paradigm (Fedorenko et al., 2010; Lipkin et al., 2022). We then used the remaining 20% of the data to generate predictions for unseen stimuli, and these predictions were compared with the brain measurements for the same stimuli using Pearson correlation. This was done for each participant separately, resulting in a score per participant based on the median (over voxels/ROIs) model-to-brain correlation. We then computed a median across participants and divided it by an estimated ceiling value to get a normalized score (see SI section 7 in Schrimpf et al., 2021 for additional details).

Model comparison to the behavioral benchmark

Following Schrimpf et al. (2021), from the last layer of each model, we extracted the representation for all the stimuli that were used in the behavioral experiment. We then built a regression model and computed correlations between model predictions and human behavioral responses for unseen stimuli, in a similar manner to the fMRI benchmarks.

Model perplexity

Following standard practice (e.g., Jelinek et al., 1977), we used perplexity as a measure of model performance on the language prediction tasks (next-word prediction for the GPT-2 models and missing-word prediction for the miniBERTa models). For both experiments, we used the test set from the wikitext-103-raw-v1 dataset (Merity et al., 2016) to compute perplexity. Perplexity was computed using a context size of 1,024 tokens and a stride of 512 tokens.

Code Accessibility

The human benchmarks and the code for relating model representations to the benchmarks are publicly available at https://github.com/mschrimpf/neural-nlp. For Experiment 1, the GPT-2 models and the representations extracted for all the benchmarks are available upon request (and will soon be uploaded to the Hugging Face library); the miniBERTa models are available upon request; the checkpoints are available at https://huggingface.co/nyu-mll. (The training corpora used for Experiment 1 have copyright constraints so cannot be made publicly available.) For Experiment 2, the model checkpoints are available at: https://huggingface.co/stanford-crfm; the training corpus is available through the Hugging Face library.

Results

1a. Models trained on relatively small amounts of training data can predict human neural and behavioral responses to language

We started by examining the performance of a unidirectional transformer model trained on the standard language modeling task in predicting human responses during language processing. Specifically, we tested a GPT-2 architecture (**Figure 1A**), which has previously been shown to best capture human neural and behavioral responses (e.g., Schrimpf et al. 2021). We trained four independent models on 1 million, 10 million, 100 million, and 1 billion tokens, respectively (**Suppl. Figure 1**). This range includes the estimated level of language exposure during early life: ~100 million (Hart & Risley, 1992). After training, we selected the checkpoint with the best perplexity on the validation set and tested how well the model representations capture human neural responses in the language-selective network (Fedorenko et al., 2011) and behavioral responses (**Figure 1B**).

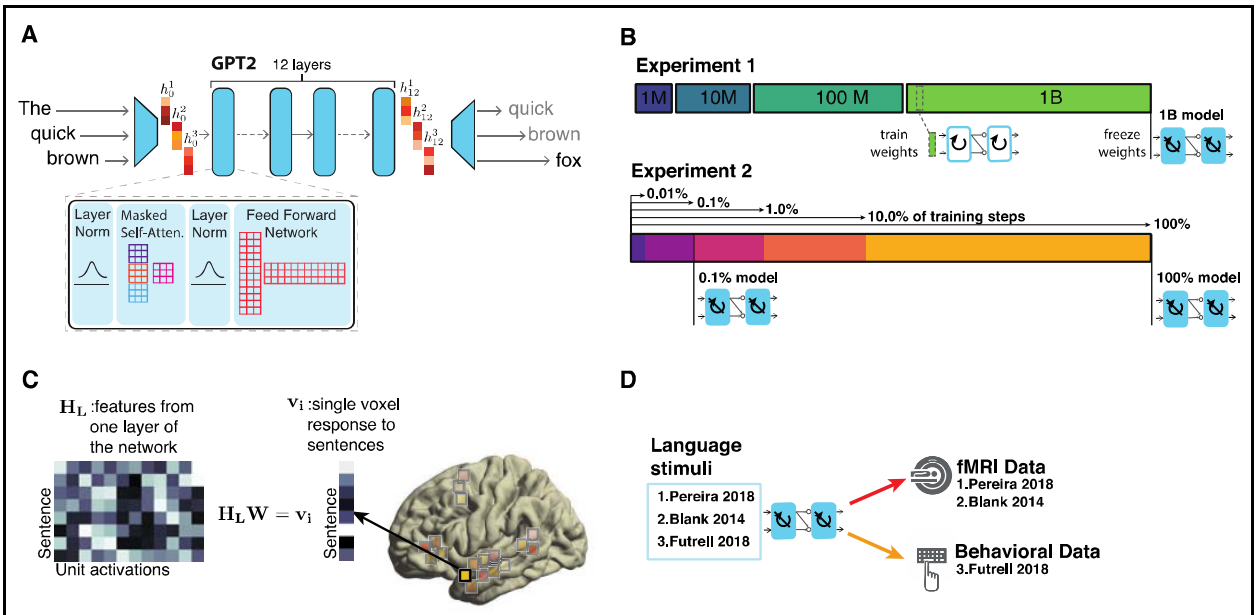


Figure 1. Methodological approach. A. Unidirectional-attention transformer architecture. Text input is processed sequentially to predict the next likely word at each step. **B.** The set-up for Experiments 1 and 2. In Experiment 1, four models were trained using different-size datasets, and for each model, the weights with the best validation perplexity were frozen and used in the model-to-brain comparison. In Experiment 2, the GPT-2 model was trained using a very large dataset, and the weights were frozen at different steps during training and used in the model-to-brain comparison. **C.** Model representations were related to human representations by building a linear regression between unit activations for each layer of the model and voxel/region activity (in the language-selective network; Fedorenko et al., 2011) or reading times for the stimuli used in each of the benchmarks. This regression was then used to make predictions about human neural/behavioral responses for unseen language stimuli, and a Pearson correlation was computed between these predictions and the observed responses. **D.** The general pipeline for predicting human brain and behavioral responses. For each benchmark, each model was exposed to the same language stimuli as humans, and the model-to-brain match was evaluated as shown in C.

For the Pereira2018 fMRI benchmark (Pereira et al. 2018; Schrimpf et al. 2021), we observed a consistent increase in performance with an increase in the size of the training set (**Figure 2A**; see **Suppl. Figure 4** for evidence—for this and the other benchmarks—of robustness of this pattern to seed choice during model initialization; cf. Mehrer et al., 2021). Critically, however, a model trained on just 100 million tokens already exhibits brain predictivity similar to that reported in Schrimpf et al. (2021) for the fully trained GPT-2 model. The untrained model performance differs between the two initializations (see Methods - Training Procedure; **Supp. Figure 5**). The version initialized with the standard Hugging Face parameters performs well above chance, as reported in Schrimpf et al. (2021; see also Caucheteux & King, 2022). However, the version initialized with the alternative parameters (all weights set to a normal distribution with a mean of 0 and a standard deviation of 0.02) performs around 0 (**Figure 2A**).

These results generalize to the Blank2014 fMRI benchmark, except that the model trained on 10 million tokens exhibits a drop relative to the model trained on 1 million tokens, and then, for 100 million tokens, the model performance recovers (**Figure 2B**). For the Futrell2018 behavioral benchmark, the results are similar to those observed for the Pereira2018 benchmark but plateauing in performance earlier, at the model trained with only 10 million tokens (**Figure 2C**).

For the Pereira2018 and Futrell2018 benchmarks, some aspects of the results also generalize to a bidirectional transformer model (miniBERTa; Liu et al. 2019) (**Suppl. Figure 2**). In particular, similar to the GPT-2 models, we observed a consistent increase in model performance with an increase in the training dataset size, which suggests that this pattern is robust to architecture. However, for the Pereira2018 benchmark, the 100-million-token model still performs below the fully trained model reported in Schrimpf et al. (2021). For the Blank2014 fMRI benchmark, even the 1-billion-token model performs well below the fully trained model as reported in Schrimpf et al. (2021). This difference between the GPT-2 and miniBERTa models in the amount of training they require to align with human data is likely due to the difference in the directionality of the attention mechanisms, with unidirectional-attention mechanisms being more sample efficient. Generalizing these results to other minimally different variants of uni- vs. bidirectional-attention transformer models will help strengthen this conclusion.

In exploratory analyses, we investigated the patterns of model-to-brain alignment across model layers. Prior work in vision (Geiger et al., 2020; Storrs et al., 2021) has suggested that training affects model performance differently across layers, with early layers already reaching close to maximal performance with a limited amount of training, but later layers continuing to benefit from increasingly more training. In line with these prior observations, for the Pereira2018 benchmark, we observed that for layers 4-9, performance peaks for the model trained on 1 million tokens, and for the last three layers (layers 10-12), a consistent improvement in performance is observed with larger datasets (**Figure 2G**). The pattern is similar for the Blank2014 benchmark (**Suppl. Figure 3A**), with layers 3-9 peaking for the model trained on 1 million tokens, and the last three layers showing better performance for models trained on larger datasets. (The pattern observed for the first three layers is less clear and varies between the two benchmarks.)

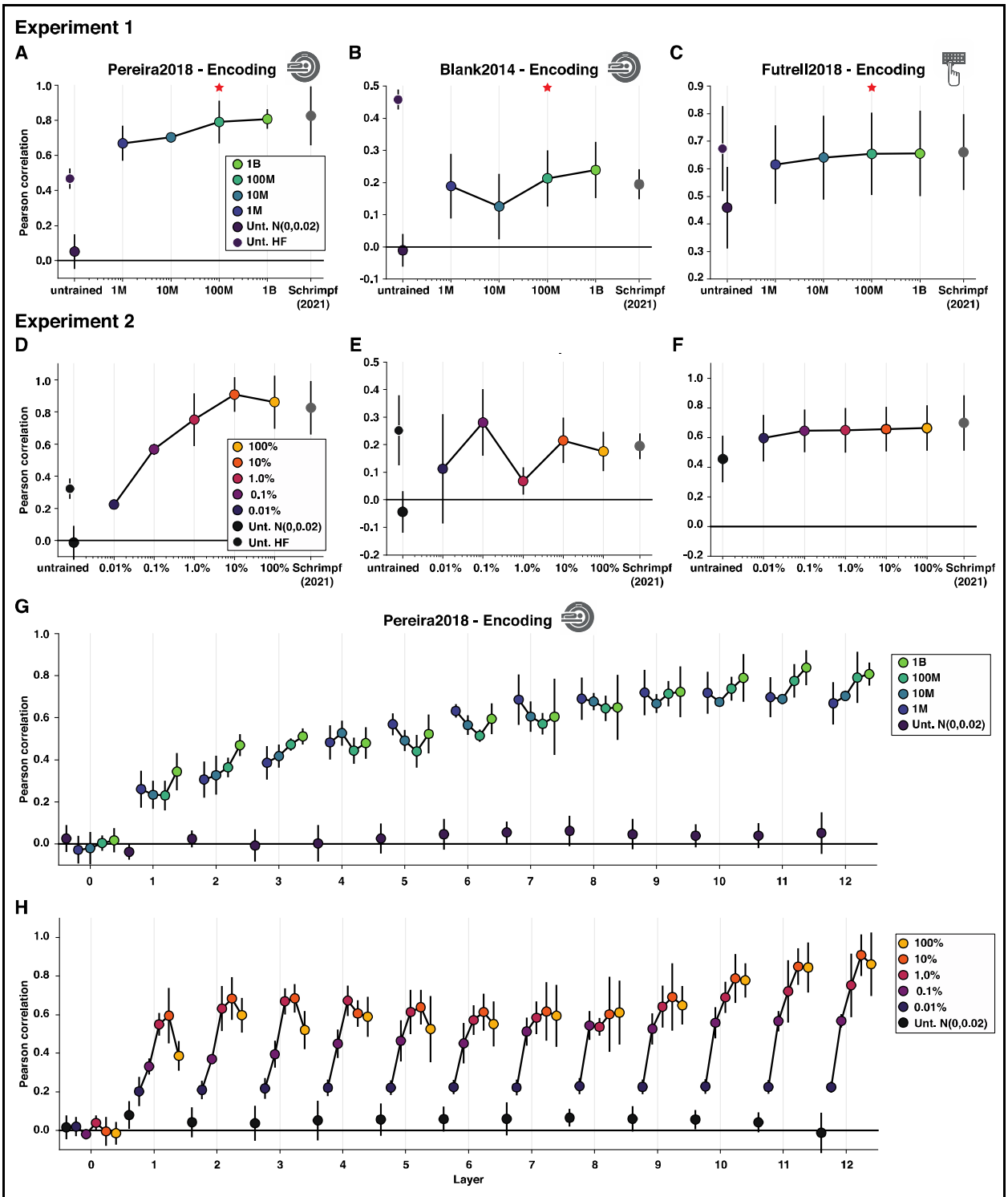


Figure 2. Model performance on the neural and behavioral human benchmarks as a function of training.

A-C. Experiment 1 results: performance of the best-performing GPT-2 layer, as reported in Schrimpf et.al. (2021), in predicting language-responsive voxels' activation in the Pereira2018 fMRI benchmark (A), language regions' responses in the Blank2014 fMRI benchmark (B), and reading times in the Futrell2018 behavioral benchmark (C). The results are shown for i) two

versions of an untrained (Unt.) model (initialized in two different ways; see Methods; Unt. N(0,0.02) corresponds to the untrained model initialized with a mean of 0 and a standard deviation of 0.02; and Unt. HF corresponds to the untrained model initialized with the Hugging Face parameters) (black dots); ii) four models trained on datasets of different sizes (1M, 10M, 100M, and 1B tokens) (blue-to-green dots connected by a line; the model trained on a developmentally plausible amount of data—100M—is marked with a red asterisk); and iii) a fully trained model, as reported in Schrimpf et al. (2021) (grey dots).
D-F. Experiment 2 results: performance of the best-performing GPT-2 layer, as reported in Schrimpf et al. (2021), in predicting human responses in the Pereira2018 fMRI benchmark (D), the Blank2014 fMRI benchmark (E), and the Futrell2018 behavioral benchmark (F). The results are shown for i) two versions of an untrained model (initialized in two different ways; see Methods) (black dots); ii) a model trained on a large dataset examined at different points during the training (0.01%, 0.1%, 1%, 10%, and 100% of training steps) (purple-to-yellow dots connected by a line); and iii) a fully trained model, as reported in Schrimpf et al. (2021) (grey dots).
G-H. Exploratory analyses of individual model layers: performance of the 12 GPT-2 model layers in predicting human neural responses in the Pereira2018 fMRI benchmark in Experiment 1 (G) and Experiment 2 (H). The results are shown for i) an untrained model (black dots); and ii) four models trained on datasets of different sizes (blue-to-green dots connected by a line in G) or a model trained on a large dataset at different points during the training (purple-to-yellow dots connected by a line in H). (Layer 0 is the token embedding layer, and layer 12 is the last transformer layer.)

1b. In the presence of a large amount of training data, models only need a small amount of training to predict human data

In the previous section, we investigated how models that are trained on a limited amount of data (until they reach their best performance on the target language modeling task) perform in predicting human data. However, humans, including children learning a language, are continuously exposed to new words and constructions. To better simulate such scenarios, as well as to evaluate the robustness of the results to approach, we examined how the ability of a model to predict human responses to language changes over time as the model is being trained on a very large dataset. To do so, we used a GPT-2 model that was trained on a dataset with over 9 billion tokens and selected several checkpoints during the training process (0.01, 0.1, 1.0, 10.0, and 100% of training steps, where 100% of training steps corresponds to 3 complete passes over the training data). At each of these checkpoints, we tested how well the model representations capture human responses to language.

For the Pereira2018 fMRI benchmark, the performance of the fully trained model (i.e., 100% of training steps) closely matches the results reported in Schrimpf et al. (2021), which suggests that model-to-human alignment is robust to the details of model implementation (as one would hope). Critically, mirroring the results from Experiment 1, we observed a consistent and nearly linear (on the log scale) increase in how well the model predicts neural data until the model reaches the 10% checkpoint, at which point the performance plateaus (**Figure 2D**; see **Suppl. Figure 4** for evidence of robustness to seed choice during model initialization). The slight decrease in performance with more training suggests that more training does not necessarily lead to better alignment with human brain data, although it is possible that this result is due to the relatively spatially and temporally coarse nature of our neural measurements (i.e., fMRI

recordings) and that for finer-grained neural data, we might continue to see improvements with more training.

For the Blank2014 fMRI benchmark, we observed a sharp increase in performance with only 0.1% of training steps, followed by a drop at the 1% checkpoint and recovery of performance for the 10% and 100% checkpoints (**Figure 2E**). (The drop in performance after initial training resembles the drop we observed for a model trained on 10 million tokens in Experiment 1.) Similar to the Pereira2018 benchmark, the performance of the fully trained model closely replicates the results reported in Schrimpf et al. (2021). Critically, this performance level is lower than model performance at the 0.1% checkpoint. The Futrell2018 behavioral benchmark closely follows the pattern we observed with limited-size training datasets in Experiment 1, plateauing after the 0.1% checkpoint (**Figure 2F**).

In exploratory analyses of the individual model layers, we observed that for the Pereira2018 benchmark, performance is consistent across layers, with all layers showing a drop in performance after 10% of the training steps (**Figure 2H**). Additionally, as in Experiment 1 and in line with prior work in vision (e.g., Storrs et al. 2021; Geiger et al. 2022), early layers reach close to maximal performance earlier in the training (at the 1% checkmark) whereas later layers reach their peak close to the 10% checkmark (**Figure 2G**).

2. Model perplexity predicts model performance on neural and behavioral benchmarks

For ANN language models, perplexity (a measure of performance on the next-word prediction task) is a reliable predictor of model performance on diverse NLP benchmarks (e.g., Radford et al. 2019; Brown et al. 2020). Schrimpf et al. (2021) further found that (off-the-shelf) models that perform better on the next-word prediction task are also better able to capture human neural and behavioral responses (cf. Pasquiou et al., 2022). Here, we examined the relationship between model perplexity and model ability to predict human data for models that only differ in the size of the training corpus and for a model at different stages of training, in order to test whether better performance on the next-word prediction task is associated with more human-like representations.

As expected, perplexity is lower (i.e., the ability to predict upcoming words is better) for models that are trained on larger datasets (**Figure 3A-C**) and for a given model at the later stages of training (**Figure 3D-F**). Critically, for the Pereira2018 fMRI benchmark and the Futrell2018 behavioral benchmark, across both Experiments 1 and 2, we observed a consistent relationship between perplexity and neural/behavioral predictivity, such that lower perplexity is associated with higher predictivity (**Figure 3A,C,D,F**). For the Blank2014 fMRI benchmark, the perplexity-predictivity relationship is less stable, reflecting the drop in predictivity early in the training (for the model trained on 10 million tokens in Experiment 1, and at the 1% checkmark in Experiment 2); importantly, this drop in predictivity is not due to an increase in perplexity (**Figure 3B,E**).

We speculate that model perplexity may not be strongly predictive of human responses in the low-value range (e.g., **Figures 3D-E**) because the models may enter a state where they surpass humans in next-word prediction performance[2], which could negatively affect their ability to capture human neural/behavioral data.

[2] https://www.alignmentforum.org/posts/htrZrxduciZ5QaCjw/language-models-seem-to-be-much-better-than-humans-at-next

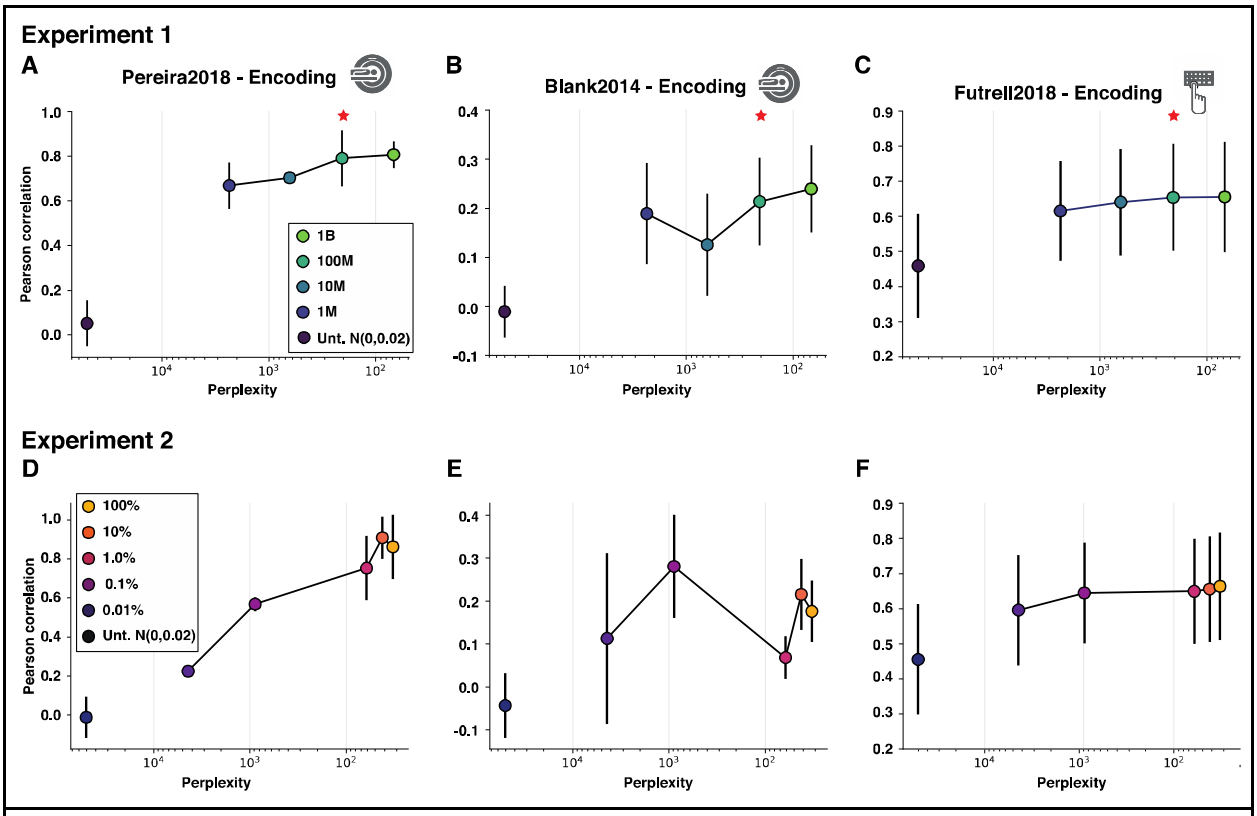


Figure 3. Relationship between model perplexity and model ability to predict human data.

A-C. Experiment 1 results: the relationship between perplexity, i.e., the model's ability to predict the next word on an independent dataset (wikitext-103-raw-v1), shown on the x-axis, with lower values corresponding to better ability, and model performance in predicting language-responsive voxels' activation in the Pereira2018 fMRI benchmark (A), language regions' responses in the Blank2014 fMRI benchmark (B), and reading times in the Futrell2018 behavioral benchmark (C). The results are shown for i) an untrained model (black dots); and ii) four models trained on datasets of different sizes (1M, 10M, 100M, and 1B tokens) (blue-to-green dots connected by a line; the model trained on a developmentally plausible amount of data—100M—is marked with a red asterisk).

D-F. Experiment 2 results: the relationship between perplexity (shown on the x-axis, lower values correspond to better ability) and model performance in predicting human responses in the Pereira2018 fMRI benchmark (D), the Blank2014 fMRI benchmark (E), and the Futrell2018 behavioral benchmark (F). The results are shown for i) an untrained model (black dots); and ii) a model trained on a large dataset examined at different points during the training (0.01%, 0.1%, 1%, 10%, and 100% of training steps) (purple-to-yellow dots connected by a line).

Discussion

In this work, we investigated the relationship between the amount of training data and brain predictivity for state-of-the-art artificial neural network (ANN) language models. Our study makes several contributions, as summarized next.

Even when trained on a developmentally realistic amount of data, ANN language models align with human data.

Using two fMRI benchmarks and a behavioral benchmark, we established that even with a realistic amount of training data (~100 million words, comparable to what humans get during the first 10 years of life; Hart & Risley, 1992), a GPT-2 model achieves near-maximal brain predictivity. This effect mostly generalizes to bidirectional-attention transformers (miniBERTa), although compared to GPT-2, such models appear to be less sample-efficient, requiring more training data to achieve peak predictivity. In a complementary approach, we showed that when trained on a large dataset, a GPT-2 model already achieves near-maximal predictivity with only 10% of training steps, well before a full pass over the dataset.

These results align with prior work in vision. For example, Geiger et al. (2020) found that even a small amount of training can result in model representations that are predictive of neural responses in macaques. Moreover, the logarithmic nature of the increase in brain predictivity between a model trained on 1 million tokens and a model trained on 1 billion tokens aligns with prior Natural Language Processing (NLP) results (e.g., see Kaplan et al., 2020 for evidence of a logarithmic relationship between training data size and the loss in training, and between model size and loss), as well as with vision research (e.g., see Geiger et al. 2020 for evidence of a logarithmic relationship between training data size and brain predictivity).

The key implication of these findings is that although state-of-the-art language models are trained on vast amounts of data (and performance on some NLP benchmarks continues to improve with more training), this large amount of training is not necessary for these models to acquire human-like representations. The fact that ANN models trained on a developmentally plausible amount of data can accurately capture responses to language helps address one of the most common criticisms of these models as models of human language processing.

Alignment between untrained ANN language models and human neural data is strongly affected by the initial unit weight configuration.

By relating different versions of untrained models to human data, this work clarifies the contributions of architecture to brain predictivity. Schrimpf et al. (2021; see also Caucheteux & King, 2022; Pasquiou et al., 2022) have found that untrained models predict neural data quite well, albeit worse than trained models. They speculated that good performance of untrained models might be due to the smoothing of word embeddings across layers in a way that enables the embeddings to capture some aspects of statistical regularities of language (perhaps something as general as nearby words being likely to be related to one another). However, what counts as 'untrained' is important to clarify.

'Untrained' models come with a particular setting of their unit weights. A particular weight configuration may get 'baked into' a model during the process of model development, aimed at maximizing learning efficiency for the target task. Such potential 'biases' in initial, pre-trained weights may be akin to innate, evolution-shaped, aspects of brain structure, which may filter information in specific ways as it travels within or across brain areas, even before any learning

of the input regularities has occurred (e.g., Zador, 2019). We showed that initializing a model with a normal distribution for all weights leads to the model being unable to predict neural data (predictivity is at ~0; of course, such a model is also unable to perform the next-word prediction task). (This inability to predict neural data for models initialized with a normal distribution is not due to the lack of activity propagation across layers, as shown in **Suppl. Figure 7**).

In summary, reliable brain predictivity reported for untrained models in previous studies should not be taken as evidence that model architecture alone (i.e., the units and the patterns of connections among them) can capture human responses to language, or at least, it should be acknowledged that these effects are due to the particular pre-trained *weight configurations*. Furthermore, if a model can (at least partially) match human data with a few bits of information in the form of the initialization parameters (see **Suppl. Figure 7** for evidence that above-baseline human predictivity for some initializations may result from the representations for different sentences being more similar), then any results at that alignment level or below for trained models are not meaningful and we should focus on progress beyond that alignment level. Another important implication is that future attempts to align ANN models with human data should generalize their findings across different weight initializations.

Model perplexity predicts brain scores.

In line with Schrimpf et al.'s (2021) claim that models that perform better on next-word prediction are better at predicting brain data (see also Caucheteux & King, 2022), we found that model perplexity for different amounts of the training data is a good proxy for model performance in predicting human data. We observed this relationship both in Experiment 1, where we varied the size of the training dataset, and in Experiment 2, where we tested model representations at different points during the training on a large training dataset. These findings provide further evidence that optimizing for predictive representations—through training the models on the next word prediction task—may be critical for ANN models to acquire human-like representations.

One recent study (Pasquiou et al., 2022) did not observe a relationship between perplexity and model performance on human neural data. We speculate that the lack of this relationship in Pasquiou et al.'s data may be because of the use of an extended-narrative stimulus (the entire 'The Little Prince' book) rather than single sentences or short passages. In our experiments, the relationship between perplexity and brain predictivity is also weakest for the Blank2014 fMRI benchmark, which uses story stimuli. Why might this relationship be weak or non-existent for long narratives? One possible explanation is that the overall low encoding performance for such stimuli imposes a ceiling on the relationship between model-to-brain alignment and model perplexity (or other variables) (cf. Oh et al., 2022 for another hypothesis having to do with humans and models using different information for predicting upcoming words, especially in extended linguistic stimuli).

Why models struggle with predicting neural responses to long narratives is a separate and important question. We offer a speculation. In the human brain, division of labor exists between i) the language-selective network, which integrates information within clauses/sentences but does not track longer-range contexts (e.g., I. A. Blank & Fedorenko, 2020), and ii) the Default network(s) (Buckner & DiNicola, 2019), which integrates information over extended temporal contexts (Lerner et al., 2011). Importantly, the Default network does not operate over word sequences; instead, the information that this system represents is likely abstract, as evidenced by the fact that it processes long contexts in both linguistic and non-linguistic stimuli (e.g., Baldassano et al., 2017; Simony et al., 2016). As a result, the ANN language models (like those used in current work and in Pasquiou et al., 2022) may simply lack representations that are

sufficiently abstract (removed from the stimulus) to match those in the Default network. Some of the newer models, like GPT-3, seem to be able to handle a greater degree of abstraction (e.g., Brown et al., 2020) and thus may be promising for future attempts to capture human neural responses to long and complex linguistic stimuli.

Limitations and future directions

We here have focused on the effects of the *amount* of training data on the ANN language models' ability to capture human responses. However, the *nature* of the training data is also likely important. For example, training models on data that are similar to what children are exposed to could lead to improved neural predictivity (Chang and Bergen 2021; Warstadt and Bowman 2022). Indeed, this approach has been shown to improve vision models' ability to capture primate neural responses (Mehrer et al. 2021). Further, it will be important to investigate the role of the *learning algorithms* that the models use and their *training objective*, as both likely affect the representations that the models learn (e.g., see Zhuang et al., 2022 for evidence from vision).

Another aspect of the ANN models that is important for building accurate models of human language processing is the model *architecture*. We here generalized our training effects across uni- and bidirectional-attention transformers, but a systematic investigation of the effects of diverse architectural parameters (e.g., the number and size of layers, number of attention heads, etc.) on the models' ability to predict human data would be valuable. Tightly controlled comparisons between different classes of model architectures are more challenging but creating numerous model variants all trained on the same dataset (e.g., Storrs et al. 2021) could enable identification of architectural motifs that are essential for a good match with human neural and behavioral data.

In future work, we aim to address these gaps to build increasingly more accurate and interpretable models of language processing in the brain.

References

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron*, *95*(3), 709-721.e5.

Blank, I. A., & Fedorenko, E. (2020). No evidence for differences among language regions in their temporal receptive windows. *NeuroImage*, *219*, 116925.

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.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2005.14165

Buckner, R. L., & DiNicola, L. M. (2019). The brain's default network: updated anatomy, physiology and evolving insights. *Nature Reviews. Neuroscience*, *20*(10), 593–608.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience*, *14*(5), 365–376.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134.

Chang, T. A., & Bergen, B. K. (2021). Word Acquisition in Neural Language Models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2110.02406

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1810.04805

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59.

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(39), 16428–16433.

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S., & Fedorenko, E. (2018, May). The Natural Stories Corpus. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. https://aclanthology.org/L18-1012

Geiger, F., Schrimpf, M., Marques, T., & DiCarlo, J. J. (2020). Wiring Up Vision: Minimizing Supervised Synaptic Updates Needed to Produce a Primate Ventral Stream. In *bioRxiv* (p. 2020.06.08.140111). https://doi.org/10.1101/2020.06.08.140111

Gokaslan, A., & Cohen, V. (2019). *OpenWebText Corpus*.

Hart, B., & Risley, T. R. (1992). American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. In *Developmental Psychology* (Vol. 28, Issue 6, pp. 1096–1105). https://doi.org/10.1037/0012-1649.28.6.1096

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

Huebner, P. A., & Willits, J. A. (2021). Scaffolded input promotes atomic organization in the recurrent neural network language model. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 408–422.

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241.

Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, *62*(S1), S63–S63.

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.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2001.08361

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (2022). Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. In *bioRxiv* (p. 2022.06.08.495348). https://doi.org/10.1101/2022.06.08.495348

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(8), 2906–2915.

Linzen, T., & Leonard, B. (2018). Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1807.06882

Lipkin, B., Tuckute, G., Affourtit, J., Small, H., Mineroff, Z., Kean, H., Jouravlev, O., Rakocevic, L., Pritchett, B., Siegelman, M., Hoeflin, C., Pongos, A., Blank, I. A., Struhl, M. K., Ivanova, A., Shannon, S., Sathe, A., Hoffmann, M., Nieto-Castañón, A., & Fedorenko, E. (2022). LanA (Language Atlas): A probabilistic atlas for the language network based on fMRI data from >800 individuals. In *bioRxiv*. https://doi.org/10.1101/2022.03.06.483177

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1907.11692

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(8). https://doi.org/10.1073/pnas.2011417118

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). Pointer Sentinel Mixture Models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1609.07843

Pannitto, L., & Herbelot, A. (2020). Recurrent babbling: evaluating the acquisition of grammar from limited input data. *Proceedings of the 24th Conference on Computational Natural Language Learning*, 165–176.

Pasquiou, A., Lakretz, Y., Hale, J., Thirion, B., & Pallier, C. (2022). Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2207.03380

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, *9*(1), 963.

Pérez-Mayos, L., Ballesteros, M., & Wanner, L. (2021). How much pretraining data do language models need to learn syntax? In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2109.03160

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., & Yarkoni, T. (2017). Scanning the horizon: towards

transparent and reproducible neuroimaging research. *Nature Reviews. Neuroscience*, *18*(2), 115–126.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*, 8.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(45). https://doi.org/10.1073/pnas.2105646118

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, *7*, 12141.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse Deep Neural Networks All Predict Human Inferior Temporal Cortex Well, After Training and Fitting. *Journal of Cognitive Neuroscience*, *33*(10), 2044–2064.

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 14954–14964). Curran Associates, Inc.

van Schijndel, M., Mueller, A., & Linzen, T. (2019). Quantity doesn't buy quality syntax with neural language models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1909.00111

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1706.03762

Warstadt, A., & Bowman, S. R. (2019). Linguistic analysis of pretrained sentence encoders with acceptability judgments. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1901.03438

Warstadt, A., & Bowman, S. R. (2022). What Artificial Neural Networks Can Tell Us About Human Language Acquisition. In *arXiv [cs.CL]*. arXiv. https://doi.org/10.1073/pnas.2021865119

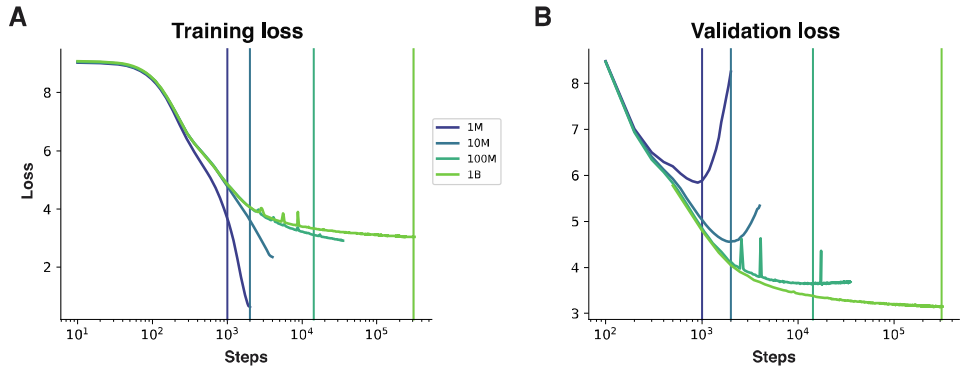
Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, *10*(1), 3770.

Zhang, Y., Liu, H., Li, H.-S., Warstadt, A., & Bowman Samuel, R. (2020, July 2). *The MiniBERTas: Testing what RoBERTa learns with varying amounts of pretraining*. https://wp.nyu.edu/cilvr/2020/07/02/the-minibertas-testing-what-roberta-learns-with-varying-amounts-of-pretraining/

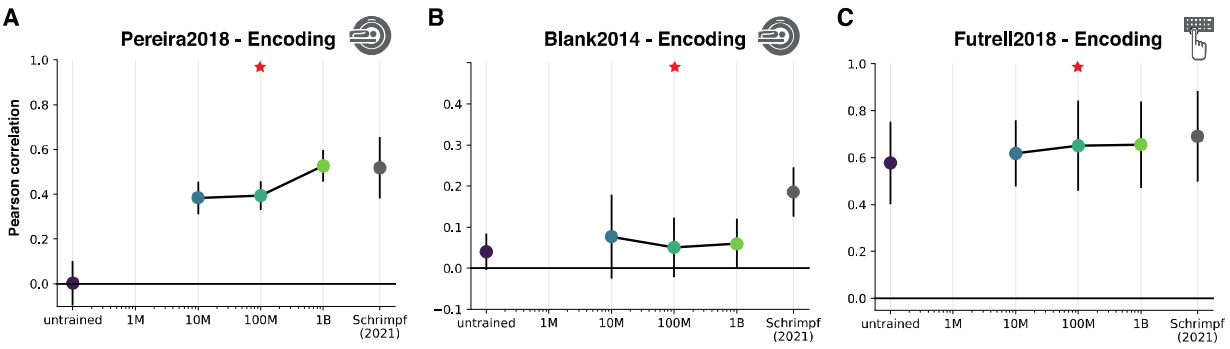
Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1506.06724

Zhuang, C., Xiang, V., Bai, Y., Jia, X., Turk-Browne, N., Norman, K., DiCarlo, J. J., & Yamins, D. L. K. (2022, September 23). How Well Do Unsupervised Learning Algorithms Model Human Real-time and Life-long Learning? *36th Conference on Neural Information Processing Systems*. https://openreview.net/pdf?id=c0l2YolqD2T

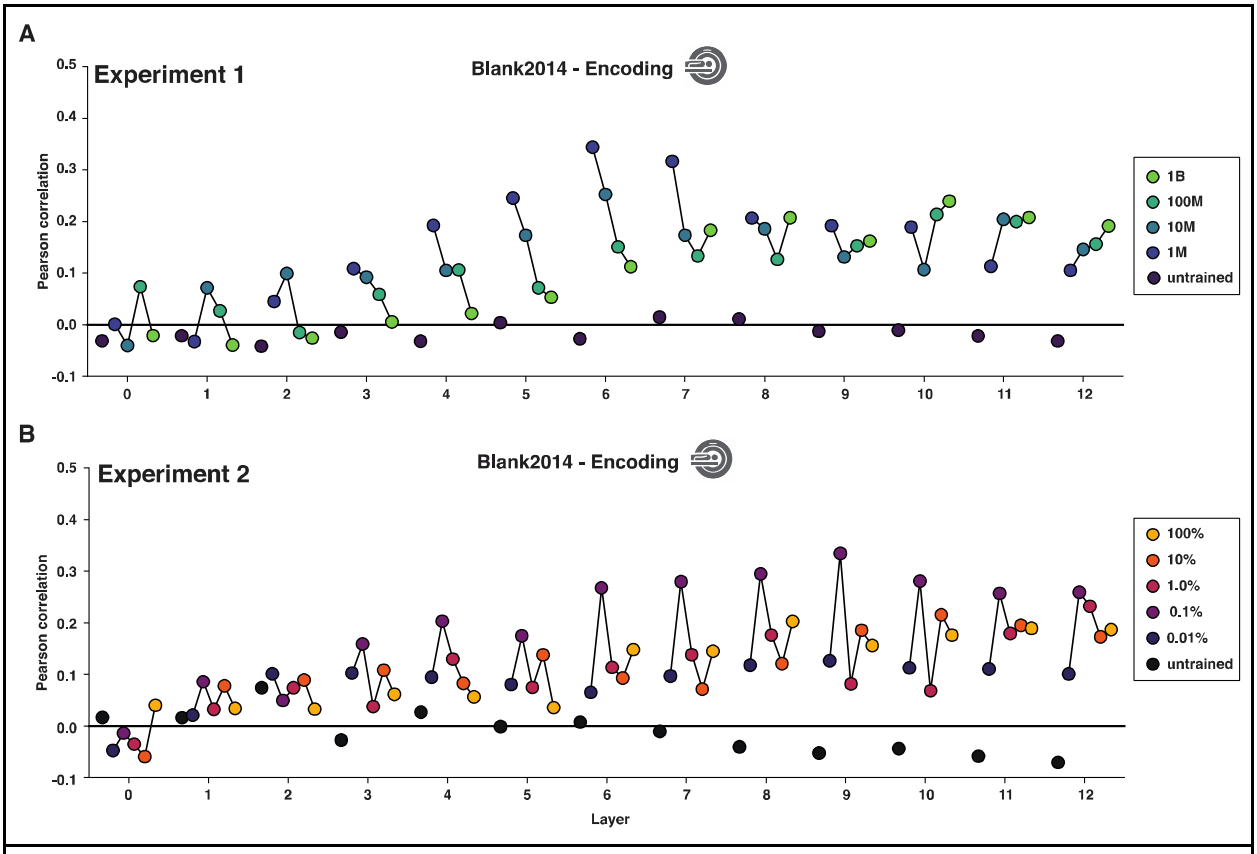
Supplementary Figures



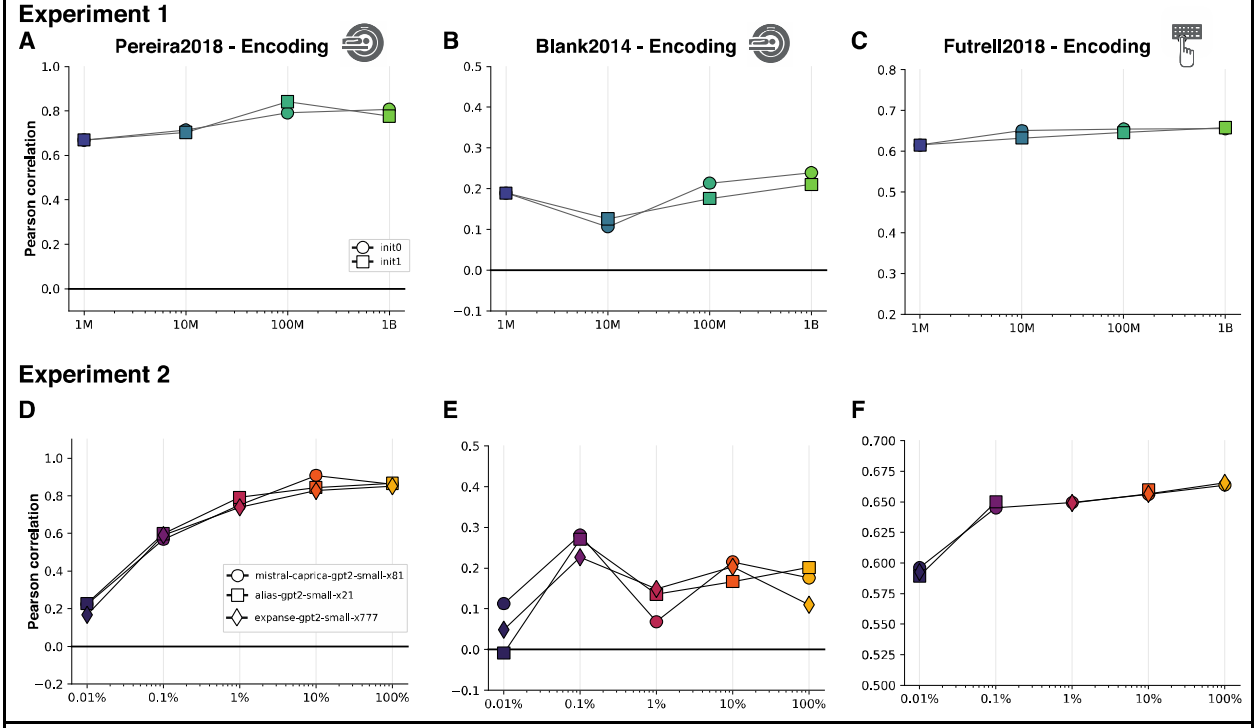
Supplementary Figure 1. Training and validation loss for the models in Experiment 1. For all models, a training batch size of 64 was used (each element in a batch is a sequence of 1,024 tokens). Training was done with a standard Adam optimizer (LR: 0.0006, beta: [0.9,0.999], eps:1e-8), and cosine learning rate (starting at 0.0006 and decaying over 320000 training steps), with gradient accumulation and clipping, attention and hidden dropout ratio of 0.1, and no weight decay. **A.** Training loss: as expected, all models show a drop in loss with additional training, and for models trained on smaller datasets, the rate of learning is faster. **B.** Validation loss: as expected, all models show a drop initially, and models trained on smaller datasets reach their minimum loss earlier in the training process and have higher values (corresponding to worse performance) than models trained on larger datasets. In both A and B, the vertical lines mark the step with the minimum validation loss (best model performance). The representations for the steps shown in B were used for testing against the human benchmarks.



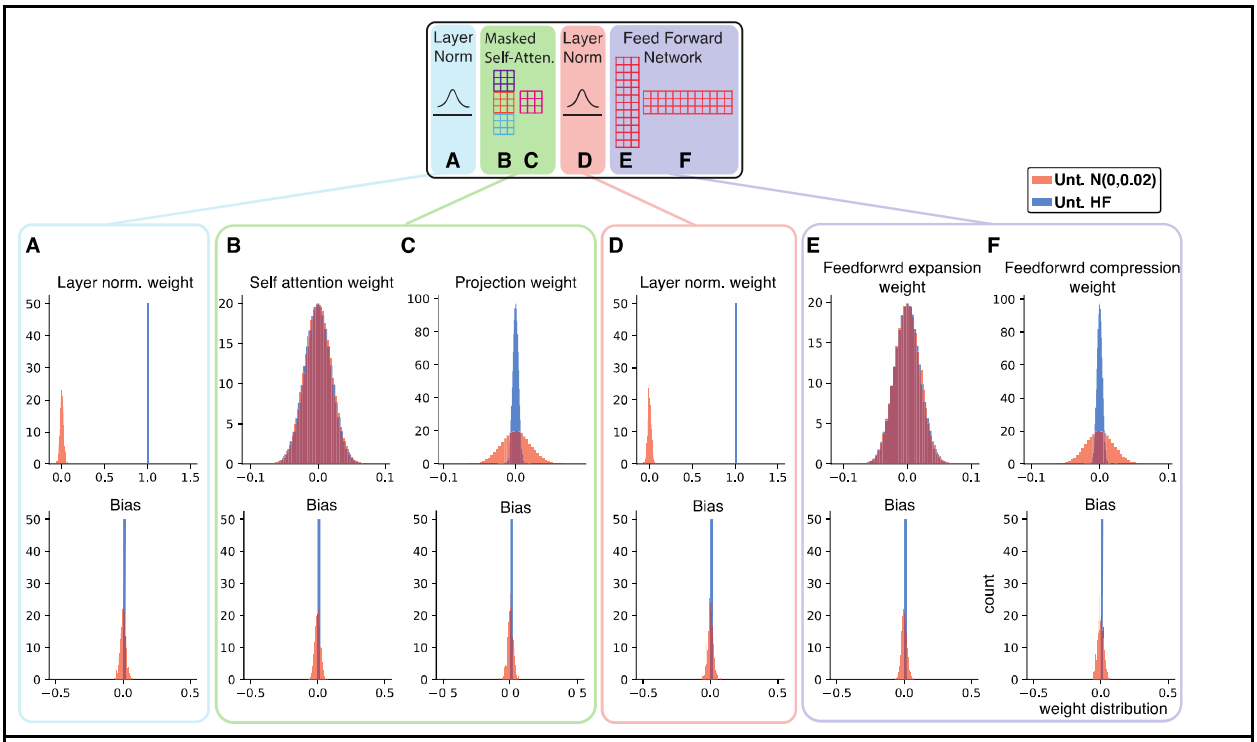
Supplementary Figure 2. Model performance on the human benchmarks for the miniBERTa models as a function of training in Experiment 1. Performance of the best-performing layer, as reported in Schrimpf et al. (2021), in predicting language-responsive voxels' activation in the Pereira2018 fMRI benchmark (A), language regions' responses in the Blank2014 fMRI benchmark (B), and reading times in the Futrell2018 behavioral benchmark (C). The results are shown for i) an untrained model (black dots); ii) three models trained on datasets of different sizes (10M, 100M, and 1B tokens) (teal-to-green dots connected by a line; the model trained on a developmentally plausible amount of data—100M—is marked with a red asterisk); and iii) a fully trained model, as reported in Schrimpf et al. (2021) (grey dots).



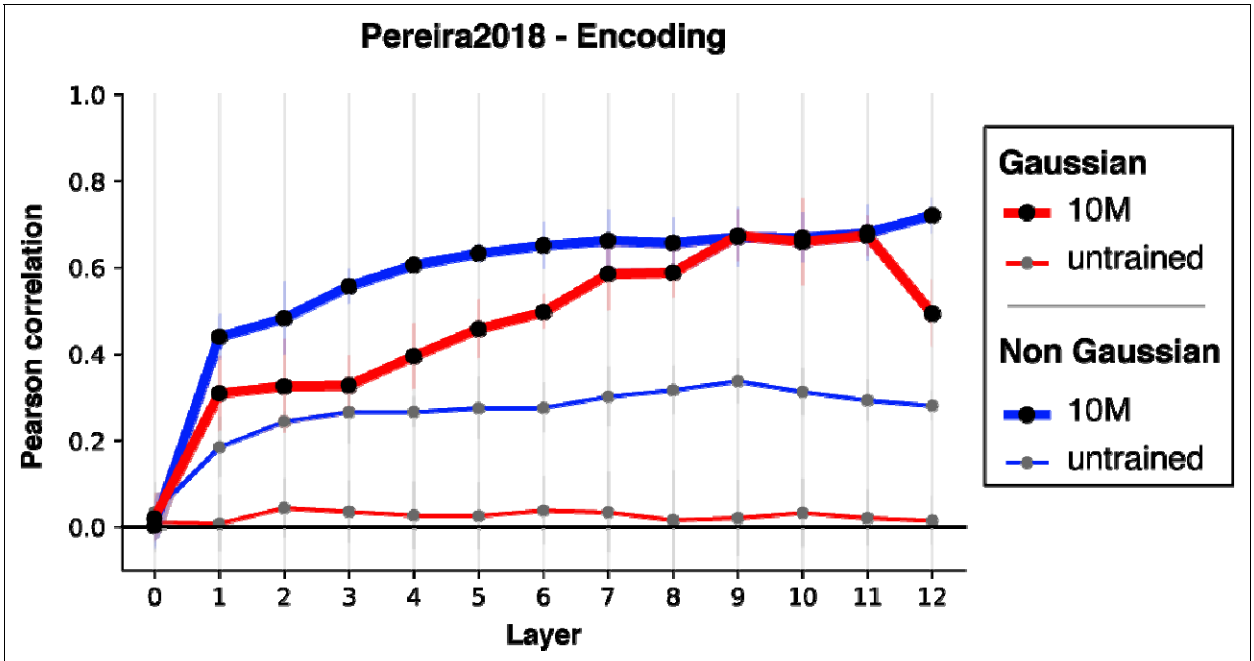
Supplementary Figure 3. Exploratory analyses of individual model layers for the Blank2014 fMRI benchmark. Performance of the 12 GPT-2 model layers in predicting human neural responses in Experiment 1 (A) and Experiment 2 (B). The results are shown for i) an untrained model (black dots); and ii) four models trained on datasets of different sizes (blue-to-green dots connected by a line in A) or a model trained on a large dataset at different points during the training (purple-to-yellow dots connected by a line in B). (Layer 0 is the token embedding layer, and layer 12 is the last transformer layer.)



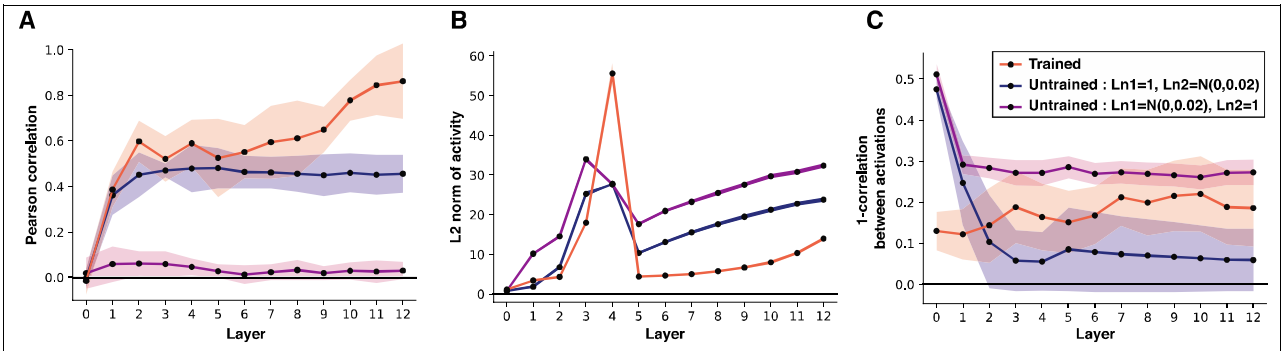
Supplementary Figure 4. Robustness of the critical results to seed choice during initialization. The effects of different seeds used for model initialization (e.g., Mehrer et al., 2020) on model performance on the human benchmarks in Experiment 1 (fMRI benchmarks: A-B; behavioral benchmark: C) and Experiment 2 (fMRI benchmarks: D-E; behavioral benchmark: F). For Experiment 1, two different seeds were used; for Experiment 3, three different seeds were used.



Supplementary Figure 5. A detailed comparison of the two model initializations for the untrained models in Experiment 2 (for a sample layer (layer 1)). The distribution of weight and bias values for the first layer normalization (A), two operations in self attention (B-C), the second layer normalization (D), two operations in feedforward processing (E-F). Unt. N(0,0.02) corresponds to the untrained model initialized with a mean of 0 and a standard deviation of 0.02; and Unt. HF corresponds to the untrained model initialized with the Hugging Face parameters. (The model initializations used for the untrained models in Experiment 1 are similar.)



Supplementary Figure 6. Performance of an untrained GPT-2 model and a GPT-2 model trained on the 10M tokens dataset, each initialized in two different ways, across layers. Performance of four GPT-2 models (an untrained model and a model trained on a 10M tokens dataset, each initialized with two different weight distributions; see Methods and **Suppl. Figure 5**) in predicting language-responsive voxels' activation in the Pereira2018 fMRI benchmark. The untrained model initialized with a gaussian distribution (mean: 0; standard deviation: 0.02) performs close to 0 across layers. In contrast, the untrained model initialized with the Hugging Face (non-gaussian) parameters already achieves ~0.4 predictivity for some layers. After training, both models reach a similar level of alignment with the human data for their later layers.



Supplementary Figure 7. Effect of differences in model initialization for untrained models on model performance on the Pereira2018 benchmark. We created two variants of untrained models. Both used a gaussian weight distribution for the self-attention and feedforward components of the model, but critically one model (blue curve), the first layer normalization (Ln1) was set to 1 (as in the Hugging Face initialization), and the second layer. normalization (Ln2) was set to a gaussian distribution, and the other model (maroon curve), Ln1 was set to a gaussian distribution, and Ln2 – to 1 (as in the Hugging Face initialization). The third curve (orange) corresponds to a trained model for reference.

A. Model performance on the Pereira2018 fMRI benchmark. The model with Ln1=1 exhibits a higher performance compared to the model with Ln2=1, suggesting that the first layer normalization plays a bigger role in contributing to above-zero performance for untrained models on the human benchmarks.

B. Amplitude of activity in the three models. This graph demonstrates that the difference in performance between the two untrained models is not simply due to lack of activity propagation across layers; both untrained models and the trained model show similar scale of activation.

C. Similarity of model representations among the linguistic stimuli in the Pereira2018 benchmark (higher values correspond to lower similarity). This graph demonstrates that setting Ln1 to 1 (compared to setting Ln2 to 1 or using a trained model) results in more similar representations for the different linguistic stimuli (effectively, removing stimulus-specific encoding). This pattern may explain the above-zero predictivity of neural responses for the untrained model initialized with the Hugging Face parameters.