Language models and brains align due to more than next-word prediction and word-level information

Gabriele Merlin
MPI for Software Systems
Saarbrücken, Germany
gmerlin@mpi-sws.org

Mariya Toneva
MPI for Software Systems
Saarbrücken, Germany
mtoneva@mpi-sws.org

Abstract

Pretrained language models have been shown to significantly predict brain recordings of people comprehending language. Recent work suggests that the prediction of the next word is a key mechanism that contributes to this alignment. What is not yet understood is whether prediction of the next word is necessary for this observed alignment or simply sufficient, and whether there are other shared mechanisms or information that are similarly important. In this work, we take a step towards understanding the reasons for brain alignment via two simple perturbations in popular pretrained language models. These perturbations help us design contrasts that can control for different types of information. By contrasting the brain alignment of these differently perturbed models, we show that improvements in alignment with brain recordings are due to more than improvements in next-word prediction and word-level information.

1 Introduction

Language models (LMs) that have been pretrained to predict the next word over billions of text documents have also been shown to significantly predict brain recordings of people comprehending language (Wehbe et al., 2014b; Jain and Huth, 2018; Toneva and Wehbe, 2019; Caucheteux and King, 2020; Schrimpf et al., 2021; Goldstein et al., 2022). Understanding the reasons behind the observed similarities between representations of language in machines and representations of language in the brain can lead to more insight into both systems. Recent studies suggest that the prediction of the next word is a key mechanism that contributes to the alignment between the two (Goldstein et al., 2022). What is not yet understood is whether prediction of the next word is necessary for this ob-
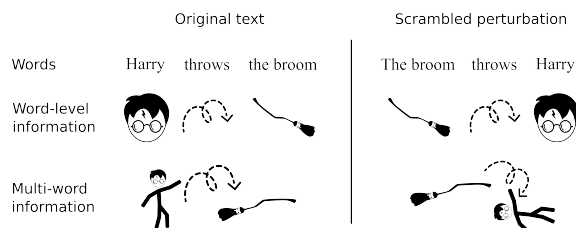

Figure 1: An illustration of additional information that may be important for alignment between language models and brain recordings. Our approach is largely agnostic about the exact linguistic information contained in the conceptual quantities "word-level information" and "multi-word information", and the only assumption is that "word-level information" is not affected by word order.

served alignment or simply sufficient, and whether other shared information is similarly important.

Understanding the impact of other kinds of information on brain alignment is complicated by correlations with next-word prediction (NWP). Because NWP is the LM training objective, better NWP may also be related to improved representations of other types of information (Piantadosi and Hill, 2022) that the human brain is sensitive to, such as word-level or multi-word information (Lerner et al., 2011). Neuroscientists are still investigating the exact linguistic features at the word- and multi-word levels that are important for processing in the brain, so for the remainder of the paper, we take an abstract approach and refer to "word-level information" as the non-contextualized representation of the word, and to "multi-word information" as relating to multiple words (e.g. syntax). For example, each word in "Harry throws the broom" has a non-contextualized meaning and the phrase has a different meaning depending on the word order ("Harry throws the broom" vs. "The broom throws Harry", see Figure 1). We note that these are conceptual quantities and not ones we are claiming to be able to quantify directly. The only assumption

---

Code available at
github.com/gab709/brain-llm-beyond-next-word.

key to our argument is that word-level information is not impacted by word order. Other than that, our methods are agnostic about the specific linguistic information contained by these conceptual quantities. Both word-level and multi-word information may contribute to brain alignment, but their effect cannot be disentangled from that of next-word prediction using previous approaches.

In this work, we aim to disentangle the contributions of next-word prediction and word-level information from other factors, such as multi-word information, in the brain alignment of GPT-2-based models (Radford et al., 2019). Our methodology builds upon the traditional neuroscientific approach of constructing *contrasts* between brain activity elicited by different conditions. A contrast reveals the processing of a specific property P in the brain, by subtracting the brain activity elicited by two conditions (condition A and B) that are tightly controlled to contain similar information except for the target property P. We leverage this approach and design contrasts between the *predicted* brain activity by two related models: an original model and its perturbed version. By contrasting the brain alignment of these two models, we can conclude that any difference in brain alignment is due to the perturbation. Conversely, if the two contrasted conditions are controlled for some factor, then any difference in the *predicted* brain activity between the two conditions cannot be due to this factor.

Our key insight is to design a contrast that controls for both information related to next-word prediction and word-level information. This contrast is enabled by two proposed perturbations. The first perturbation, which we name *input scrambling*, scrambles the order of the input words at inference time. This perturbation controls for the word-level information when contrasting a model's brain alignment related to the original vs. the scrambled inputs, because, by definition, the word-level information encoded in the model representations remains the same. Any remaining brain alignment after the contrast must therefore be due to factors beyond word-level information, such as next-word prediction or multi-word information. The second perturbation further disentangles the contribution of next-word prediction to brain alignment. This perturbation, which we name *stimulus-tuning*, fine-tunes a model to predict the next word in the specific naturalistic stimulus text corresponding to the brain recordings. The fine-tuning is done until the next-word prediction performance matches a pre-

defined level above its input-scrambled version, similar to the baseline model's improvement over its input-scrambled version. Contrasting brain predictions from pairs of models—-baseline vs. scrambled and stimulus-tuned vs. scrambled—-controls for both word-level information and next-word prediction. Any residual brain alignment is then due to other factors, such as multi-word information. We note that our methodology is based on models trained to predict the next word; however, we only control for next-word prediction performance at inference time. Therefore, we cannot draw conclusions about the importance of the training objective itself, but rather about the information that has emerged after the training process.

After controlling for word-level and next-word prediction in the final contrast, we still observe residual brain alignment. Across three types of models (GPT-2-small, GPT-2-medium (Radford et al., 2019) and GPT-2-distilled (Sanh et al., 2019)), we find consistent residual brain alignment in two specific brain areas that are thought to process language (Fedorenko et al., 2010; Fedorenko and Thompson-Schill, 2014)–the inferior frontal gyrus (IFG) and the angular gyrus (AG)– suggesting that the brain alignment between the language model and these brain regions is due to more than next-word prediction and word-level information. We speculate that this alignment is due to multi-word information, which is consistent with previous findings about processing in these regions (Friederici, 2012; Humphreys et al., 2021).

Our main contributions are as follows: (i) propose perturbations to pretrained language models that, when combined in suitable contrasts, can control for the effects of next-word prediction and word-level information on brain alignment; (ii) demonstrate that a proposed perturbation, which consist of tuning a language model on a validation stimulus text, can increase the alignment with brain recordings that correspond to a heldout text; (iii) reveal that the brain alignment with language regions, in particular in the inferior frontal gyrus (IFG) and the angular gyrus (AG), is due to more than next-word prediction and word-level information.

2 Methods

2.1 Baseline models

We use GPT-2-based language models (Radford et al., 2019) as the baseline pretrained language models. In particular, we investigate GPT-2-small,

GPT-2-medium (Radford et al., 2019) and GPT-2-distilled (Sanh et al., 2019). GPT-2-based models achieve strong results on a variety of natural language processing tasks such as question answering, summarization, and translation, without any specific training beyond next-word prediction. Furthermore, we analyze GPT-2-based models to allow for a direct comparison with prior brain alignment research (Goldstein et al., 2022; Schrimpf et al., 2021). We will present the averaged results across the three types of models in the main paper, and the individual results in the Appendix H, I, J. We observed that as the model size increases, the effect of the stimulus-tuning perturbation as well as the residual effect after the contrast decrease. This reduction in effect is likely due to the small size of the dataset that we use for fine-tuning, which limits the learning capacity of larger models that already have a better next-word prediction ability. Therefore, we did not include additional larger models. For the baseline models we use the checkpoints provided by Huggingface (Wolf et al., 2020b)[1][2][3].

2.2 fMRI data

To evaluate the brain alignment of GPT-2 and of its perturbations, we use publicly available fMRI data provided by Wehbe et al., 2014a, one of the largest publicly available fMRI datasets in terms of samples per participant. fMRI data were obtained from eight participants as they read chapter 9 of Harry Potter and the Sorcerer's Stone (Rowling et al., 1998) word-by-word. One fMRI image (TR) was acquired every 2 seconds (TR = 2 sec). The chapter was divided into four runs of approximately equal length and participants were allowed a short break at the end of each run. Each word of the chapter was presented for 0.5 seconds, after which a new word was presented immediately.

2.3 Evaluation tasks

We use two tasks to evaluate models: prediction of the next-word and brain alignment. Importantly, both tasks are evaluated using the same text, which corresponds to the fMRI stimulus. For consistency, we use the same setting to evaluate both next-word prediction and brain alignment: we evaluate each metric as described below using sliding windows of 20 consecutive words (overlapping by 16 words,

which corresponds to 4 TRs). We choose this window length because previous work has shown that using contexts larger than 20 words does not substantially improve brain prediction performance with similarly-sized language models (Toneva and Wehbe, 2019). We empirically verified that this also holds in our setting.

Next-word prediction. To generate the next token, we follow best practices for GPT-2-based models which consist of a linear prediction head with weights tied to the input embeddings (Wolf et al., 2020a). We evaluate the next-word prediction performance using the cross-entropy measure.

Brain alignment. To measure the brain alignment between a GPT-2-based model and the fMRI recordings, we employ a standard linear prediction head on top of the last transformer block. This prediction head learns a function that maps input stimulus representations to output brain recordings and is frequently used to measure how well word representations obtained from a language model can predict brain recordings (Jain and Huth, 2018; Toneva and Wehbe, 2019; Schrimpf et al., 2021). Similarly to previous work, we parameterize this function as a linear function, regularized using the ridge penalty. We train this function in a cross-validated way and test its performance on the data that was heldout during training. We select the ridge parameter via nested cross-validation. As a result, for each participant, we train four functions, then we aggregate the predictions and evaluate the brain alignment. We provide further details about this prediction head in Appendix A. We evaluate the brain alignment using Pearson correlation, computed between the predictions of heldout fMRI recordings and the corresponding true data. Specifically, for a model q and voxel v_j with corresponding heldout fMRI y_j, the brain alignment is computed as

$$\texttt{brain alignment}(q, v_j) = \texttt{corr}(\hat{y}_j, y_j),$$

where $\hat{y}_j = q(X)W_{q,j}$, X is the input text sample to model q, and $W_{q,j}$ are the learned prediction weights corresponding to the voxel. All voxel-wise brain alignment scores are visualized on each participant's brain surface using PyCortex (Gao et al., 2015).

2.4 Perturbations

We aim to disentangle the effects on brain alignment of different types of information contained

---

[1] https://huggingface.co/openai-community/gpt2
[2] https://huggingface.co/openai-community/gpt2-medium
[3] https://huggingface.co/distilbert/distilgpt2

in language models, that we describe in Section 1: next-word prediction, word-level information, and multi-word information. To achieve this, we designed two perturbations that isolate the contributions of these different types of information to the brain alignment when used as part of carefully designed contrasts (see Section 2.5).

Input scrambling. The aim of the first perturbation is to control for the effect of word-level information on brain alignment. This perturbation consists of scrambling the words at inference time in each text sequence that we use to predict one fMRI TR image (i.e., 20 consecutive words). The order of words has been shown to be marginally important for other downstream tasks at inference time (Sinha et al., 2021). Therefore, if the words are scrambled, even though the next-word prediction ability will decrease, we expect the model to still predict the next word above chance level, using information from the 20-word context.

Stimulus-tuning. The second perturbation fine-tunes the baseline pretrained model with the next-word prediction objective on a training portion of the stimulus text. To perform the stimulus-tuning, we select training samples that consist of non-overlapping sequences of 80 consecutive words. For each baseline model, we trained four models, one for each held-out run of the fMRI data (see Appendix B for more details on model training).

We expect stimulus-tuning to improve all three brain-relevant types of information we consider (next-word prediction, word-level information, and multi-word information). Therefore, we also expect that the stimulus-tuned model will also exhibit better brain alignment than the baseline. However, stimulus-tuning itself is not sufficient to investigate the independent effect of either type of information on brain alignment. This perturbation is useful when combined with the input scrambling perturbation. By ensuring that the drop in next-word prediction accuracy after the scrambling perturbation is similar to the drop in the baseline after scrambling, we can control for both next-word prediction and word-level information (see Section 2.5).

2.5 Contrasts to disentangle brain alignment factors

Baseline $-$ Baseline scrambled: We first consider the contrast of brain alignment between the baseline model and its scrambled counterpart. Any change in brain alignment between the two can be due to changes in word-level information, next-word prediction, or other factors, such as multi-word information:

$$\Delta^{base} = \Delta^{base}_{WL} + \Delta^{base}_{NWP} + \Delta^{base}_{*}, \qquad (1)$$

where Δ^{base} is the change in brain alignment of the baseline model due to scrambling, Δ^{base}_{WL} is the change in alignment related to differences in word-level information, Δ^{base}_{NWP} is the change in alignment related to differences in next-word prediction, and Δ^{base}_{*} is the change in alignment related to other factors, such as multi-word information. By definition, word-level information is not affected by context, so perturbing the order of words in the input does not affect word-level information. Therefore, $\Delta^{base}_{WL} = 0$ which simplifies Eq. 1:

$$\Delta^{base} = \Delta^{base}_{NWP} + \Delta^{base}_{*}. \qquad (2)$$

(Stimulus-tuned$-$Stimulus-tuned scrambled)

$-$ (Baseline$-$Baseline scrambled):

Any residual brain alignment in the previous contrast (i.e. Δ^{base}) may still be due to differences in next-word prediction information (Δ^{base}_{NWP}). Therefore, while the previous contrast controls for the effect of word-level information on brain alignment, it is not able to additionally control for the effect of next-word prediction. To control for the next-word prediction effect, we designed a second-level contrast i.e. (stimulus-tuned $-$ stimulus-tuned scrambled) vs (baseline $-$ baseline scrambled) (Δ^{stim} - Δ^{base}). Similarly to the baseline model, we can show that $\Delta^{stim} = \Delta^{stim}_{NWP} + \Delta^{stim}_{*}$. Therefore, the residual brain alignment of the proposed second-level contrast is as follows:

$$\Delta^{stim} - \Delta^{base} = \Delta^{stim}_{NWP} + \Delta^{stim}_{*} \qquad (3)$$
$$- \Delta^{base}_{NWP} - \Delta^{base}_{*}.$$

To control for the next-word prediction, we specifically select a fine-tuning checkpoint of the stimulus-tuned model such that the change in next-word prediction performance due to scrambling is similar to that of the baseline (i.e. $\delta^{stim} \approx \delta^{base}$, where δ signifies the change in next-word prediction between a model and its scrambled counterpart). Therefore, $f(\delta^{stim}) - f(\delta^{base}) \approx 0$, when f is a linear function. Previous work has shown a linear relationship f between next-word prediction and brain alignment (Schrimpf et al., 2021), which also holds

18434

with high correlation in our setting (0.61 Pearson correlation, see Appendix E). Therefore:

$$\Delta_{NWP}^{stim} - \Delta_{NWP}^{base} \approx 0. \qquad (4)$$

Combining Eq. 3 and Eq. 4, we see:

$$\Delta^{stim} - \Delta^{base} \approx \Delta_*^{stim} - \Delta_*^{base}. \qquad (5)$$

Therefore, if any brain alignment remains after this second-level contrast, it must be due to factors beyond next-word prediction and word-level information, such as multi-word information.

3 Results

In this section, we report the results averaged across models. The results for the individual models can be found in Appendices H, I, J.

3.1 Next-word prediction

In Figure 2A, we report the next-word prediction performances of the GPT-2-small model and the corresponding perturbed models. The results for GPT-2-distilled and GPT-2-medium are consistent and reported in Appendix Figures 14, 20. We observe that the stimulus-tuned model performs better than the baseline. This verifies that stimulus-tuning indeed improves the model's ability to predict the next word in the heldout stimulus set.

As expected, the performance of the scrambled models is worse than their unscrambled counterparts (i.e. baseline scrambled vs. baseline and stimulus-tuned scrambled vs. stimulus-tuned). We further observe that the next-word prediction performance of the stimulus-tuned scrambled model is still better than that of the baseline, indicating that the information gained by stimulus-tuning is not entirely counteracted by the scrambling perturbation.

3.2 Brain alignment

Figures 2C-F visualize the brain alignment of the baseline model (i.e., Pearson correlation between predicted and true brain recordings) and the percentage change between pairs of models for one participant for GPT-2-small. Results for the remaining participants and models are largely consistent and are shown in Appendix Figs. 9, 21, 15.

Effects of stimulus-tuning. In Figure 2D, we observe that the stimulus-tuned model aligns better with the brain recordings than the baseline, particularly in many brain areas that have been previously implicated in language-specific processing

(Fedorenko et al., 2010; Fedorenko and Thompson-Schill, 2014) and word semantics (Binder et al., 2009) (visualized in Figure 2B and listed in Appendix C). We quantify the improvement across models and participants in brain alignment due to stimulus-tuning in language processing regions versus non-language processing regions in Figure 3. Here, we demonstrate that the stimulus-tuned model has higher brain alignment in language-related regions than in non-language-related regions. This indicates that the stimulus tuning perturbation contributes to an improvement in the model's performance, particularly in language-related ROIs. The results shown are computed using all voxels in the brain, including a large number of noisy voxels. Therefore, the reported numbers are numerically low. To focus on more informative voxels, we quantify the difference in each language-related ROI and use an estimate of the noise ceiling in each voxel to discard noisy voxels (see Appendix G for details). In Figure 4, we present the average percentage gain in brain alignment due to stimulus-tuning across models in each specific language ROI. Here we include only voxels that have estimated noise ceiling values > 0.05. The figure reveals a positive gain of the stimulus-tuned model over the baseline in every language region.

The results show that stimulus-tuning leads to both an improved ability to predict the next word and an improved alignment with fMRI recordings, but we are not able to conclude that the improvement in alignment with the brain is due to the improved prediction of the next word. The reason is that improving a model's ability to predict the next word may also improve other aspects of the model that are brain-relevant, such as its ability to represent word-level or multi-word information, that are specific to the stimulus narrative.

Effects of scrambling. In Figure 2E-F, we observe that for both the baseline and stimulus-tuned model the scrambling perturbation affects the alignment, particularly in the language-related ROIs (see Figure 2B).

However, both the baseline scrambled and stimulus-tuned scrambled models still align with the fMRI recordings, particularly in the language ROI, as shown in Appendix Figures 9, 15, 21. This suggests that even when perturbing the next-word prediction capability and multi-word information, a language model is able to strongly align with brain areas that are thought to process language.
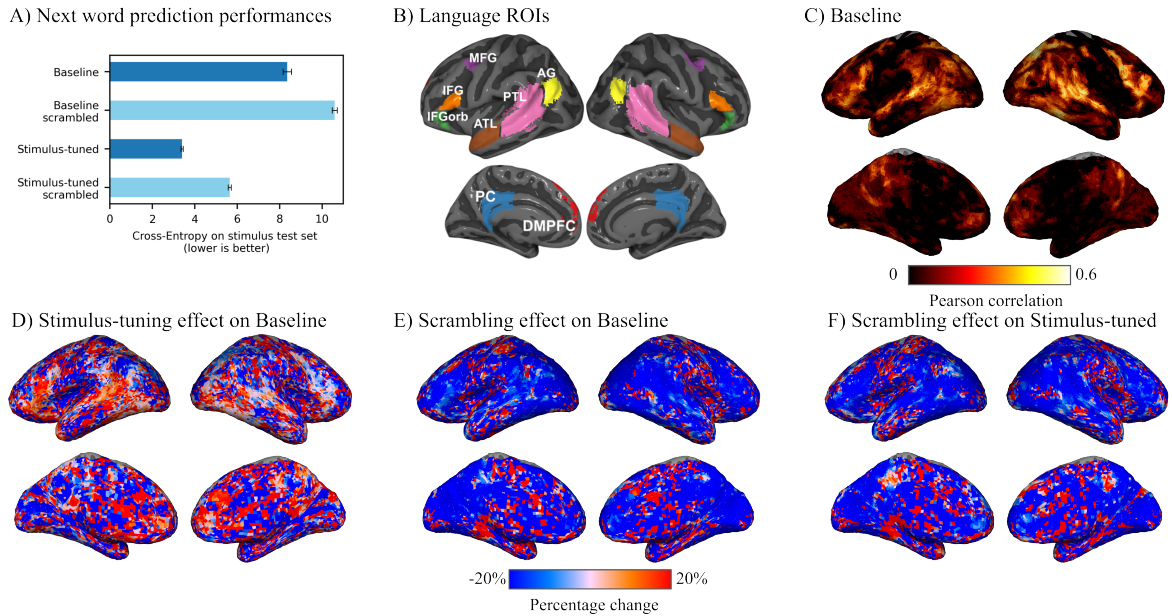
Figure 2: Performances of the GPT-2-small baseline and perturbed models at next-word prediction averaged across runs with standard deviation (A) and brain alignment (C-F). Stimulus-tuning improves both the next-word prediction (stimulus-tuned vs baseline in (A)) and brain alignment (D). Instead, scrambling reduces the next-word prediction (baseline vs baseline scrambled in (A)) and reduces the brain alignment (E and F). Despite the reduction in alignment due to the scrambling perturbation, all four models exhibit alignment in language processing regions (B) (see Appendix Figure 9 for brain alignment plots for all participants and Appendix Figures 15, 21 for other models.
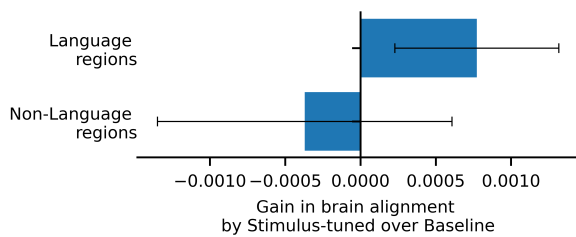
Figure 3: Impact of the stimulus-tuning perturbation on the baseline model. For each model (GPT-2-small, medium, distill) we computed the median difference in language and non-language regions across participants. Here we display the average difference across models as well as the standard deviation. Results for the single models are reported in Appendix Figures 10, 16, 22.

Interestingly, we observe that the effect of scrambling on the stimulus-tuned model is much larger for brain alignment than for next-word prediction. For next-word prediction, the stimulus-tuned scrambled performs worse than the stimulus-tuned model but better than the baseline. In Appendix Figures 9, 15, 21, we see that the stimulus-tuned scrambled performs worse at brain alignment, not only with respect to the stimulus-tuned model but also to the baseline. This is an initial hint that next-word prediction is not the only key information in

aligning language models and brain recordings.

We show that scrambling affects both the next-word prediction ability and the brain alignment of language models. However, we are not able to draw a conclusive link yet. The reason is that the scrambling procedure only controls for word-level information, but not for any possible changes in multi-word information, which may also contribute to the decrease of alignment with the language processing brain areas.

3.3 Controlling for both word-level information and next-word prediction

In Figure 5, we report the average percentage gain by (stimulus-tuned - stimulus-tuned scrambled) over (baseline - baseline scrambled) (Δ^{stim} - Δ^{base}), across GPT-2 models, in each specific ROI, including voxels that have estimated noise ceiling values > 0.05 (see Figures 13, 19, 25 in Appendix for the corresponding brain plot of each model). Given the high variability across subjects and the number of subjects in our setting, obtaining statistically significant results is challenging (see Appendix D for details). Despite these challenges we observe that there is still a positive residual brain alignment after the contrast for two language processing regions, the Inferior Frontal Gyrus (IFG)
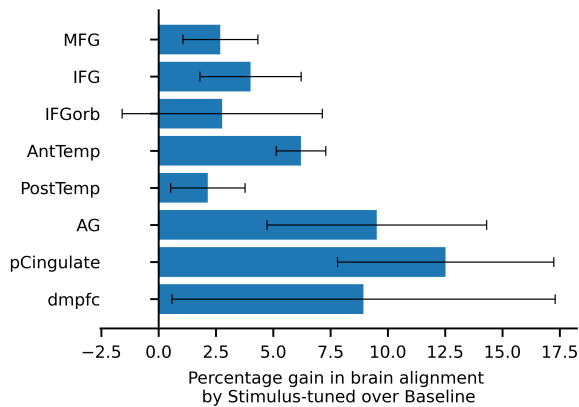
Figure 4: Impact of the stimulus-tuning perturbation on the baseline model. For each model (GPT-2-small, medium, distill) we computed the median percentage gain by stimulus-tuned over baseline in language regions across participants. Here we display the average percentage gain across models as well as the standard deviation. We include only voxels with estimated noise ceiling values >0.05. Results for the single models are reported in Appendix Figures 11, 17, 23.
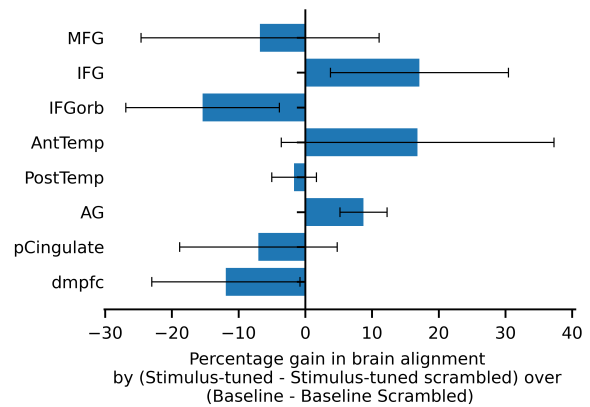

Figure 5: Impact of the scrambling perturbation on the stimulus-tuned model versus its impact on the baseline model. For each model (GPT-2-small, medium, distill) we computed the median percentage gain by (stimulus-tuned - stimulus-tuned scrambled) over (baseline - baseline scrambled) in language regions across participants. Here we display the average percentage gain across models, as well as the standard deviation. We include only voxels with estimated noise ceiling values >0.05. Results for the single models are reported in Appendix Figures 12, 18, 24.

and Angular Gyrus (AG), even when controlling for next-word prediction and word-level information, across three different models. This is evidence that the alignment with the language model in these areas is due to more than next-word prediction and word-level information (See Eq. 5).

4 Related Works

Several previous studies have investigated the alignment between pretrained language models and brain recordings of people comprehending language, finding significant similarities (Wehbe et al., 2014b; Jain and Huth, 2018; Toneva and Wehbe, 2019; Abdou et al., 2021; Schrimpf et al., 2021; Hosseini et al., 2024). Our work builds on these and further studies the reasons for these similarities. The work of Goldstein et al. (2022) is most directly related to our research question, as they suggest that the prediction of the next word is a key reason for the alignment between language models and brain recordings, based on evidence that ECoG electrodes can predict the neural network representation of upcoming words in a story. Our work uses perturbations to disentangle next-word prediction from other types of information that may affect brain alignment (word-level and multi-word information) and offers an additional account of the necessary information for brain alignment.

Our work also relates to a growing body of re-

search on disentangling the contributions of different types of information to the alignment between brain recordings and language models. Toneva et al. (2022a) present an approach to disentangle supra-word meaning from lexical meaning showing that the supra-word meaning is predictive of fMRI recordings in two language regions (Anterior and Posterior Temporal Lobes), which was further adapted by Oota et al. (2024a) and Oota et al. (2024b) to disentangling effects of other linguistic properties. Caucheteux et al. (2021) and Reddy and Wehbe (2021) aim to disentangle alignment due to syntactic and semantic processing. Toneva et al. (2022b) examine whether representations obtained from a language model align with different language processing regions in similar or different ways. Kauf et al. (2023) investigate the contribution of word-level semantics to the brain alignment of language models, showing that syntactic perturbations have a lesser impact on brain alignment compared to semantic perturbations. Gauthier and Levy (2019) demonstrate that fine-tuning language models on scrambled data has been shown to be beneficial for brain decoding. Our experiments reveal that the scrambling perturbation influences brain alignment. Despite this, when sentences are scrambled at inference time, the model retains the capability to predict brain responses in regions as-

sociated with language processing. However, the aim of our work is not to directly evaluate the significance of semantics or syntax, but rather to investigate the effects on brain alignment when information relevant to next word prediction is controlled for. Therefore, our proposed perturbations are complementary to these previous works and may yield additional insights if combined.

Other studies have used perturbations related to word order to investigate some properties of language models. Pandia and Ettinger (2021) introduced distracting content to test how robustly language models retain and use that information for prediction, showing that language models are particularly susceptible to semantic similarity and word position. Papadimitriou et al. (2022) applied a perturbation (scrambling method) to investigate where the semantic and syntactic processing is taking place in BERT, revealing that early layers care more about the lexicon, while the latter layer care more about word order. Our current work contributes to this research direction by examining the effects of scrambling on both brain alignment and language modeling.

Finally, a work by Aw and Toneva (2023) finetunes language models to summarize narratives and finds improved brain alignment, despite a lack of improvement in next-word prediction. While this finding suggests a similar conclusion to the one from our work–that next-word prediction performance is not necessary for improved brain alignment–the perturbation approach in our work allows additional control over the language model representations and is complementary to this previous work.

5   Discussion

We showed that the perturbation that we termed *stimulus-tuning* (i.e., finetuning a pretrained model on a validation stimulus text) can increase the alignment with brain recordings that correspond to a heldout text, particularly in several language processing brain areas. We quantified this improvement by comparing the stimulus-tuned model and the baseline in these brain areas. Stimulus-tuning may improve brain alignment due to improved ability to represent the next word, previously seen individual words, or multi-word information that are specific to the stimulus narrative.

Using the perturbation that we termed *input scrambling*, we showed that the improved next-

word prediction capabilities of the stimulus-tuned model is not the only reason for improved brain alignment. We showed that leveraging a contrast that controls for word-level information and next-word prediction, we still obtain a residual brain alignment. Specifically, we show that, across multiple GPT-2 models, improvements in alignment with brain recordings in two language regions—Inferior Frontal Gyrus (IFG), Angular Gyrus (AG)—(see Figure 5) are due to more than improvements in next-word prediction and word-level information.

One possible reason for this improvement in brain alignment is improved capabilities to represent multi-word information that are specific to the stimulus text. This hypothesis aligns with previous work that has found the Inferior Frontal Gyrus (IFG) to be sensitive to syntax (Friederici et al., 2003; Friederici, 2012) and the Angular Gyrus (AG) to multi-word event structure (Ramanan et al., 2018; Humphreys et al., 2021). Note that the fact that we do not find strong effects in other language regions does not necessarily mean that they do not process multi-word information.

6   Conclusion

This work aims to deepen our understanding of the existing alignment between language models and brain recordings. We proposed two perturbations to pretrained language models that, when used together, can control for the effects of next-word prediction and word-level information on the alignment with brain recordings. Using these controls, we show that improvements in brain alignment are due to more than improvements in next-word prediction and word-level information. Our findings are relevant for both cognitive neuroscientists as well as natural language processing researchers.

The findings are relevant for cognitive neuroscientists because they suggest that accurate prediction of the next word is not a necessary condition for brain alignment. It is possible that learning to accurately predict the next word is sufficient for inducing other properties in the language model representations that are necessary for brain alignment, such as multi-word information, and future work can further examine this hypothesis.

Our findings are also relevant to NLP researchers who examine what language models can learn from only text. We show that finetuning a language model with small amounts of text can increase its alignment with never-before-seen brain recordings,

and that this improvement in brain alignment is not purely due to next-word prediction or word-level information. This finding suggests that training a language model with little additional text can improve its representations of multi-word information in a brain-relevant way. We note that while our methodology controls for next-word prediction ability at inference time, it still relies on the next-word prediction objective during training. However future work can investigate alternative training objectives, that may improve their ability to represent multi-word information in other ways. One example is work by Aw and Toneva (2023) that shows that finetuning a language model using a summarization objective can further improve brain alignment.

7 Limitations

We have attempted to address potential limitations in our research design, however, it is important to acknowledge the inherent limitations of our study. Firstly, we use GPT-2 based models to compare with previous work using the same model family. However, analyzing additional language models, such as larger language models, or ones trained with a masked language modeling objective, is an important next step for insights that can be generalized to larger families of models, even if previous work has suggested that larger language models could diverge from human-like representations (Oh and Schuler, 2023). Moreover, in our highly controlled setting, stimulus-tuning larger language models could lead to a smaller increase in next-word prediction and brain alignment, given the relatively small amount of data available for finetuning. This is visible when comparing the stimulus-tuning improvements of GPT-2-distilled and GPT-2-small with GPT-2-medium: as the model size increases, the effect of stimulus tuning and the residual effect after the contrast decreases.

Secondly, our experiments were conducted with one fMRI dataset. Even though the dataset we chose is a well studied dataset, is one of the largest ones available, and care was taken to test the generalization performance to never-before-seen brain data, the effects we observe may still be specific to this dataset. Testing datasets that differ in text genre (we use only a narrative dataset) and language (our conclusions are drawn for English text only) would be particularly interesting.

Thirdly, our findings are based on some experimental choices, such as the scrambling method. For instance, Mollica et al. (2020) showed that fMRI activity in humans reading scrambled sentences remains relatively stable under certain perturbations. Therefore, further investigation into different scrambling methods and their effects could provide additional insights. Furthermore, despite the presence of a strong positive correlation between next-word prediction and brain alignment reported by Schrimpf et al. (2021) and our experiments (0.61 Pearson correlation, see Appendix E), this relationship is not perfectly linear so it is possible that the subtraction that we employ does not perfectly control for the effect of the next-word prediction capabilities.

Fourthly, our results are based on observing changes in NWP on a held-out stimulus test set. However, changes in next-word prediction can be influenced by multiple factors. The model could acquire general English knowledge or knowledge specific to the Harry Potter chapter. Since we are using a language model heavily pretrained on general English knowledge, we believe that the improvement in next-word prediction ability is primarily due to learning about the specific domain we are fine-tuning for. However, investigating further the causes behind the increase in next-word prediction performance could be an insightful next step.

8 Ethics and Broader Impact

Our research impact is closely related to its potential social implications. We propose a method to analyze language models, with the aim of gaining a better understanding of the differences and similarities between the human brain and neural networks. This understanding can serve two key purposes: firstly, shedding light on the reasons behind the impressive power of neural networks; secondly, enhancing our comprehension of the underlying mechanisms governing brain functions. A deeper understanding of both artificial neural networks and biological neural networks can significantly benefit society, especially considering the prevalence of black-box artificial intelligence systems. By leveraging insights from the human brain, we can strive to integrate these systems more consciously into human activities. This integration is essential for ensuring transparency, interpretability, and ethical use of AI, thereby fostering a positive and responsible impact on society.

Acknowledgements

References

Mostafa Abdou, Ana Valeria González, Mariya Toneva, Daniel Hershcovich, and Anders Søgaard. 2021. Does injecting linguistic structure into language models lead to better alignment with brain recordings? *arXiv preprint arXiv:2101.12608*.

Khai Loong Aw and Mariya Toneva. 2023. Training language models to summarize narratives improves brain alignment. In *The Eleventh International Conference on Learning Representations*.

Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. 2009. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, 19(12):2767–2796.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. 2021. Decomposing lexical and compositional syntax and semantics with deep language models. *arXiv preprint arXiv:2103.01620*.

Charlotte Caucheteux and Jean-Rémi King. 2020. Language processing in brains and deep neural networks: computational convergence and its limits. *BioRxiv*.

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

Evelina Fedorenko and Sharon L Thompson-Schill. 2014. Reworking the language network. *Trends in cognitive sciences*, 18(3):120–126.

Angela D Friederici. 2012. The cortical language circuit: from auditory perception to sentence comprehension. *Trends in cognitive sciences*, 16(5):262–268.

Angela D Friederici, Shirley-Ann Rüschemeyer, Anja Hahne, and Christian J Fiebach. 2003. The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. *Cerebral cortex*, 13(2):170–177.

James S Gao, Alexander G Huth, Mark D Lescroart, and Jack L Gallant. 2015. Pycortex: an interactive surface visualizer for fmri. *Frontiers in neuroinformatics*, page 23.

Jon Gauthier and Roger Levy. 2019. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 529–539, Hong Kong, China. Association for Computational Linguistics.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Eghbal A. Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. 2024. Artificial Neural Network Language Models Predict Human Brain Responses to Language Even After a Developmentally Realistic Amount of Training. *Neurobiology of Language*, 5(1):43–63.

Gina F Humphreys, Matthew A Lambon Ralph, and Jon S Simons. 2021. A unifying account of angular gyrus contributions to episodic and semantic cognition. *Trends in Neurosciences*, 44(6):452–463.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

Shailee Jain and Alexander Huth. 2018. Incorporating context into language encoding models for fMRI. In *Advances in Neural Information Processing Systems*, pages 6628–6637.

Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas, and Evelina Fedorenko. 2023. Lexical semantic content, not syntactic structure, is the main contributor to ann-brain similarity of fmri responses in the language network. *bioRxiv*.

Yulia Lerner, Christopher J Honey, Lauren J Silbert, and Uri Hasson. 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915.

Francis Mollica, Matthew Siegelman, Evgeniia Diachek, Steven T. Piantadosi, Zachary Mineroff, Richard Futrell, Hope Kean, Peng Qian, and Evelina Fedorenko. 2020. Composition is the Core Driver of the Language-selective Network. *Neurobiology of Language*, 1(1):104–134.

Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. 2011. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646.

Byung-Doh Oh and William Schuler. 2023. Transformer-based lm surprisal predicts human reading times best with about two billion training tokens. *Preprint*, arXiv:2304.11389.

Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. 2024a. Speech language models lack important brain-relevant semantics. *ACL*.

Subba Reddy Oota, Manish Gupta, and Mariya Toneva. 2024b. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 36.

Lalchand Pandia and Allyson Ettinger. 2021. Sorting through the noise: Testing robustness of information processing in pre-trained language models. *arXiv preprint arXiv:2109.12393*.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying grammatical role, bert doesn't care about word order... except when it matters. *arXiv preprint arXiv:2203.06204*.

Steven Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Siddharth Ramanan, Olivier Piguet, and Muireann Irish. 2018. Rethinking the role of the angular gyrus in remembering the past and imagining the future: the contextual integration model. *The Neuroscientist*, 24(4):342–352.

Aniketh Janardhan Reddy and Leila Wehbe. 2021. Can fmri reveal the representation of syntactic structure in the brain? *Advances in Neural Information Processing Systems*, 34:9843–9856.

J.K. Rowling, M. GrandPre, M. GrandPré, T. Taylor, Arthur A. Levine Books, and Scholastic Inc. 1998. *Harry Potter and the Sorcerer's Stone*. Harry Potter. A.A. Levine Books.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45).

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7329–7346, Online. Association for Computational Linguistics.

Mariya Toneva, Tom M Mitchell, and Leila Wehbe. 2022a. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, 2(11):745–757.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, 32.

Mariya Toneva, Jennifer Williams, Anand Bollu, Christoph Dann, and Leila Wehbe. 2022b. Same cause; different effects in the brain. *Causal Learning and Reasoning*.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014a. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 2014b. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 233–243.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020a. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020b. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

A Prediction head

Similarly to previous work, to predict the fMRI recordings corresponding to a given TR, we use a prediction head that maps from the model representation to the fMRI space. We parameterize this function as a linear function, regularized using the ridge penalty. We train this function in a cross-validated way and test its performance on the data that was heldout during training. We select the ridge parameter via nested cross-validation. For each participant, we train four functions, each one using three of the four runs in the fMRI recordings, and reserve the remaining run for testing. To generate the models representation we average the embeddings corresponding to each fMRI image (i.e., TR) and uses a concatenation of the previous 5 averaged TR embeddings. The averaging is done in order to down-sample the word embeddings (words presented at 0.5 seconds) to the TR rate (2 seconds). The features of the words presented in the previous TRs are included to account for the lag in the hemodynamic response that fMRI records. Because the response measured by fMRI is an indirect consequence of brain activity that peaks about 6 seconds after stimulus onset, predictive methods commonly include preceding time points (Nishimoto et al., 2011; Wehbe et al., 2014a; Huth et al., 2016). This allows for a data-driven estimation of the hemodynamic response functions (HRFs) for each voxel, which is preferable to assuming one because different voxels may exhibit different HRFs.

B Training hyperparameters

To perform the stimulus-tuning, we select training samples that consist of non-overlapping sequences of 80 consecutive words. We train the models with a batch size of 32 and for 2 epochs and we saved checkpoints of the models for each batch. The stimulus text is divided into 4 consecutive sections to enable cross-validation. For each GPT-2-based models we stimulus-tuned four models, one for each held-out run. We train the models using the default hyperparameters provided by Huggingface. As mentioned in Section 2.5 we selected the checkpoint for the stimulus-tuned model that best satisfies $\delta^{stim} \approx \delta^{base}$.

C List of ROI related to language processing and word semantics

These regions include: Middle Frontal Gyrus (MFG), Inferior Frontal Gyrus (IFG), Inferior Frontal Gyrus par Orbitalis (IFGorb), Anterior Temporal Lobe (AntTemp), Posterior Temporal Lobe (PostTemp), Angural Gyrus (AG), Posterior Cingulate Cortex (pCingulate), Dorsomedial Prefrontal Cortex (dmPFC) (see Figure 2B).

D Significance testing and Participants variability

We designed the experiments to compare different models and their perturbations, testing their capabilities in brain alignment and displaying the percentage gain of one model over another. For each comparison, we conducted significance testing using ROI-level Wilcoxon signed-rank tests with $p < 0.05$ and Holm-Bonferroni correction (Holm, 1979). For the final contrast the significance test revealed statistically significant residual alignment in the IFG (pvalue of 0.027). After correcting for multiple comparisons across ROI, this pvalue was no longer statistically significant at a threshold of 0.05. Similarly for the baseline vs stimulus-tuned contrast the results are not significative. Given the high variability across subjects and the number of subjects in our setting, obtaining statistically significant results is challenging. Moreover, due to our controlled experimental design, we only finetuned the baseline model using the stimulus text, that is composed by few samples, which could result in a minimal effect from this perturbation. Still, positive residual alignment in the IFG and AG is observed across three different models: GPT-2-distill, GPT-2-small, and GPT-2-medium. Despite these challenges, we believe that our results are informative and are strengthened by the analysis across three different models.

E Linear relation brain alignment and cross-entropy loss

Schrimpf et al. (2021) shows a linear relationship between brain alignment and the next-word prediction capability of language models. Specifically, to test this relationship, they used Pearson correlation (normalized using estimated noise ceiling) for brain alignment and the exponentiated cross-entropy, i.e.,

perplexity, to evaluate next-word prediction capability. However, language models are typically fine-tuned using cross-entropy loss, and our experiments during fine-tuning showed that it is not feasible to achieve a similar difference in perplexity between (stimulus-tuned and stimulus-tuned scrambled) and (baseline and baseline scrambled) models. Therefore, in our final contrast, we employed the difference in cross-entropy between models. Although Schrimpf et al. (2021) suggested a linear relationship between perplexity and brain alignment, this does not guarantee a linear relationship between cross-entropy and brain alignment, even if the two metrics are related. For this reason, we tested whether the linear relationship also holds for cross-entropy loss. We demonstrate in Figure 6 that there is indeed a linear relationship between cross-entropy and brain alignment (0.61), albeit slightly lower than the one between brain alignment and perplexity (0.67). More experiments with additional models and datasets are necessary to verify if this relation holds in general, but for our analysis, it is sufficient to apply the subtraction explained in Section 2.5.
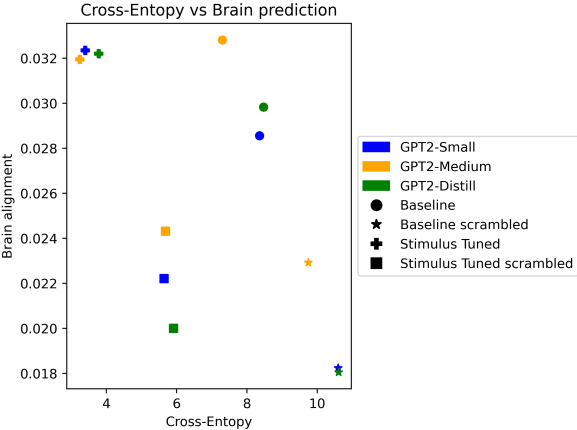
Figure 6: Correlation between the brain predictions capability of GPT-2-small, GPT-2-medium and GPT-2-distilled (on a held-out test set) and their cross-entropy loss. The correlation between these two measures is 0.61, similar to the correlation between the brain predictions capability and the perplexity 0.67.

F   Cross perturbation contrast illustration
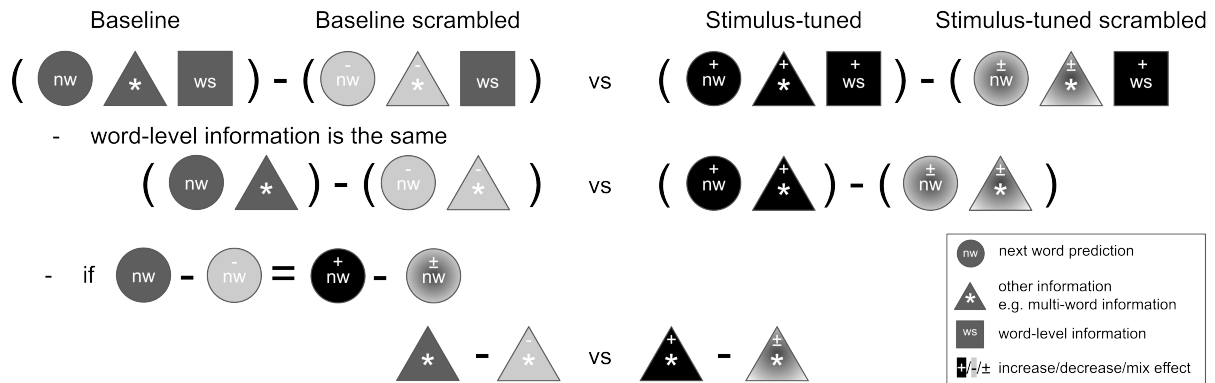
Figure 7: Illustration of the contrast (baseline - baseline scrambled) vs (stimulus-tuned - stimulus-tuned scrambled). Any observed effect of this contrast is controlled for word-level information information. Additionally, if the next-word prediction differences are equal then the contrast control also for next-word prediction. Therefore any observed effect would then be due to more than next-word prediction and word-level information information. An example of other source of information is the multi-word information.

G   Noise Ceiling estimation

The noise ceiling estimation is employed to assess the signal quality of fMRI data. fMRI data are inherently noisy, and the noise ceiling estimation provides an estimate of the variance that can be explained by an ideal data-generating model. The method relies on predicting the fMRI activity of a target participant using linear models trained on data from another participant. Linear encoding models are utilized. For a more detailed explanation, refer to (Schrimpf et al., 2021). We employed this approach because our method relies on contrasts between different models aimed at predicting the brain activity of each subject, necessitating a consistent set of voxels for each encoding model.

18444
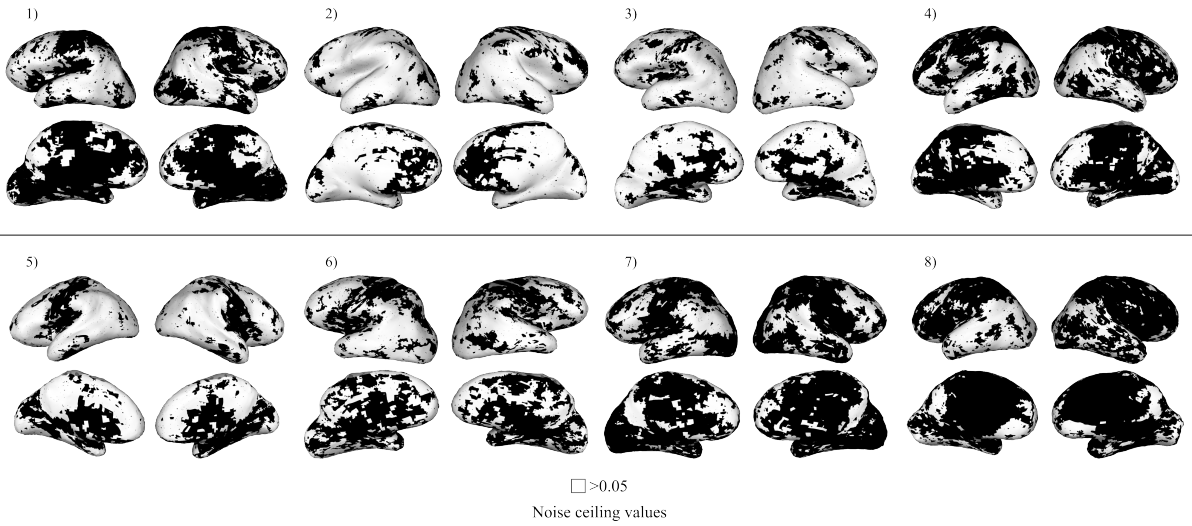
Figure 8: Voxel-wise estimated noise ceiling values. To exclude noisy voxels, we selected, for each participant, those with noise ceiling estimates above 0.05.

H GPT-2-small
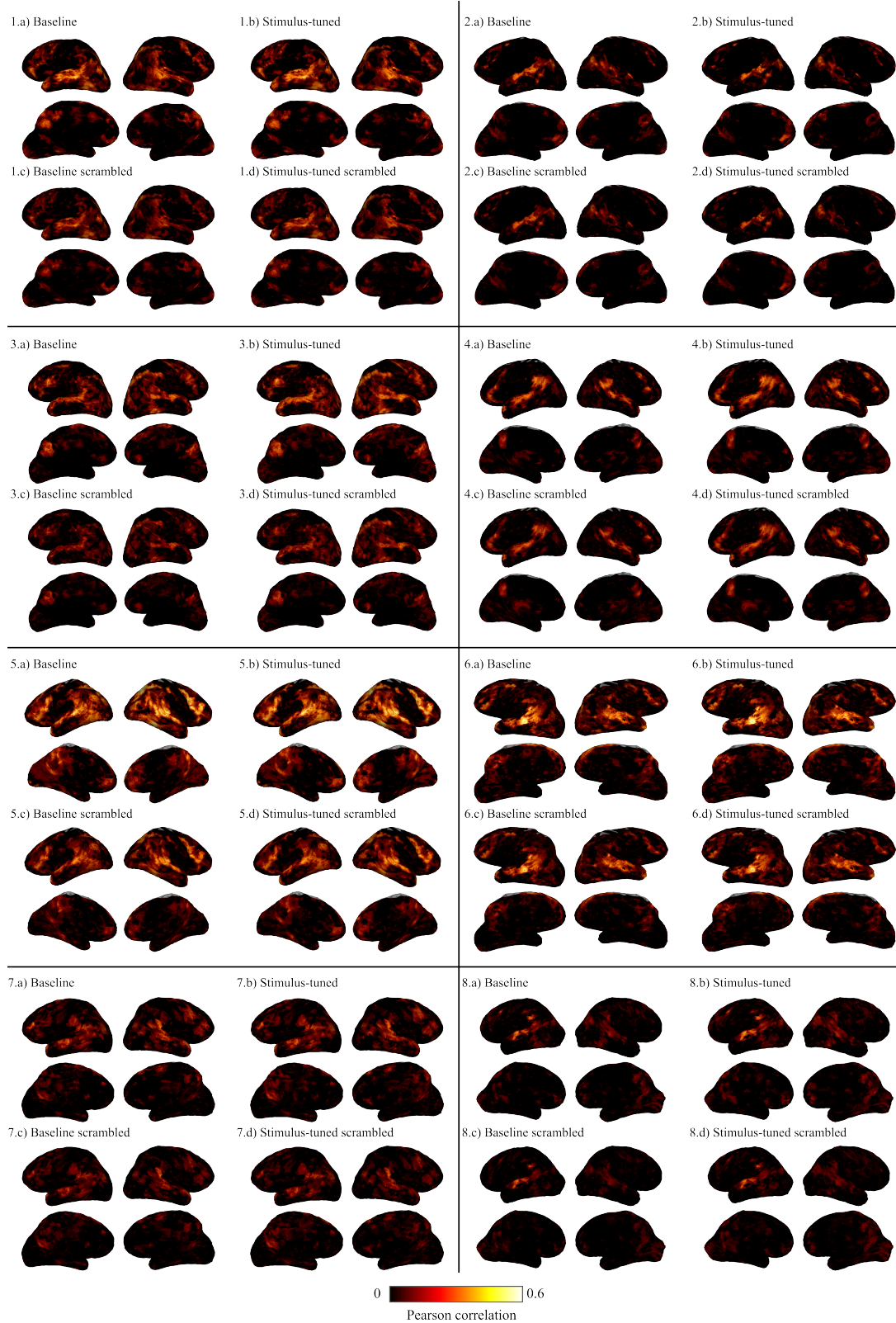
Figure 9: Performances of the GPT-2-small baseline and perturbed models of all participants at the brain alignment task. Stimulus-tuning improves the brain alignment (stimulus-tuned in (.b) vs baseline in (.a)) for almost all participants. In contrast, scrambling reduces the brain alignment (baseline in (.a) vs baseline scrambled in (.c)). Despite the reduction in alignment due to the scrambling perturbation, all four models (.a,.b,.c,.d) exhibit alignment in language processing regions.
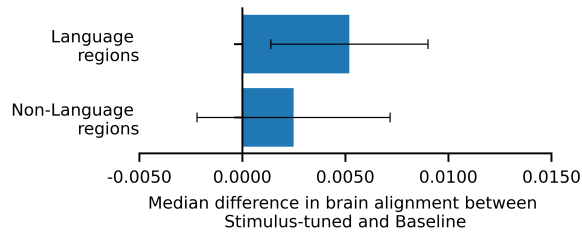
Figure 10: Median difference in brain alignment between GPT-2-small stimulus-tuned and baseline models. We display the median difference in language and non-language regions and the median absolute deviation across the 8 participants.
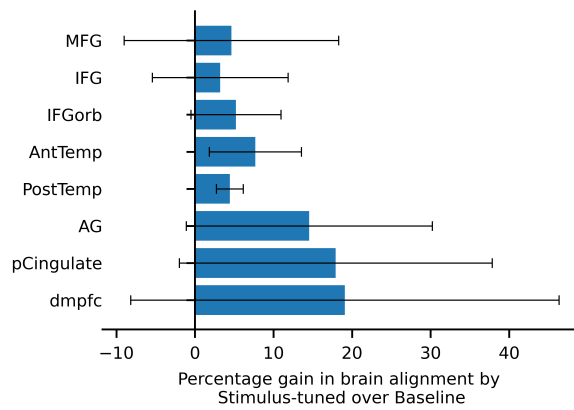
Figure 11: Impact of the stimulus-tuning perturbation on the brain alignment of the GPT-2-small baseline model. We show the median percentage gain as well as the median absolute deviation across participants. We include only voxels with estimated noise ceiling values >0.05.
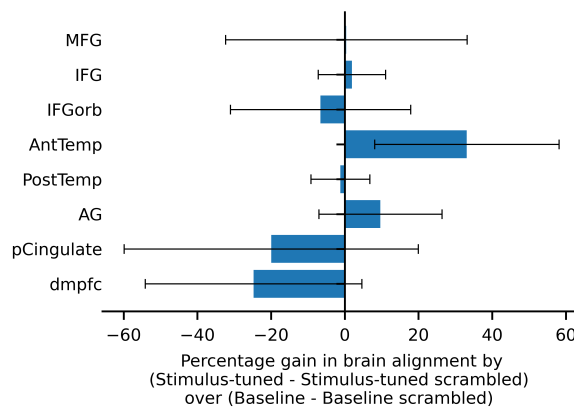
Figure 12: Impact of the scrambling perturbation on the stimulus-tuned model versus its impact on the baseline model for GPT-2-small model. We show the median percentage gain as well as the median absolute deviation across participants by (stimulus-tuned - stimulus-tuned scrambled) over (baseline - baseline scrambled) in language regions. We include only voxels with estimated noise ceiling values >0.05.

-20% [color bar] 20%

Percentage change
(Stimulus-tuned - Stimulus-tuned scrambled)
vs (Baseline - Baseline scrambled)
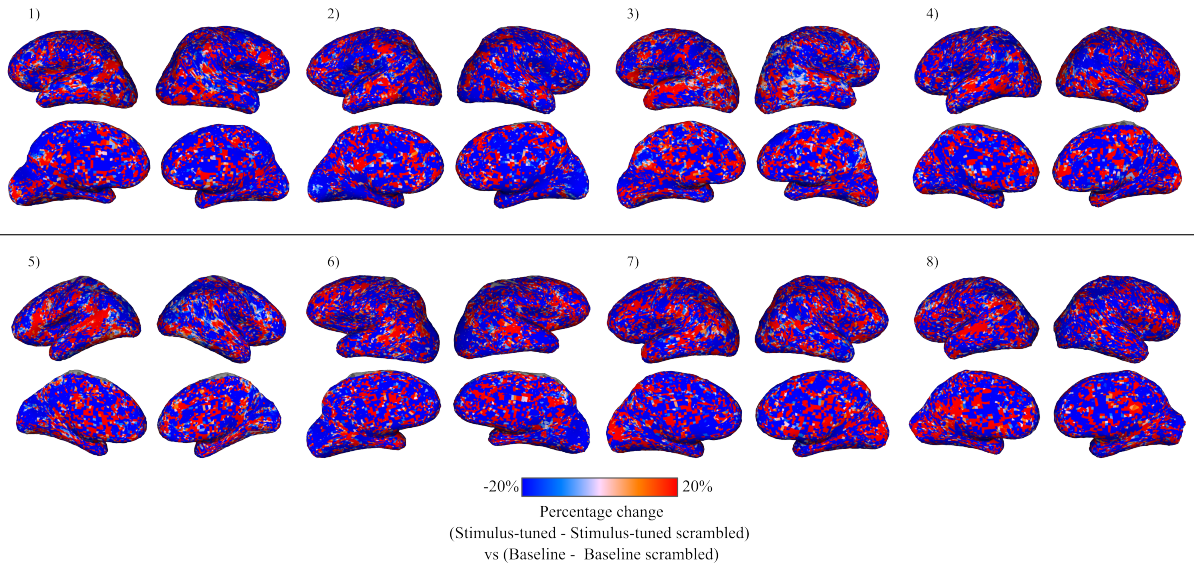
Figure 13: Voxel-wise brain alignment for each participant from contrast that controls for the effect of next-word prediction and word-level information on brain alignment: (stimulus-tuned - stimulus-tuned scrambled) vs (baseline - baseline scrambled). Voxels that appear in red are better predicted by the stimulus-tuned model, even when accounting for next-word prediction and word-level information. Voxels that appear in blue are better predicted by the baseline model. Despite some variation across participants, several language regions appear in red. We quantify these observations in Figure 12.

I GPT-2-Distilled
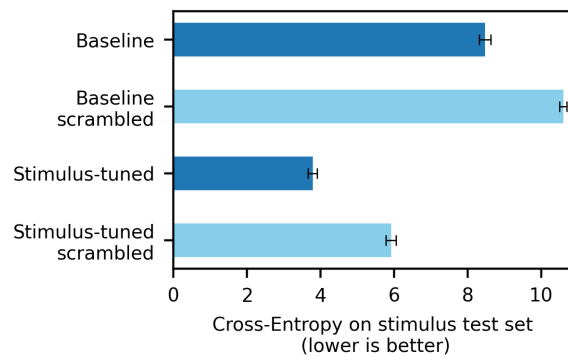
Figure 14: Performances of the GPT-2-distilled baseline and perturbed models at next-word prediction averaged across runs with standard deviation.

18448
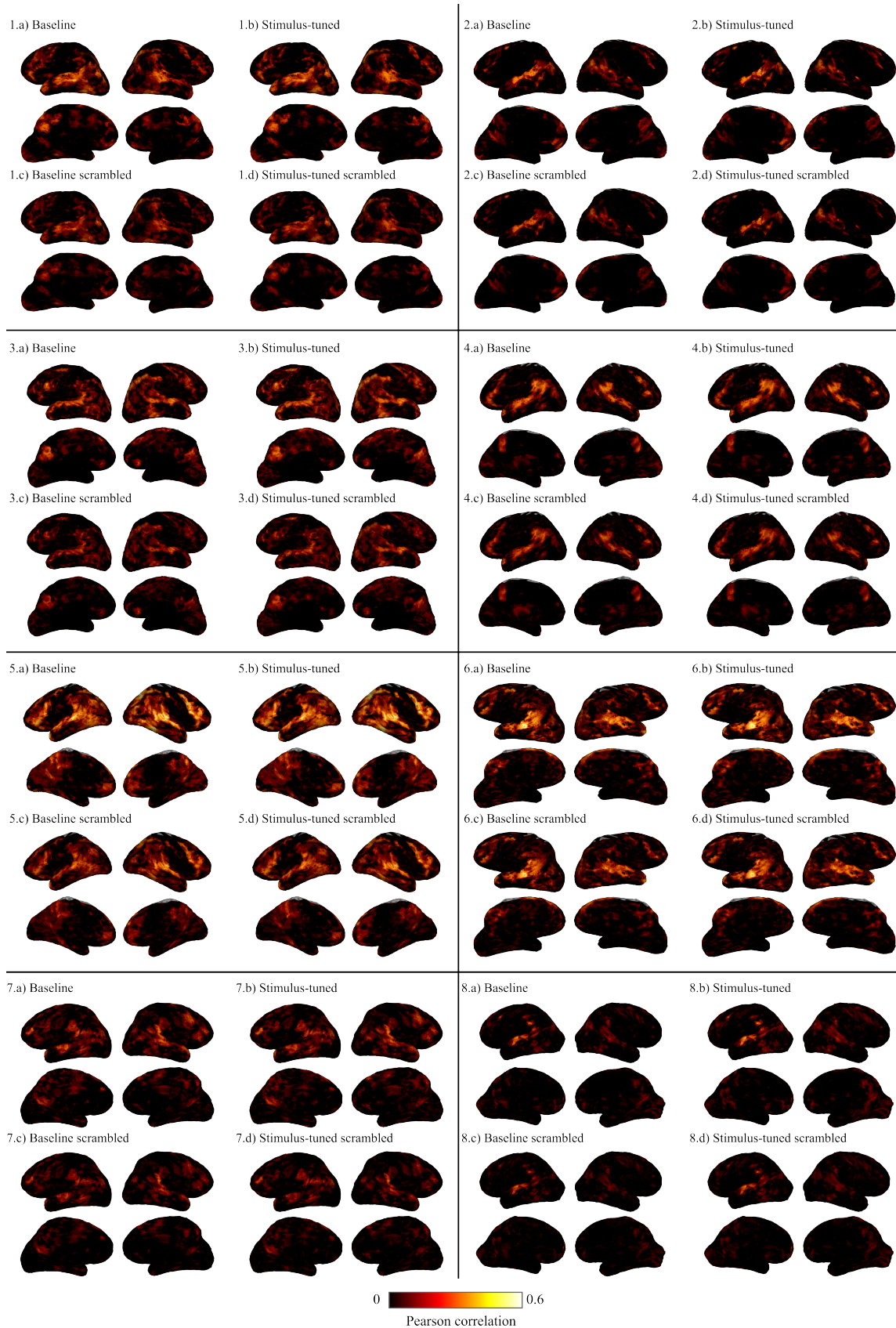
Figure 15: Performances of the GPT-2-distilled baseline and perturbed models of all participants at the brain alignment task. Stimulus-tuning improves the brain alignment (stimulus-tuned in (.b) vs baseline in (.a)) for almost all participants. In contrast, scrambling reduces the brain alignment (baseline in (.a) vs baseline scrambled in (.c)). Despite the reduction in alignment due to the scrambling perturbation, all four models (.a,.b,.c,.d) exhibit alignment in language processing regions.
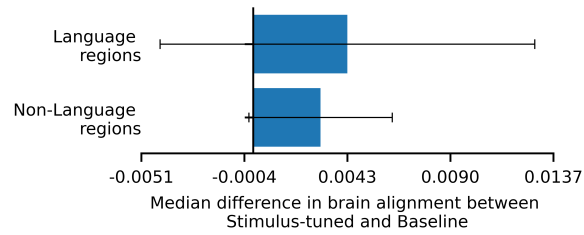
Figure 16: Median difference in brain alignment between GPT-2-distilled stimulus-tuned and baseline models. We display the median difference in language and non-language regions and the median absolute deviation across the 8 participants.
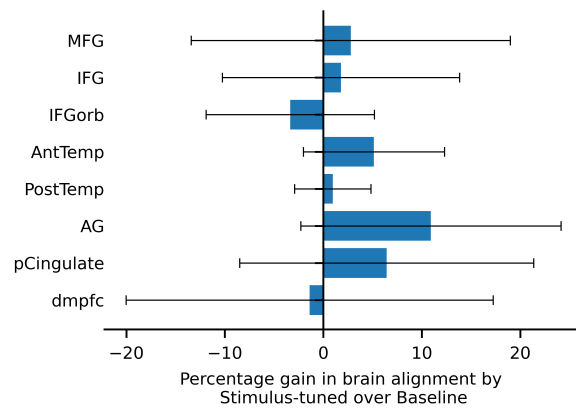
Figure 17: Impact of the stimulus-tuning perturbation on the brain alignment of the GPT-2-distilled baseline model. We show the median percentage gain as well as the median absolute deviation across participants. We include only voxels with estimated noise ceiling values >0.05.
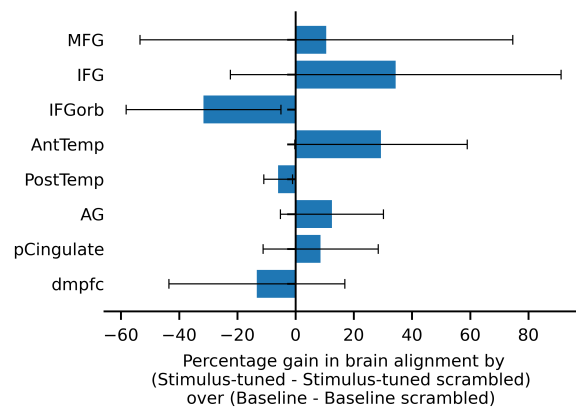
Figure 18: Impact of the scrambling perturbation on the stimulus-tuned model versus its impact on the baseline model for GPT-2-distilled model. We show the median percentage gain by (stimulus-tuned - stimulus-tuned scrambled) over (baseline - baseline scrambled) as well as the median absolute deviation across participants in language regions. We include only voxels with estimated noise ceiling values >0.05.

-20% ■■■■■■■■■■ 20%

Percentage change
(Stimulus-tuned - Stimulus-tuned scrambled)
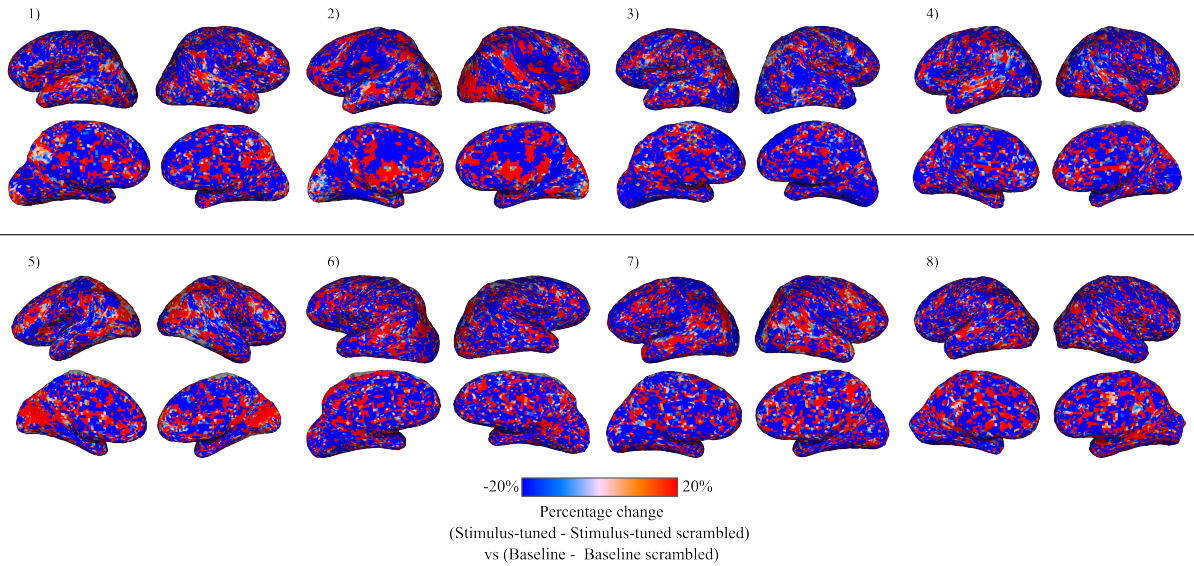vs (Baseline - Baseline scrambled)

Figure 19: Voxel-wise brain alignment for each participant from contrast that controls for the effect of next-word prediction and word-level information on brain alignment: (stimulus-tuned - stimulus-tuned scrambled) vs (baseline - baseline scrambled) for GPT-2-distill. Voxels that appear in red are better predicted by the stimulus-tuned model, even when accounting for next-word prediction and word-level information. Voxels that appear in blue are better predicted by the baseline model. Despite some variation across participants, several language regions appear in red. We quantify these observations in Figure 18.
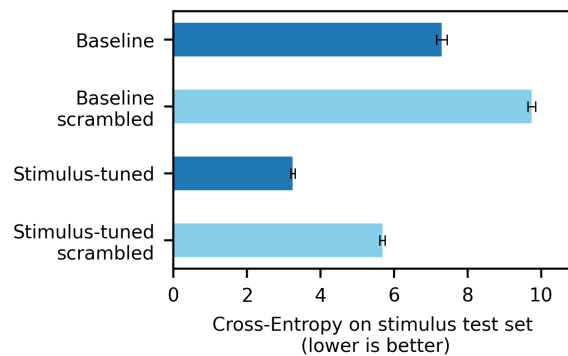
J   GPT-2-Medium

Figure 20: Performances of the GPT-2-medium baseline and perturbed models at next-word prediction averaged across runs with standard deviation.
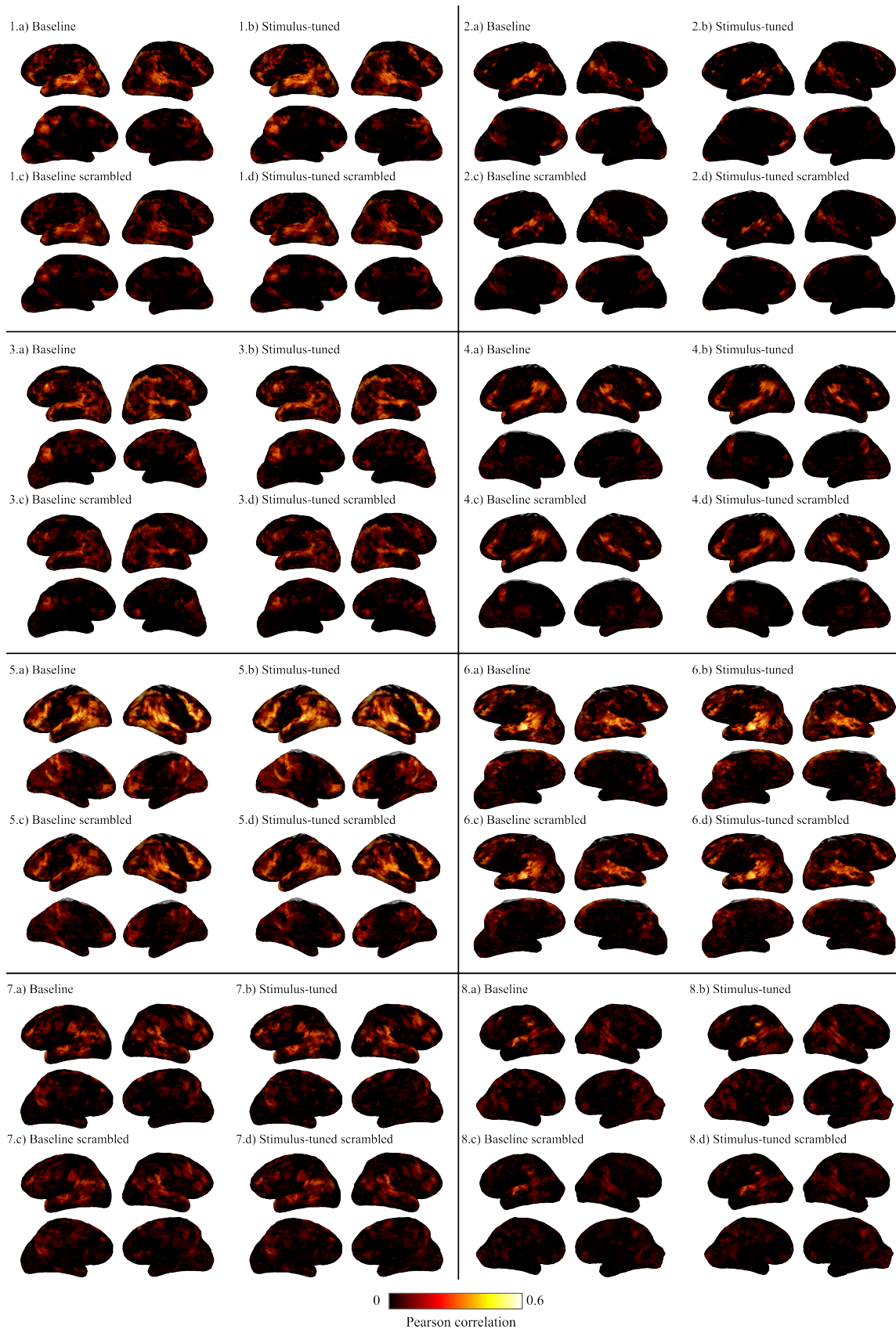
Figure 21: Performances of the GPT-2-medium baseline and perturbed models of all participants at the brain alignment task. Stimulus-tuning improves the brain alignment (stimulus-tuned in (.b) vs baseline in (.a)) for almost all participants. In contrast, scrambling reduces the brain alignment (baseline in (.a) vs baseline scrambled in (.c)). Despite the reduction in alignment due to the scrambling perturbation, all four models (.a,.b,.c,.d) exhibit significant alignment in language processing regions.
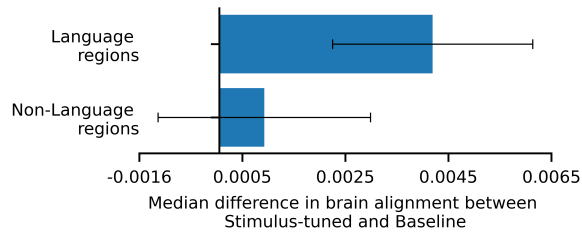
18452

Figure 22: Median difference in brain alignment between GPT-2-median stimulus-tuned and baseline models. We display the median difference in language and non-language regions and the median absolute deviation across the 8 participants.
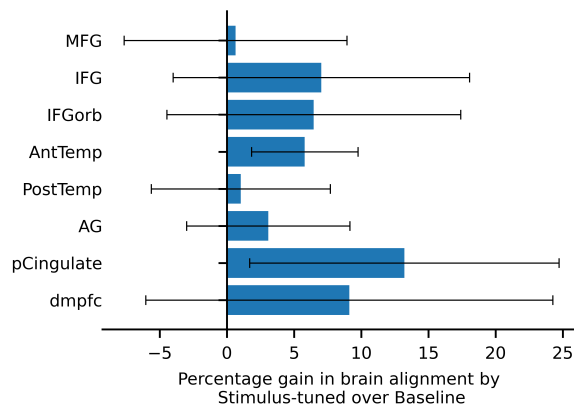

Figure 23: Impact of the stimulus-tuning perturbation on the brain alignment of the GPT-2-medium baseline model. We show the median percentage gain as well as the median absolute deviation across participants. We include only voxels with estimated noise ceiling values >0.05.
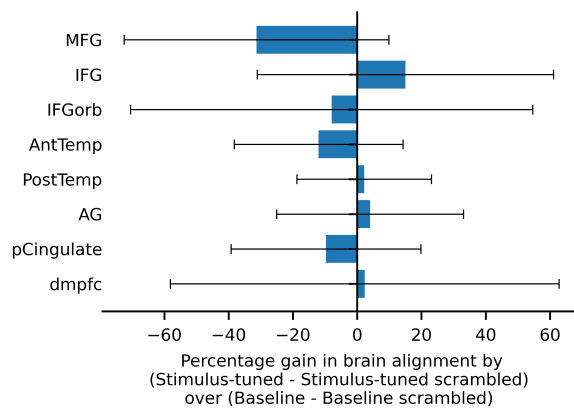

Figure 24: Impact of the scrambling perturbation on the stimulus-tuned model versus its impact on the baseline model for GPT-2-medium model. We show the median percentage gain by (stimulus-tuned - stimulus-tuned scrambled) over (baseline - baseline scrambled) as well as the median absolute deviation across participants in language regions. We include only voxels with estimated noise ceiling values >0.05.

-20% 20%

Percentage change
(Stimulus-tuned - Stimulus-tuned scrambled)
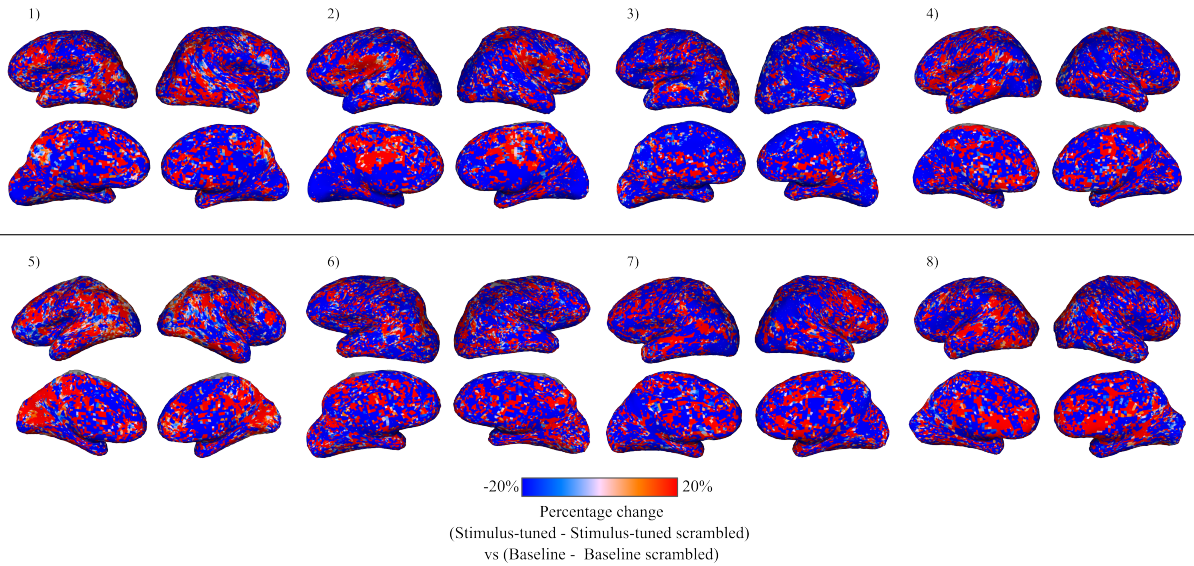vs (Baseline - Baseline scrambled)

Figure 25: Voxel-wise brain alignment for each participant from contrast that controls for the effect of next-word prediction and word-level information on brain alignment: (stimulus-tuned - stimulus-tuned scrambled) vs (baseline - baseline scrambled) for GPT-2-medium. Voxels that appear in red are better predicted by the stimulus-tuned model, even when accounting for next-word prediction and word-level information. Voxels that appear in blue are better predicted by the baseline model. Despite some variation across participants, several language regions appear in red. We quantify these observations in Figure 24.