Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps

Alexandre Pasquiou, Yair Lakretz, John Hale, Bertrand Thirion, Christophe Pallier

Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps

Alexandre Pasquiou [*1,2] **Yair Lakretz** [*1] **John Hale** [3] **Bertrand Thirion** [2] **Christophe Pallier** [1]

Abstract

Neural Language Models (NLMs) have made tremendous advances during the last years, achieving impressive performance on various linguistic tasks. Capitalizing on this, studies in neuroscience have started to use NLMs to study neural activity in the human brain during language processing. However, many questions remain unanswered regarding which factors determine the ability of a neural language model to capture brain activity (aka its 'brain score'). Here, we make first steps in this direction and examine the impact of test loss, training corpus and model architecture (comparing GloVe, LSTM, GPT-2 and BERT), on the prediction of functional Magnetic Resonance Imaging timecourses of participants listening to an audiobook. We find that (1) untrained versions of each model already explain significant amount of signal in the brain by capturing similarity in brain responses across identical words, with the untrained LSTM outperforming the transformer-based models, being less impacted by the effect of context; (2) that training NLP models improves brain scores in the same brain regions irrespective of the model's architecture; (3) that Perplexity (test loss) is not a good predictor of brain score; (4) that training data have a strong influence on the outcome and, notably, that off-the-shelf models may lack statistical power to detect brain activations. Overall, we outline the impact of model-training choices, and suggest good practices for future studies aiming at explaining the human language system using neural language models.

[*]Equal contribution [1]Cognitive Neuroimaging Unit, INSERM, CEA, Neurospin, Gif-sur-Yvette, France [2]Parietal, INRIA, CEA, Neurospin, Gif-sur-Yvette, France [3]Dept. of Linguistics, U. of Georgia, Athens, GA, USA. Correspondence to: Alexandre Pasquiou <alexandre.pasquiou@inria.fr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

1. Introduction

In the last few years, Transformer-based language models have revolutionized the field of natural language processing in virtually all areas. Although these models were developed for applications in language technology, their impressive success has raised interest in whether these models could also shed light on language processing in the human brain. Promising results in this direction suggest that brain activations and transformer-based models converge to similar linguistic representations (Caucheteux et al., 2021b) showing that brain activity can be significantly predicted from linear combinations of model activations, as was shown for fMRI (Toneva et al., 2020; Caucheteux et al., 2021b;a), MEG (Caucheteux & King, 2021), and intracranial data (Goldstein et al., 2021).

However, several differences between Transformer-based models and the human brain raise questions about how far we can advance our understanding of brain function using these models. First, the architecture of Transformers is based on multi-head self-attention modules, which does not clearly map on neural computations in biological networks (e.g., Dayan & Abbott, 2005). Does this architecture contribute to or hinder the ability of the model to predict brain activity compared to other, possibly more brain-like, architectures (e.g., recurrent neural networks)? Second, the data used to train Transformer-based models is often different from that available for children, both in type and size. Training a Transformer-based model requires massive corpora, on the order of billions of words, whereas children require orders of magnitudes less words to achieve comparable or better linguistic performance. How does the training corpus (type and size) affect the model's ability to fit brain activity? Finally, the learning and evaluation objective commonly used with these models, such as masked or next-word prediction, is at most a rough approximation of the computational problem the human brain solves during language acquisition and processing. Can one consider that a well-trained model (according to perplexity loss) is a good model for brain activity in language tasks?

We investigate these questions by contrasting several types of language models in their ability to fit functional Magnetic Resonance Imaging (fMRI) timecourses of participants lis-

tening to the 'The Little Prince' audiobook. Importantly, we conduct the model comparison while controlling for various aspects of the architecture of the models and the type and size of the corpus on which they are trained. To address the first question about the architecture of the models, we study the ability of untrained models to fit brain activity. We obtain significant differences across architectures, with that of recurrent neural networks achieving highest scores. Next, we study brain-score gains brought by training across models, and find a network of brain regions, in which brain activity is consistently better fitted by various types of models. Moreover, running a comprehensive comparison of neural language models, we find that the effect of training is stronger in the case of Transformer-based models. We next question the relationship between perplexity and brain score, and study it across models and across training epochs during convergence. In contrast to previous studies, we find that perplexity is not a reliable predictor of model's brain score. Finally, we show the impact of training data on the model ability to fit brain data, notably, that off-the-shelf models, such as ones trained only on Wikipedia, may lack statistical power to fit brain activation.

Taken together, we conclude that while the starting point of Transformer-based models is less advantageous compared to that of recurrent neural networks, due to differences in their architectures, training leads to them outperforming all other models in predicting brain data.

2. Related Literature

Current knowledge about the cerebral basis of language mostly comes from brain imaging studies that have used tightly controlled stimuli, typically isolated words or sentences out of context (see Price, 2012; Hickok & Small, 2015, for reviews). As conclusions from such studies may be bounded to the peculiarity of the task and setup used in the experiment (Varoquaux & Poldrack, 2019), researchers have become more and more interested in data using "Ecological Paradigms", in which participants are engaged in more natural tasks, such as conversation or story listening (e.g. Regev et al., 2013; Lerner et al., 2011; Wehbe et al., 2014).

Ecological paradigms commonly require methodologies of machine learning based on predictive modeling, to account for the high number of degrees of freedom in the complex system that is the brain. It has been shown that representations extracted from computational models can explain part of the signal acquired in brain neuroimaging. This was shown in early studies by using non-contextualized semantic representations (Mitchell et al., 2008; Huth et al., 2016), moving in later studies to recurrent neural networks to extract context-based word representations (Jain & Huth, 2018; Jain et al., 2021), and more recently to Transformer-based language models (e.g., Toneva et al., 2020; Caucheteux et al., 2021b;a; Goldstein et al., 2021) - see Hale et al. (2022) for a review.

Interestingly, the architecture of neural language models has been shown to substantially contribute to the ability of the model to fit brain data. Untrained neural language models fitted human brain activity surprisingly well (Schrimpf et al., 2021). Training was shown to improve brain scores by around 50% on average, across different architectures. This was suggested as evidence that the human cortex might already provide a sufficiently rich structure for relatively rapid language acquisition. However, conclusions in Schrimpf et al. (2021) were based on relatively small datasets, from no more than 9 participants. Also, different models in the comparison had different number of units, layers, and were trained on different datasets with varying vocabulary sizes. In our work, we suggest a more controlled study of the effect of architecture and training on brain score, comparing different types of models, while controlling for the aforementioned factors, using a larger brain-imaging dataset, from 51 participants.

The performance of neural language models on a next-word prediction task, but not on other linguistic tasks, was shown to correlate with their ability to fit human brain data (Schrimpf et al., 2021). This was suggested as evidence that predictive processing shapes language-comprehension mechanisms in the brain. Here, we question this conclusion and study the relation between next-word prediction and brain score in various types of models, training corpora and training steps.

3. Analysis Setting: Fitting Brain Data with Modern NLP Models

Investigating the ability of neural language model to capture brain activity, we (1) first defined the three families of model architectures that we explored: non-contextualized word embeddings (GloVe; Pennington et al. 2014), a recurrent neural network (LSTMs; Hochreiter & Schmidhuber 1997) and two Transformer-based models (GPT-2; Radford et al. 2019 and BERT; Devlin et al. 2019); (2) we then trained and tested each model as described in the following paragraphs; before (3) presenting the story *The Little Prince* (Eckert-Boulet, 2011) to both human participants and artificial neural language models. Their activation in response to each word and punctuation sign of *The Little Prince* was extracted and (4) used to fit participants fMRI brain activations thanks to regularized linear encoding models. (5) Finally, at the end of the analysis pipeline, we had for each model: a test loss evaluated on our test set and a volumic R maps containing, for each brain voxel, the cross-validated correlation between the encoding model prediction and the observed fMRI response.

3.1. Datasets

Brain-imaging data. The brain data consisted of functional Magnetic Resonance Imaging (fMRI) scans from 51 participants who listened to the entire audiobook of *The Little Prince*[1]. For each participant, there were 9 runs of fMRI acquisition lasting about 10 minutes. Whole brain volumes were acquired every 2 seconds. A global brain mask was computed to only keep voxels containing useful signal across all runs for at least 50% of all participants (26,164 voxels). Finally, linear detrending and standardization (mean removal and scaling to unit variance) were applied to each voxel's time-series. The analysis pipeline relies on Nilearn (v.0.8.1) for data access and visualization. Encoding and subsequent statistical analyses were run with custom Python code using sklearn.

The acoustic onsets and offsets of the 15,435 spoken words were marked to align the audio recording with the text of *The Little Prince*. In addition to the words, the token streams fed to the neural language models included punctuation signs (commas, dots, ...).

Text Corpora. We designed several datasets on which we trained and evaluated our models. In total, we created 6 training datasets from Wikipedia (425M) and Project Gutenberg [2], using up to 2 thirds of the entire Project Gutenberg in the *xlarge* version and splitting the remaining 1/3 left into equal size validation (1.1G) and test sets (1.1G). The datasets created from the Gutenberg Project's data are nested ($small(240M) \subset medium(737M) \subset large(2.2G) \subset xlarge(4.4G) \subset full(4.8G; xlarge + Wikipedia)$).

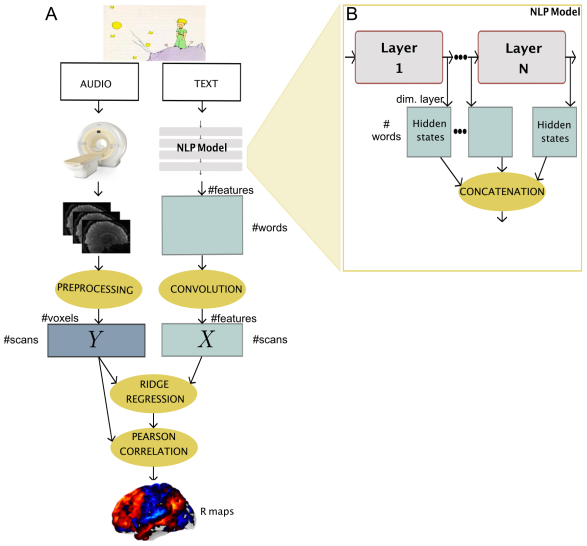

Figure 1. **Overview of the pipeline:** (A) Human participants were presented with an auditory version of *The Little Prince* story while their brain activity was recorded with fMRI. Neural models were presented with a text transcription and the entire state of the network was recorded for each word and punctuation sign. Pre-processing steps were applied to both brain and model activations before aligning the two signals using a nested cross-validated Ridge-regression model. Finally, brain maps of correlation coefficients between models' predictions and fMRI time-series were computed. (B) In the case of models with several layers, model activations were extracted from each layer and were concatenated into a single activation matrix.

3.2. Pipeline

Models. Given common training, validation and test datasets, we trained several instances of GloVe, LSTM, GPT-2 and BERT.

- GloVe was trained using the open-source code made available by Pennington and al. [3],
- GPT-2 and LSTM were trained on a Language Modeling task,
- while BERT was trained on a Masked-Language Modeling task with a 15% masking-rate.

Each model had a vocabulary of 50,001 tokens. GloVe and LSTM were trained for 23 epochs, while GPT-2 and BERT were trained during 5 epochs. Convergence assessment and comparisons with off-the-shelf models are provided in Supplementary Material. For computational cost reasons, we limited our analysis to 1, 2 and 4-layers models. In the following, we denoted MODEL.X a MODEL with X-layers.

Activation generation. (See Fig.3.2) We presented the transcription of the audio book used to acquired fMRI brain data (The Little Prince, TLP; 15.435 words) to both trained and untrained versions of the selected artificial neural language models.

For each model, we extracted the model hidden-states while it processed its input, and defined from it what we call the activation matrix (one for each run). More precisely, if we note d the dimensionality of a neural model, which corresponds to the total number of units in the model, and w the total number of words in the text. We obtained an activation matrix $A \in \mathbb{R}^{w \times d}$ after the presentation of the entire text to the model. This means that each word of *TLP* is represented by a d-dimensional vector. Then, model activations were transformed into time-series matched to the fMRI acquisition times, using the following procedure:

[1]Available from https://openneuro.org/datasets/ds003643/versions/1.0.2

[2]Project Gutenberg. (n.d.). Retrieved February 21, 2016, from www.gutenberg.org.

[3]https://nlp.stanford.edu/projects/glove/.

1. **Normalization:** To match the scale of activations across layers (for multi-layers models), the activations of each layer were normalized by dividing them by the average L2-norm over words.
2. **Convolution:** each column of the resulting activation matrix, mapped onto words' offsets times, was convolved with SPM's canonical Haemodynamic Response Function(HRF; Friston 2007) sampled at 0.2s. The outcome was resampled at 2s to match the repetition time of fMRI acquisition and then mean-centered.

Fitting brain data. The latter stage resulted for each model and each run into a design-matrix of size $\#\text{scans} \times d$. Given a neural language model, we gave the associated nine design-matrices to a nested cross-validated L2-regularized univariate linear encoding model to fit the fMRI brain data (of size $\#\text{scans} \times \#voxels$).

The *encoding model* is a function that maps a vector of stimulus features onto brain responses activity (Naselaris et al., 2011). We denote by x_t the vector of features at time t, such as predicted time-courses derived from a language model, and by y_t^v the corresponding brain responses measured at voxel v. We learnt a linear voxel-level encoding model using Ridge regression, whose general solution is given by:

$$\hat{\beta}^v_{\text{Ridge}} = arg\min_{\beta} \sum_{t=1}^{n} (y_t^v - \beta^T x_t)^2 + \lambda \|\beta\|_2^2$$

To evaluate model performance and the optimal regularization parameter λ^*, we used a nested cross-validation procedure: we split each participant's dataset into training, validation and test sets, such that the training set included 7 out of the 9 experiment runs, and the validation and test sets contained one of the two remaining sessions. We evaluated model performance using the Pearson correlation coefficient R, which is a measure of the linear correlation between models' predicted time-courses and the actual time-courses. It is defined as:

$$R(y, \hat{y})_{v,test} = \frac{\sum(\hat{y}_t^v - \bar{\hat{y}}^v)(y_t^v - \bar{y}^v)}{\sqrt{\sum(\hat{y}_t^v - \bar{\hat{y}}^v)^2 \sum(y_t^v - \bar{y}^v)^2}},$$

$$\text{where } \bar{\hat{y}}^v = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t^v \quad , \quad \bar{y}^v = \frac{1}{T}\sum_{t=1}^{T} y_t^v$$

For each subject and each voxel, we first determined λ^* by comparing R_{valid} for 10 different values of λ, linearly spaced in log-scale between 10 and 10^5. We then calculated R_{test} for λ^*. Finally, we repeated this procedure 9 times, using cross-validation. This resulted in 9 R_{test} values that we then averaged to produce a single R_{test} map for the participant.

Results of the analysis pipeline. Finally, at the end of the analysis pipeline we had for each model: a test loss evaluated on our test set derived from the Gutenberg Project, and a volume-based R map displaying for each brain voxel the correlation between the encoding model prediction and the observed time series. Volume maps are rendered on cortical surfaces by projection.

4. Methods and Experimental Setup

4.1. Assessing model fitness to brain data

Whole-brain, voxel-based, group analyses were performed, using one-sample t-tests applied to the individuals' R_{test} maps spatially smoothed with an isotropic Gaussian kernel with 6mm FWHM. In each voxel, the test assessed whether the distribution of R_{test} values across participants was significantly larger than zero. To control for multiple comparisons, all the maps displayed in this document were corrected using a Bonferroni correction of 0.1 (Bonferroni, 1936), that is, values reported on the maps (e.g. R scores) are shown only for voxels that survived this threshold.

We derived in a model-agnostic manner from a Shared-Response Model (SRM, Chen et al. 2015) the most "responsive" voxels, that is, the voxels whose R values were the 25% highest ones. This set of 6541 voxels, which we will refer to as "SRM25" is displayed on a brain surface at the bottom of Fig.2. It is used to compute the distributions of brain scores.

4.2. Comparison of untrained models and baselines

In our first analysis, we assessed whether the model class and number of layers bias its ability to fit fMRI brain data. We instantiated several untrained versions of each model class, varying the number of layers, and generated activation matrices from these models before fitting them to brain data. For each model, the activation matrices were built using all the hidden-states of all layers, including the embedding layer. We also defined a Baseline model whose activation matrices are obtained by associating a fixed embedding vector of size 768 (size of each model's layer) to each word of the text. It is equivalent to an untrained GloVe model (and will be referred to as such). For each, we obtained 3D brain maps displaying the average R_{test} values in each voxel. Finally we displayed boxplots of the R_{test} values distributions in the previously SRM-defined voxel selection.

4.3. Comparison of trained and untrained neural language models

We then investigated how training models impacts their ability to fit fMRI brain data. We first generated activation matrices from the trained language models before fitting them to brain data and finally displaying the group-level difference between each model's map and its untrained ver-

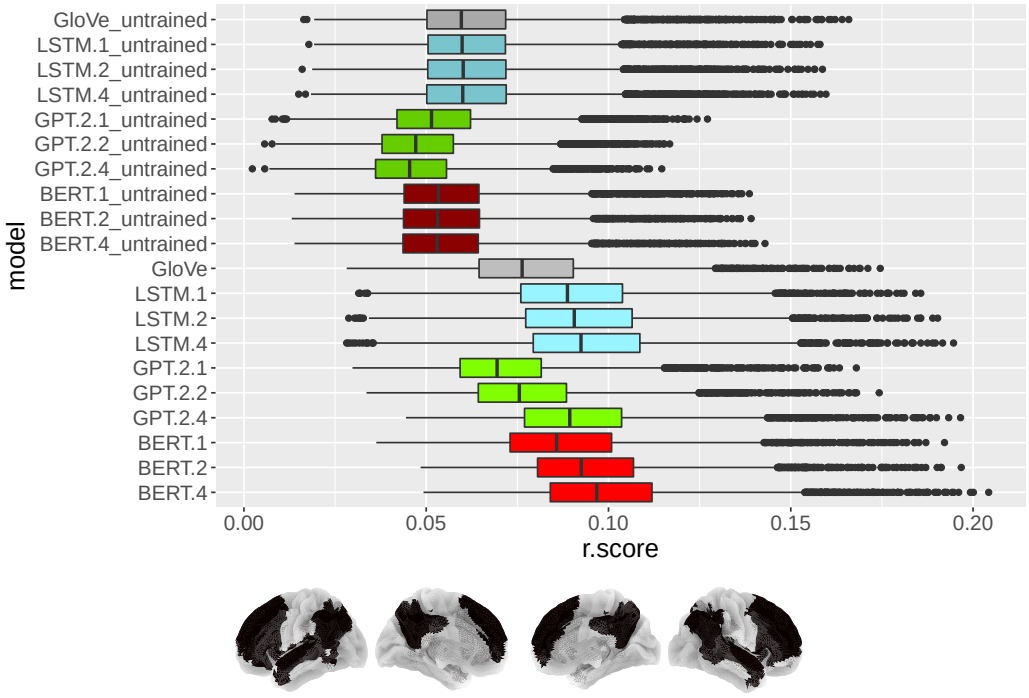

Figure 2. **Distributions of R_{test}-values across voxels in the 25% most reliable voxels across subjects (mask computed from shared response model, shown below the graphic)** for untrained and trained versions of GloVe (static word embedding), LSTM, GPT-2 and BERT models, having 1, 2 or 4 layers.

sion's map. To study the overlap between the different contrast maps, we extracted for LSTM, GPT-2 and BERT the set of voxels whose R values were the 10% highest. Then we computed the intersection of these three sets and the percentage of overlap for pairs of models and for the three models. We additionally identified the set of voxels whose R values were the 10% highest in the untrained models maps, and computed the intersection of the three sets and the percentage of overlap for the three models. Finally, we studied the intersection between the 2 overlaps of the three models and synthesized differences and similarities in Fig.4.

4.4. Mapping the brain scores of NLP models

The next step was to run a comprehensive comparison of the selected models to understand models similarities and specificities. We contrasted individual maps between pairs of models to test in each voxel: (1) the effect of incorporating context into target-word representations, by contrasting LSTM and GloVe; (2) the effect of attention vs. recurrence mechanisms on prediction, by contrasting the maps of transformers and LSTM. (3) the interaction between model architecture and training between transformers and LSTM. (4) and finally the effect of bi-directionality vs. incrementality (BERT vs. GPT-2). Compared models always have the same number of hidden-states.

4.5. Perplexity and Brain score

Finally, we investigated the relation between *perplexity* and *brain score*. Using the set of trained LSTM, GPT-2 and BERT models, we evaluated them using the standard loss, that is, the average logarithm of model perplexity, on the *test set*. For each model, we also computed the *brain score*, defined as the average R-value within the SRM25 voxelset.

We also investigated the importance of the training data on the model ability to fit brain data, comparing models trained on Wikipedia and on the Full dataset (Wikipedia + Gutenberg xlarge). These analyses were performed on LSTM and GloVe.

5. Results

Figure 2 shows the distributions of R_{test}-values across voxels for trained and untrained models of various architectures and number of layers.

5.1. Performance of untrained models

Remarkably, all untrained models, regardless of their architecture, explain signal better than chance. Untrained

LSTM and untrained GloVe (that is, Random Embeddings) perform equally well with an average score around 6.3% (SE=0.02%), and significantly better than Transformers as attested by direct comparisons between 4 layers models: LSTM.4−GPT-2.4 (1.6% SE=0.02%); LSTM.4−BERT.4 (0.7% SE=0.004%). Overall, untrained GPT-2.4 had the worst performance (BERT.4−GPT-2.4 (0.9% SE=0.01%)).

The brain regions where $\text{LSTM}_{\text{untrained}}$ performs significantly better than GPT-2 are displayed on Fig.3. They are located within the left hemispheric language network and its right counterpart (Superior Temporal Gyrus/Superior Temporal Sulcus and Inferior Frontal Gyrus (pars opercularis)).

5.2. Effect of number of layers

Next, we looked at the effect of number of layers for LSTM, GPT-2 and BERT models.

As can be seen on Fig. 2 the change in performance for untrained models is either flat (for LSTM and BERT) or negative (for GPT-2). Comparing 4-layer models to 1-layer models yields the following: LSTM (-0.02% SE=0.002%); GPT-2 (-0.6% SE=0.004%), BERT (-0.02% SE=0.003%).

For trained models, performance improves with the number of layers. The increase in performance (4-layer model's performance - 1-layer model's performance) is more marked for Transformers — GPT-2 (2% SE=0.006%) and BERT (1% SE=0.006%) — than for LSTM (0.4% SE=0.005%).

5.3. Effect of Training

Visual inspection of Fig.2 shows, unsurprisingly, that training helps: trained model fit brain data better than models initialized with random weights. To quantify this improvement for each model type, we computed, in each voxel, the difference in R_{test} between the trained model and the untrained model. All differences were statistically significant: GloVe (1.5% SE=0.02%); LSTM (3.1% SE=0.02%) ; GPT-2 (4.5% SE=0.02%); BERT (4.4% SE=0.02%); in Student T-tests, all $ps < 10^{-16}$). Fig.4A shows the distributions of these training effects.

The maps on Fig.4B show the locations of voxels where the R_{test} increases are significant. The effect of training is spatially consistent across models, that is, displays similar topographies across models; and the R-score improvements are comparable in high-order language networks across models.

To assess the similarity between the hotspots on these maps, we thresholded them, keeping the 10% of voxels (2617 voxels) showing the highest gains with training. We then computed the percentage of overlap across the resulting binarized maps. Results are presented in Table 1. The overlap between the three maps is 75% across all 3 models. We

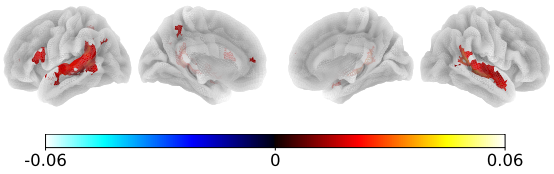

Figure 3. **LSTM vs. GPT-2 architecture:** Brain regions in which an untrained LSTM outperforms an untrained GPT-2 model. The comparison is for 2-layer models.

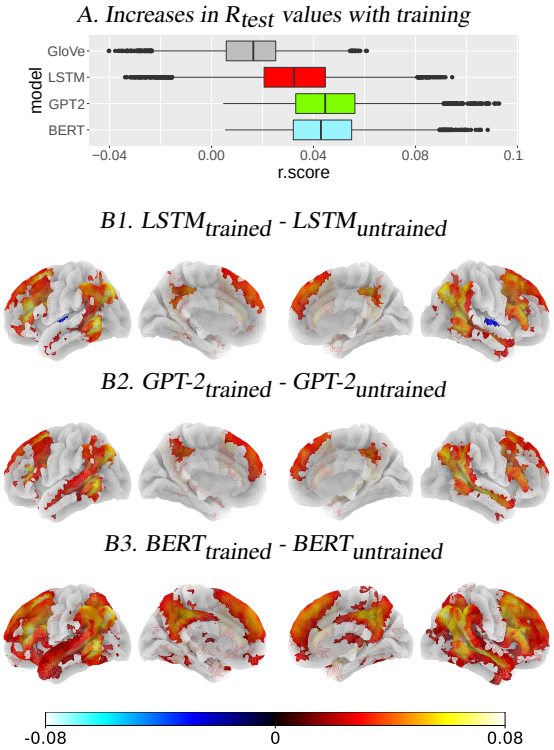

Figure 4. **Effect of model training.** A) Distributions of R_{test} increase for the 2-layer versions of the three types of models. B) Brain areas showing significant increases: LSTM (top), GPT-2 (middle) and BERT (bottom). C) Regions showing the strongest gains in R scores with training across the three models (intersection of the three previous maps thresholded at the 10% upper percentile), in green. Regions showing the strongest R scores across the three 2-layer untrained models: LSTM, GPT-2, BERT (intersection of the three maps thresholded at the 10% upper percentile), in red. There is a 18% overlap between these two highlighted networks (in blue).

then thresholded the untrained models brain maps, keeping again the 10% (2617 voxels) of voxels showing the highest r score, and computed the percentage of overlap across the resulting binarized maps. Results are presented in Appendix Table 1. The overlap between the three maps is 79% across all 3 models. The differences and similarities of these two overlaps are displayed in Fig.4C.

5.4. Comparisons between models

Then, we ran a comprehensive comparison of the models described in Section 3. Firstly, we contrasted a model that takes context into account (LSTM) to a model that does not (Glove). Importantly, to conduct a fair comparison, both models were trained on exactly the same corpus, and had vocabulary of same size (see Section 3). The contrast map displayed in Fig. 5A highlights a set of regions located in the temporo-parietal junction, similar to the trained models overlap of Fig.4 (in green), with bilateral effects in the medial and lateral Superior Frontal Gyri and Posterior Cyngulate gyri, as well as left-lateralized effects in the Temporal Pole, the Inferior Frontal and Middle Temporal gyri.

Note that this map is quite similar to the green network showing an effect of training across models (Fig.4C). This suggests the role of these regions in context processing, given that in both cases, the model benefits from using context for predicting activity in these regions.

Next, we compared a model using attention (BERT) to a recurrence-based model (LSTM). Fig. 5B1) shows that BERT outperforms LSTM mostly in the Superior Temporal Gyri, and in the auditory cortex.

Comparing the bidirectional BERT with the incremental GPT-2 in Fig.5 B2) shows that BERT outperforms GPT-2 in the entire language network. Finally, to test which model architecture best benefit from training, we investigated the interaction between model architecture (BERT vs. LSTM) and training (trained vs. untrained). The interaction map is shown on Fig. 5C. The effects are more spread than the direct difference between the trained BERT and LSTM, as expected from the fact that LSTM untrained has higher performance than BERT untrained. Transformer-based models gain more with training relative to LSTM (this is also the case for GPT2, see Appendix Fig.S6C), and explorations show that this relative gain increases with the number of layers.

In summary, *i)* LSTM model outperforms the non-contextual model GloVe in core regions of the language network; *ii)* Transformer-based models benefit more from training than the LSTM model, and achieve higher brain scores (compared to LSTM) mostly around the auditory cortex.

5.5. Relationship between Perplexity and Brain score

Fig. 6A shows the relationship between perplexities (model loss) and brain scores derived from various models, architectures, training sets and training stages. Unlike previous reports (Schrimpf et al., 2021), we did not observe a clear monotonic relationship between the two variables. For example, the average LSTMs perplexity is worse than that of GPT-2, but the average brain score is higher.

We investigated in more details the effects of model class, number of layers, training epochs and training dataset size on the relationship between brain score and perplexity. The results are presented in Appendix Fig.12. In Fig.S12 panel A, we observed that within each model class, increasing the number of layers improves perplexity and brain score. However, within a given model class, there is not always a monotonic relationship between brain score and perplexity as shown by the effect of training epochs in panels B and C for GPT-2 and panel D for LSTM. Finally, manipulating training dataset size with LSTM shows no simple relation between brain score and perplexity.

5.6. Effect of Training set

Data used for training have a strong influence on the outcome. Fig.7 presents contrasts maps obtained when training LSTM or GloVe with our custom *Full* dataset versus Wikipedia. This shows that off-the-shelf models trained on Wikipedia likely lack statistical power to detect brain activations.

Model	GPT-2	BERT
LSTM	79%	86%
GPT-2	.	85%

Table 1. **Overlap between training effect brain maps** The percentage of common voxels when the maps were thresholded at their 10% upper percentile.

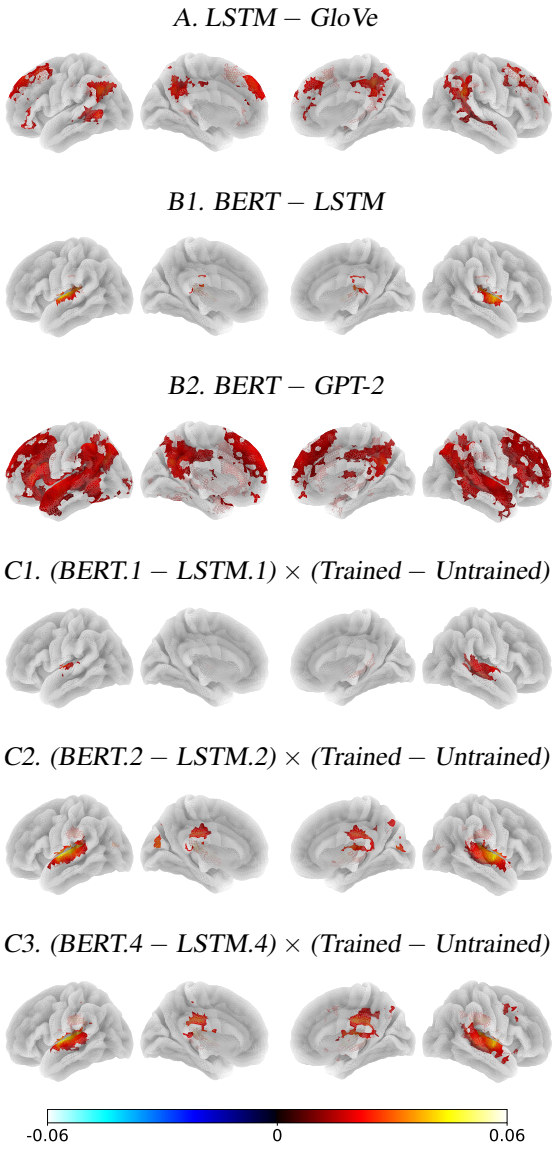

Figure 5. **Localisation of differences between models** A) Comparison between a contextual model (LSTM) and a non-contextual model (GloVe). LSTM outperforms the non-contextual model GloVe in core regions of the language network, with significant effects that are bilateral and included in the core of the previously highlighted network of regions that are better fitted with training 4C. B1) Comparison of a transformer-based model (BERT) with a recurrent neural network (LSTM). Trained BERT better fits fMRI brain data than LSTM model around Heschl's Gyri bilaterally. B2) Comparison of a bidirectional transformer-based model (BERT) with an incremental transformer-based model (GPT-2). Trained BERT better fits fMRI brain data than GPT-2 in the entire language network. C) Interaction between model architecture (BERT vs. LSTM) and Training (trained vs. untrained) for 1-layer (top), 2-layer (middle) and 4-layer (bottom) models. BERT benefits more from training than LSTM. The more layers, the more it learns.

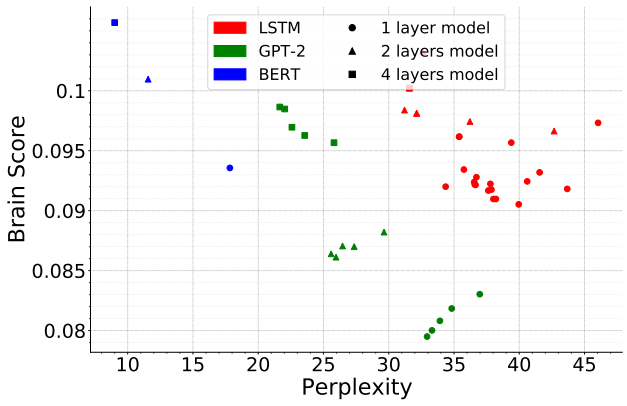

Figure 6. **Perplexity does not predict Brain score.** Among all instances of LSTM, GPT-2 and BERT that are plotted, we found no monotonic relationships, showing that Perplexity cannot serve as a simple proxy to determine Brain Score, as the relation between the two is impacted by the model class, its architecture and training. We represented the brain scores and perplexities of BERT (blue), of several instances of GTP-2 at different training stages (green) and of various instances of LSTM trained on different datasets (red).

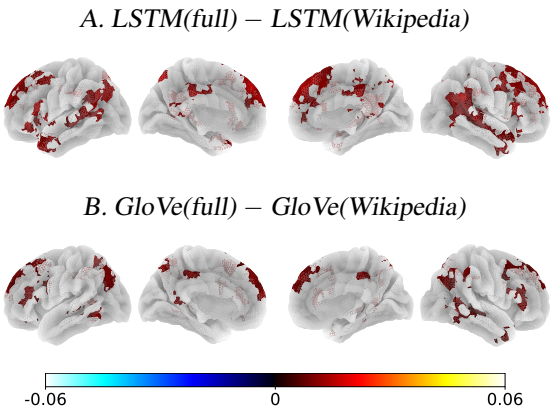

Figure 7. **Influence of Training dataset on R_{test}.** LSTM and GloVe better fit brain data when learning on more training data. This shows the dependence of models contrast on training data. Our full dataset comprised Gutenberg + Wikipedia (4.8 GB) while Wikipedia represented 425 MB.

6. Discussion

Previous work has shown that brain activity during visual or language processing can be significantly predicted from artificial neural-network activations. From a neuro-scientific point of view, the interest lies in the possibility to study language processing in the brain by manipulating the information provided to an artificial model and then to assess the impact on the model's ability to fit brain data. In the present research, we examined the impact of several factors on the fitting performance of artificial models. We studied the impact of model architecture (assessing GloVe and LSTM, GPT-2 and BERT models with varying number of layers), models' perplexity and training corpus, on their capacity to predict functional Magnetic Resonance Imaging timecourses of participants listening to an audiobook. We made several observations: (1) untrained versions of each model already explain significant amount of signal in the brain, with the untrained LSTM and GloVe outperforming the others; (2) training NLP models improves brain scores in the same set of brain regions irrespective of model's architecture; (3) Perplexity is not a good predictor of brain score; (4) training data have a strong influence on the capacity to fit brain data.

One discovery is that all architectures are not equal, but training them consistently increases brain scores in the same set of brain areas. Moreover, these very brain regions are also better predicted by models that take the context into account, such as LSTMs, compared to static models, represented here by GloVe. This provides converging evidence that these regions perform context-dependent computations. From a neuratomical point of view, they are located on the border of the core language regions (IFG and STS) and partly overlap with regions assigned to the Default mode network (Angular Gyri, Dorso mesial prefrontal cortex). These results are coherent with previous work investigating the processing of contextual information by either using LSTM models (Jain & Huth, 2018) or by scrambling the stimuli at different levels (Lerner et al., 2011), confirming that this network (in green/blue in Fig.4) is at the center of combinatorial language processing in the human brain.

We observed that even if transformers start with a disadvantage regarding the ability to fit fMRI brain data, they benefit more from training than LSTMs, and they are able to take advantage of stacks of layers to improve their fitting performance. The comparison of untrained LSTM and untrained GloVe (i.e., random embeddings) showed no significant differences, whereas the comparison of untrained GloVe and untrained transformers showed significant R-score differences in some regions. The difference between untrained LSTMs and untrained Transformers might be due to their different architectures. However, there might be an alternative explanation. Note that for untrained GloVe (random embeddings), each word in the corpus is assigned a *fixed* vector, whereas for untrained Transformers, each word is mapped to a *variable* vector, depending on the context that surrounds the word. Therefore, untrained Glove might better predict brain responses to words that occur frequently both in the training and in the test data (e.g., function words). In contrast, untrained transformers might generate very different embeddings to the same word (e.g., 'the' in the train and test sets), due to their context sensitivity. This variability could therefore reduce the brain score of untrained Transformers compared to that of untrained GloVe. Finally, our results suggest that untrained LSTM are more similar to untrained Glove, having less context sensitivity compared to Transformers. Taken together, this suggests that most of what the untrained baselines capture is similarity in brain responses to words that appear in both the train and test sets.

The discrepancy between brain score and perplexity indicates that training is not a guarantee of convergence towards brain-like representations (see also Hale et al., 2019). Relatedly, other research also seems to militate against perplexity as a royal road to cognitive models (see, e.g., Clark 2000 , chapter 11, 2^{nd} edition).

A final methodological word of caution stem from our results: data used for training have a strong influence on the outcome, showing that off-the-shelf models trained on small datasets like Wikipedia lack statistical power to capture brain activations and should be avoided to probe brain representations.

Acknowledgements

This project/research has received funding from the American National Science Foundation under Grant Number 1607441 (USA), the French National Research Agency (ANR) under grant ANR-14-CERA-0001, the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3), and the KARAIB AI chair (ANR-20-CHIA-0025-01).

References

Bonferroni, C. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

Caucheteux, C. and King, J.-R. Language processing in brains and deep neural networks: computational convergence and its limits. *bioRxiv*, 2021. doi: 10.1101/2020.07.03.186288. URL https://www.biorxiv.org/content/early/2021/01/14/2020.07.03.186288.

Caucheteux, C., Gramfort, A., and King, J.-R. Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects. In *EMNLP 2021 - Conference on Empirical Methods in Natural Language Processing*, Dominican Republic, November 2021a. URL https://hal.archives-ouvertes.fr/hal-03361430.

Caucheteux, C., Gramfort, A., and King, J.-R. Disentangling Syntax and Semantics in the Brain with Deep Networks. In *ICML 2021 - 38th International Conference on Machine Learning*, France, July 2021b. URL https://hal.archives-ouvertes.fr/hal-03361421.

Chen, P.-H. C., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J., and Ramadge, P. J. A reduced-dimension fmri shared response model. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/b3967a0e938dc2a6340e258630febd5a-Paper.pdf.

Clark, A. *Mindware: An introduction to the philosophy of cognitive science.* Oxford University Press, 2000.

Dayan, P. and Abbott, L. F. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press, 2005.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv: 1810.04805.

Eckert-Boulet, N. *The Little Prince/Le Petit Prince*. Omilia Languages, 2011. ISBN 978-0-9567215-9-4.

Friston, K. J. (ed.). *Statistical Parametric Mapping: the analysis of functional brain images*. Elsevier, Amsterdam, 1st ed edition, 2007. ISBN 978-0-12-372560-8.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, S. C., Casto, C., Fanda, L., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Reichart, R., Devore, S., Flinker, A., Hasenfratz, L., Levy, O., Hassidim, A., Brenner, M., Matias, Y., Norman, K. A., Devinsky, O., and Hasson, U. Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines. *bioRxiv*, 2021. doi: 10.1101/2020.12.02.403477. URL https://www.biorxiv.org/content/early/2021/09/30/2020.12.02.403477.

Hale, J., Kuncoro, A., Hall, K., Dyer, C., and Brennan, J. Text genre and training data size in human-like parsing. 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)*, pp. 5846–5852, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1594. URL https://aclanthology.org/D19-1594.

Hale, J. T., Campanelli, L., Li, J., Bhattasali, S., Pallier, C., and Brennan, J. R. Neurocomputational models of language processing. *Annual Review of Linguistics*, 8:427–446, 2022. URL https://doi.org/10.1146/annurev-linguistics-051421-020803.

Hickok, G. and Small, S. L. *Neurobiology of language*. Academic Press, 2015.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, April 2016. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature17637. URL http://www.nature.com/articles/nature17637.

Jain, S. and Huth, A. Incorporating context into language encoding models for fmri. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/f471223d1a1614b58a7dc45c9d01df19-Paper.pdf.

Jain, S., Vo, V. A., Mahto, S., LeBel, A., Turek, J. S., and Huth, A. G. Interpretable multi-timescale models for predicting fmri responses to continuous natural speech. *bioRxiv*, pp. 2020–10, 2021. URL https://doi.org/10.1101/2020.10.02.324392.

Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *Journal of Neuroscience*, 31(8):2906–2915, February 2011. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.3684-10.2011. URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3684-10.2011.

Mitchell, T., Shinkareva, S., Carlson, A., Chang, K.-M., Malave, V., Mason, R., and Just, M. Predicting human brain activity associated with the meanings of nouns. *Science*, 320:1191–1195, 2008.

URL https://www.science.org/doi/full/10.1126/science.1152876.

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, May 2011. ISSN 10538119. doi: 10.1016/j.neuroimage.2010.07.073. URL http://linkinghub.elsevier.com/retrieve/pii/S1053811910010657.

Pennington, J., Socher, R., and Manning, C. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL http://aclweb.org/anthology/D14-1162.

Price, C. J. A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, 62(2):816–847, August 2012. ISSN 10538119. doi: 10.1016/j.neuroimage.2012.04.062. URL http://linkinghub.elsevier.com/retrieve/pii/S1053811912004703.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., and others. Language models are unsupervised multitask learners. *OpenAI blog*, 1 (8):9, 2019. URL https://openai.com/blog/better-language-models.

Regev, M., Honey, C. J., Simony, E., and Hasson, U. Selective and Invariant Neural Responses to Spoken and Written Narratives. *Journal of Neuroscience*, 33(40):15978–15988, October 2013. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1580-13.2013. URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.1580-13.2013.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2105646118.

Toneva, M., Stretcu, O., Poczos, B., Wehbe, L., and Mitchell, T. M. Modeling task effects on meaning representation in the brain via zero-shot meg prediction. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5284–5295. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/38a8e18d75e95ca619af8df0da1417f2-Paper.pdf.

Varoquaux, G. and Poldrack, R. A. Predictive models avoid excessive reductionism in cognitive neuroimaging. *Current Opinion in Neurobiology*, 55:1–6, 2019. URL https://doi.org/10.1016/j.conb.2018.11.002.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., and Mitchell, T. Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. *PLoS ONE*, 9 (11):e112575, November 2014. doi: 10.1371/journal.pone.0112575. URL https://doi.org/10.1371/journal.pone.0123148.

A. Convergence of the Language Models During Training

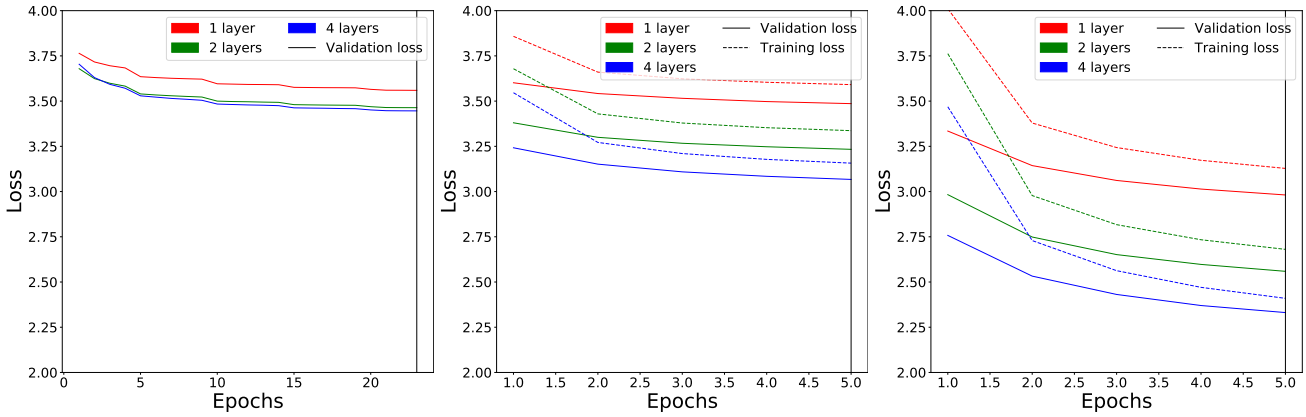

Figure 8. **Model convergence during training on the Full dataset.** Validation losses for all trained models, as well as training losses for transformers, for LSTM (left), GPT-2 (middle) and BERT (right). LSTMs and Transformer-based models were trained for 23 and 5 epochs, respectively.

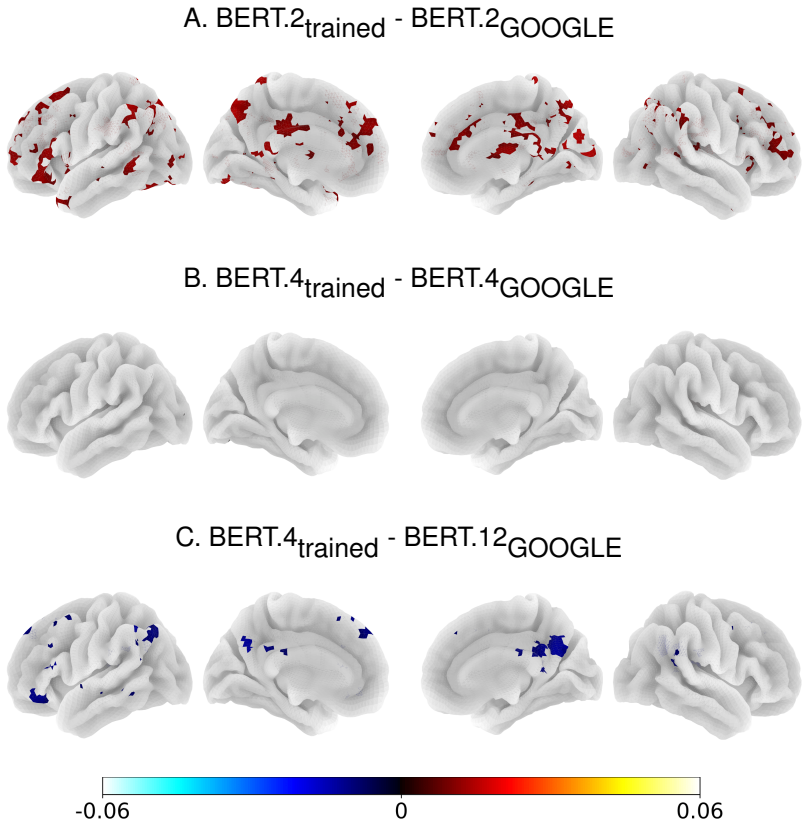

Figure 9. **Comparison of the trained BERT models with off-the-shelf baselines.** To assess the performance of our trained models, we compared their ability to predict brain data with that of off-the-shelf models (`https://github.com/google-research/bert`). The 2 and 4-layers BERT models either significantly outperform the baseline or are on a par. The 12-layers baseline, which is 3-times bigger than the 4-layers model, outperforms the latter in core regions of the language network, but only to a small extent.

B. Evaluation of Brain-Fit Performance of the Untrained Models

Model	LSTM	GPT-2	BERT
LSTM	.	81%	92%
GPT-2	.	.	86%

Table 2. **Overlap between untrained brain maps** The percentage of common voxels when the maps were thresholded at their 10% upper percentile.

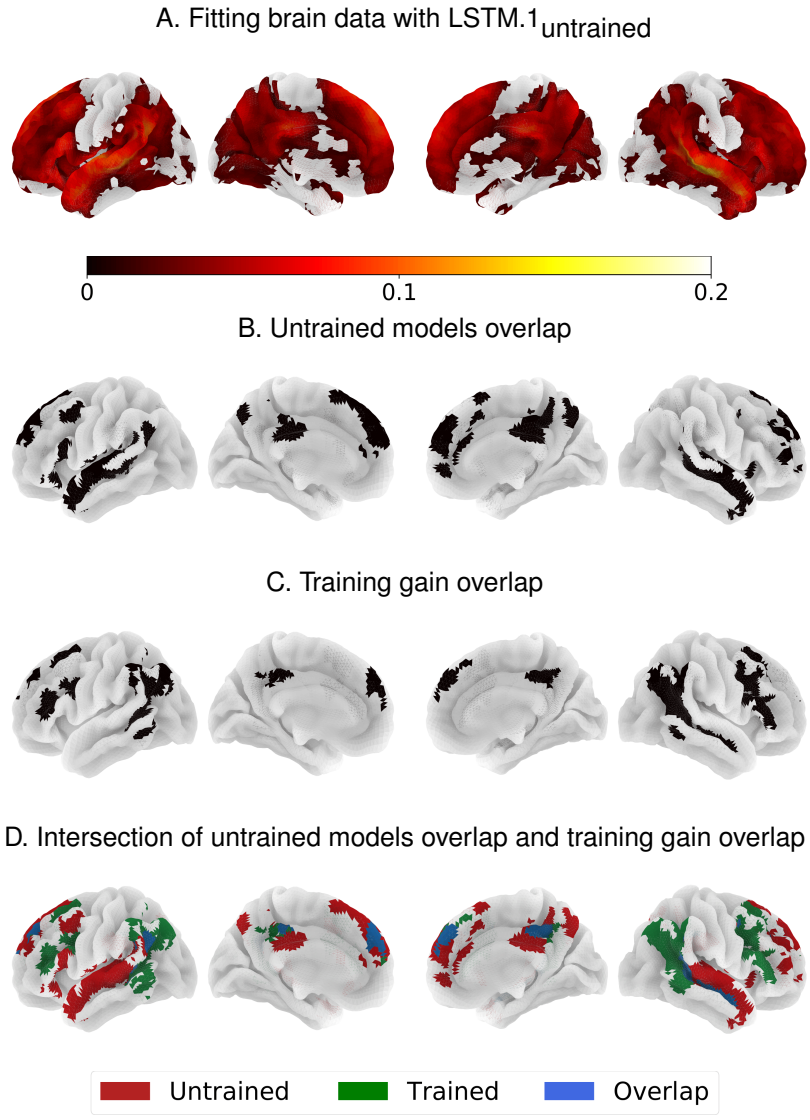

Figure 10. **A. Untrained LSTMs predict fMRI brain data better than chance across the entire brain.** Significant scores are observed in the language network, and non-significant scores in the motor cortex and the medial temporal regions. **B. Regions showing the strongest R score across the three 2-layer untrained models: LSTM, GPT-2, BERT.** (intersection of the three maps thresholded at the 10% upper percentile). There is a 79% overlap across the three untrained models. **C. Regions showing the strongest gains after training: LSTM, GPT-2, BERT.** (intersection of the three maps thresholded at the 10% upper percentile). There is a 75% overlap across the three trained models. **D. Representing shared and specific brain regions of the two overlaps.** The regions showing the strongest R scores across the three untrained models only have a 18% overlap with the regions showing the strongest gains across the three trained models.

C. Improvement in Brain Score after Training

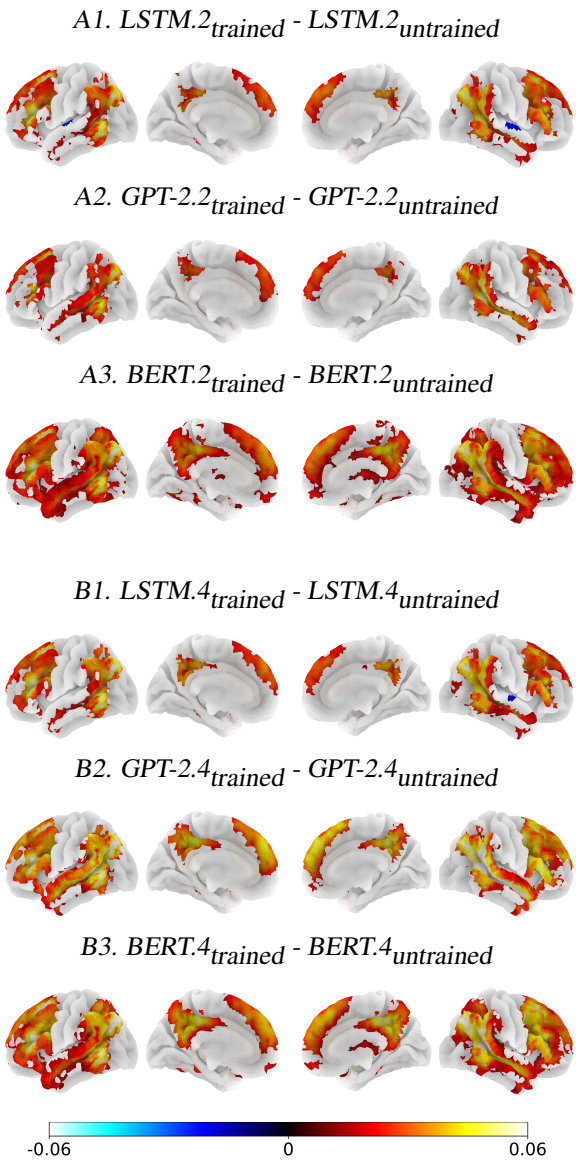

Figure 11. **A consistent increase in R scores due to training across various types of neural language models.** Contrasts between R scores of trained vs. untrained models. Contrast maps are shown for 2-layer models (panel A) and 4-layer models (panel B). In each panel, from top to bottom: LSTM, GPT-2, BERT.

D. Perplexity is not a Good Predictor of Brain Score

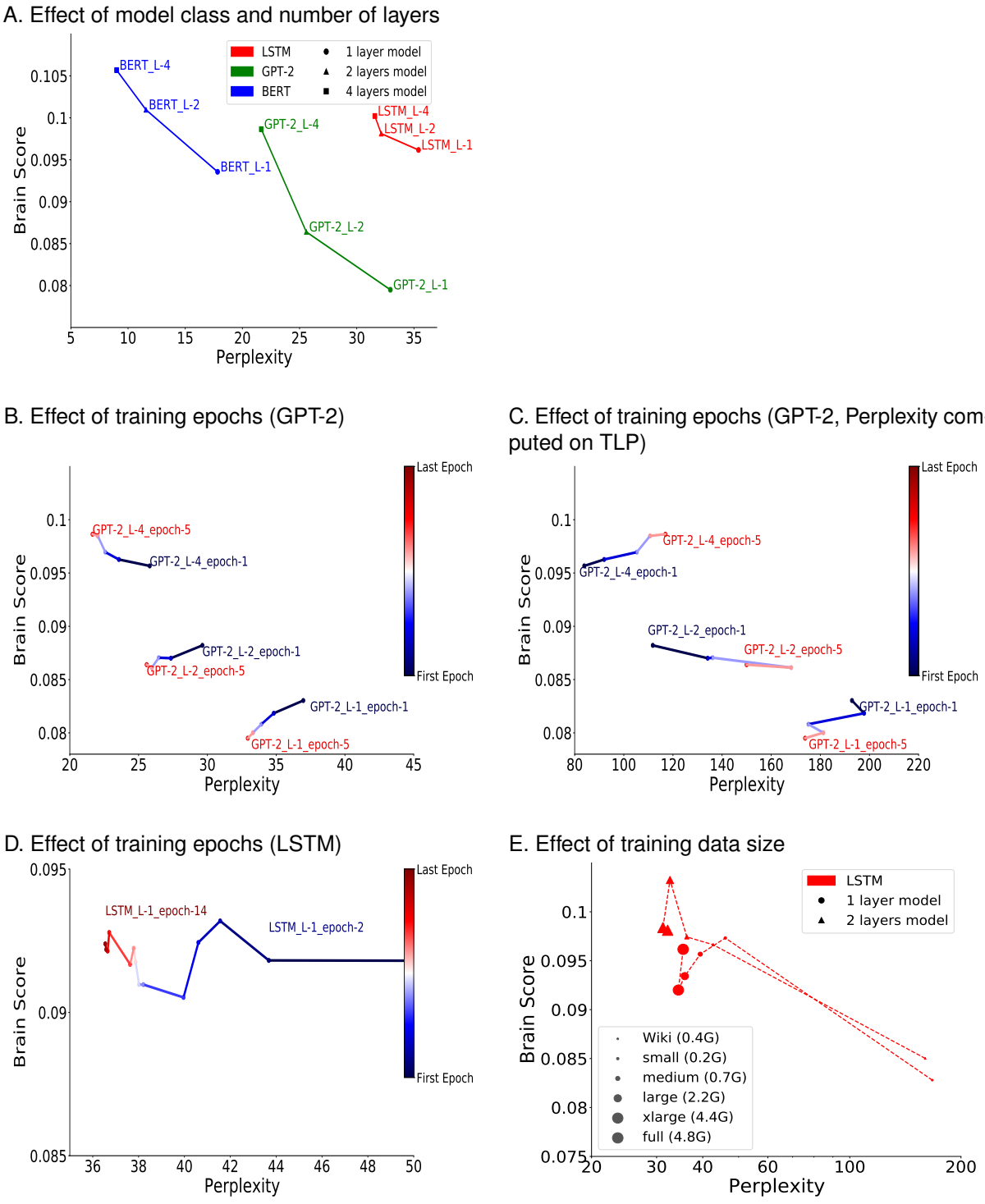

Figure 12. **Detailed analyses of the relation between brain score and perplexity as a function of model class (A), number of layers (A), training epochs (B-D) and training datasets (E).**

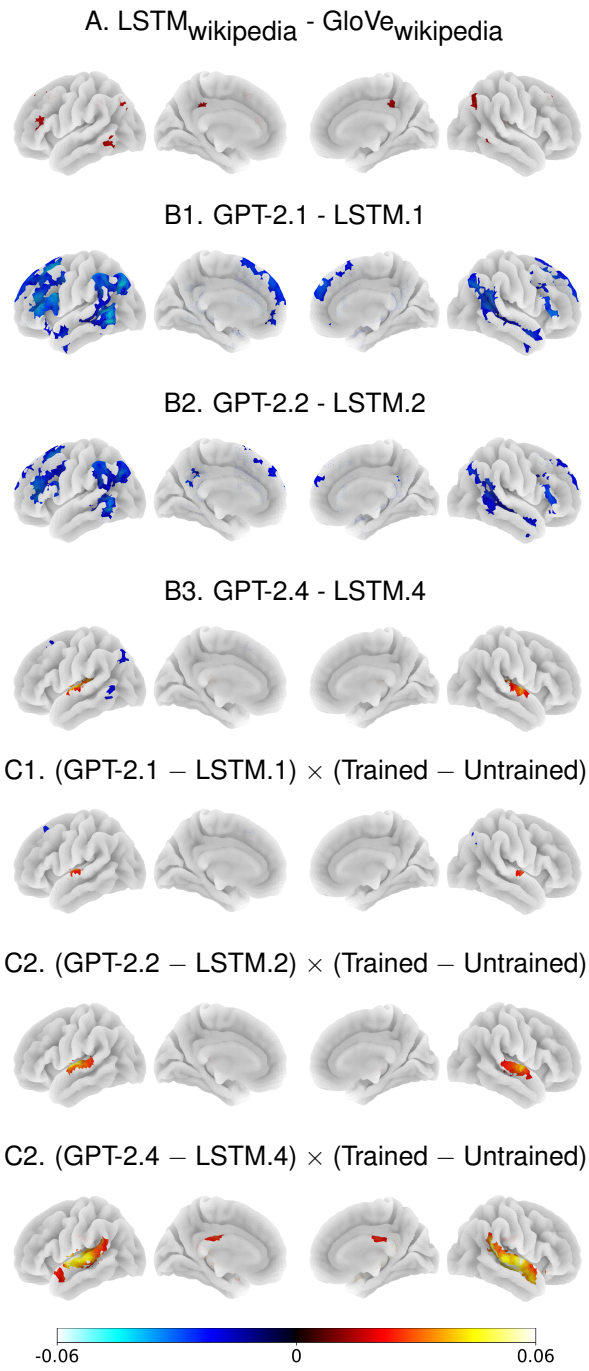

Figure 13. **Panel A) LSTM vs. GloVe when trained on the small Wikipedia dataset.** Contrast maps for LSTM vs. GloVe. Results strongly depend on the training dataset (compare panel S6A and Figure 5A), with less brain regions identified with the small dataset. **Panel B) GPT-2 vs. LSTM when trained on the full dataset.** Contrasts maps for GPT-2 vs. LSTM for 1-layer, 2-layers and 4-layers models (respectively panels B1, B2 and B3). GPT-2 better predicts brain activity around the Heschel gyri, for the 4-layer version, but LSTM outperforms GPT-2 in most of the language network.
Panel C) Interaction between model architecture (GPT-2 vs. LSTM) and Training (trained vs. untrained) for 1-layer (top), 2-layers (middle) and 4-layer (bottom) models. GPT-2 benefits more from training than LSTM. The more layers, the more it learns.

E. Relation Between Brain Score and Perplexity

Table 3. **Test Perplexities of models and their brains score in the SRM25 network.** For each model type, the best score is highlighted in bold.

MODEL	TRAINING DATASET	DATASET SIZE	PERPLEXITY	BRAIN SCORE
LSTM L-1 H-768	WIKIPEDIA	425M	167.25	0.0828
LSTM L-1 H-768	GUT. SMALL	240M	46.04	0.0973
LSTM L-1 H-768	GUT. MEDIUM	737M	39.39	0.0957
LSTM L-1 H-768	GUT. LARGE	2.2G	35.76	0.0934
LSTM L-1 H-768	GUT. XLARGE	4.4G	34.36	0.0920
LSTM L-1 H-768	GUT. XLARGE + WIKIPEDIA (FULL)	4.8G	35.40	0.0962
GPT-2 L-1 H-768	GUT. XLARGE + WIKIPEDIA (FULL)	4.8G	30.62	0.0795
BERT L-1 H-768	GUT. XLARGE + WIKIPEDIA (FULL)	4.8G	17.83	0.0898
LSTM L-2 H-768	WIKIPEDIA	425M	160.03	0.0850
LSTM L-2 H-768	GUT. SMALL	240M	42.67	0.0966
LSTM L-2 H-768	GUT. MEDIUM	737M	36.22	0.0974
LSTM L-2 H-768	GUT. LARGE	2.2G	32.61	0.1032
LSTM L-2 H-768	GUT. XLARGE	4.4G	31.21	0.0984
LSTM L-2 H-768	GUT. XLARGE + WIKIPEDIA (FULL)	4.8G	32.14	0.0981
GPT-2 L-2 H-768	GUT. XLARGE + WIKIPEDIA (FULL)	4.8G	26.22	0.0864
BERT L-2 H-768	GUT. XLARGE + WIKIPEDIA (FULL)	4.8G	11.57	0.0954
LSTM L-4 H-768	GUT. XLARGE + WIKIPEDIA (FULL)	4.8G	31.58	0.1002
GPT-2 L-4 H-768	GUT. XLARGE + WIKIPEDIA (FULL)	4.8G	23.62	0.0986
BERT L-4 H-768	**GUT. XLARGE + WIKIPEDIA (FULL)**	**4.8G**	**9.00**	**0.1057**
—	VALIDATION SET	1.1G	—	—
—	TEST SET	1.1G	—	—