We thank the reviewers for the thoughtful comments and attempt to address their questions, space permitting.

All reviewers: R1 and R2 correctly point out that relating intermediate representations in neural networks to brain activity has been previously explored (as we also state in L53-54). However, previous works make key untested assumptions about information contained in the neural network representations and use these representations to examine where/when this information is present the brain. Our work is the first, to our knowledge, to propose using brain activity for examining how this assumed information alters as the network representations change. While the brain has successfully informed computer vision (e.g. the hierarchy of CNNs is inspired from the visual system), the NLP field remains less convinced of the potential of the brain. We propose a framework to start changing this status-quo.

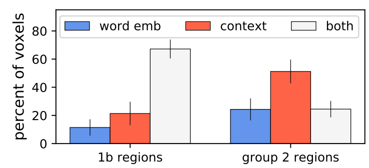
R1 and R3: The learned function f is evaluated in a classification task on held-out data using 4-fold cross-validation (L182-185). The classification task is to predict which of 2 sets of words was being read by the participant, which is a previously established metric. Our evaluation metric is the accuracy of this classification for each voxel (theoretical chance is 0.5), summarized by the mean and standard error across voxels (Fig 3). We will clarify this in the main text.

The motivation behind the attention removal experiments was to investigate how important the learned attention is in different layers. Following one of our more surprising findings that removing attention in shallow layers can actually improve brain activity prediction, we wanted to test the same network modification in an independent NLP setting.

R1: Due to typically limited data, we have observed that training non-linear predictive models for fMRI that outperform linear ones is difficult. Predicting fMRI directly from language networks is nonetheless interesting future work.

R2: To test the effect of the fixed-length evaluation window, we additionally computed ELMo features for word w_t by passing words $w_1, ..w_t$ from the sentence in which w_t appears through pretrained ELMo. We found that predictions of fMRI activity using sentence features correlate strongly with predictions from ELMo representations obtained from a fixed window longer than 5 words (0.8 ± 0.01 mean Pearson correlation across subjects).

Following R2's suggestion, we computed the percentages of voxels within the ROIs that are well explained by word embeddings (blue), long-context representations (red), or both (white). These quantitative results (mean percentages over subjects for ELMo, shown to the right) further validate our conclusions that both word embeddings and long-context can explain 1b regions well, while group 2 regions are best predicted by long-context. We will include plots for all models.



R3: We will add more references to ground our statements about the brain. We address our hypothesis that a neural representation can be decomposed across time points and locations in the brain through aligning with fMRI and MEG in two experiments. First is a proof of concept (L188), showing that the ELMo word embedding aligns with times and locations in MEG corresponding to known processing of word length and part-of-speech. These are expected properties of word embeddings that can be tested with more traditional methods, and we wanted to verify that our method is also able to expose these. The second experiment was to contrast a word embedding with sequence embeddings from 4 different NLP models in their abilities to align with different locations that are known to processes single-word information and long-range context information to different extents (results in Fig 2). We agree that there is more to be done to fully characterize the many types of information contained in a neural embedding in future work.

We agree that there are other tasks that could have led to removing attention. However, we argue that predicting brain activity is more informative as it provides additional insights, such as a decomposition of the neural representation across the brain. Further, the attention experiment supports our premise that similarity of language representations to brain activity is useful, and that it reveals what representations are more relevant for language tasks, a fact we can capitalize on in future research. While retraining with the modified architecture is a good next step, the current setup tests whether the modified representations themselves, before retraining, have more language relevant information.

The syntactic tasks measure subject-verb agreement, so the "incorrect verb" is the wrongly-numbered correct verb (e.g. incorrect verb is "are" if the correct verb is "is", as in the example in L268). We will clarify this in Section 5. Following R3's suggestion, we tested the significance of the differences in accuracy on each task in Table 1 between the base model and the uniform-attention models. The uniform-attention models in the early layers presented in Table 1 (L1, L2, L6) significantly outperform the base model (paired t-test, significance level 0.01, FDR controlled for multiple comparisons) in 8 of the 13 tasks. The two numerical improvements in Table 1 that do not survive the statistical test are for the tasks “short VP coordination” and “across an object relative clause (no that)”. We will indicate this in Table 1.

The delay in fMRI is due to the hemodynamic response. We account for it by building predictive models with features from words occurring in previous time points, which is a common way to correct for the delay [Nishimoto 2011 Cur. Biol.,Huth 2016 Nature], and in this way we avoid any confounding from the delay (section 3.1 of supplementary which we can add to main text). MEG does not suffer from such latency, as information due to the current word is detected in the recordings within 100 milliseconds after word presentation [Pulvermüller 2011 EJNeur].