On whether the relationship between large language models and brain activity is language-specific

Sertug Gürel (sertug.guerel@uni-potsdam.de)
Alessandro Lopopolo (lopopolo@uni-potsdam.de)
Milena Rabovsky (milena.rabovsky@uni-potsdam.de)
Department of Psychology, University of Potsdam, Karl-Liebknecht-Str. 24–25, 14476 Potsdam, Germany

Abstract

Using large language models (LLMs), such as GPT-2, to study language processing in both machines and humans has become increasingly prevalent. Existing literature demonstrates that these models are strong predictors of human brain activity (Schrimpf et al., 2021), which has been taken to indicate that LLMs are good models for language processing in the human brain. The current study aimed to assess whether these models' predictive performance of brain activity is specific to brain regions involved in language processing and whether or not the prediction of functionally different brain regions relies on different features of the LLMs' hidden layers. Our results suggest that LLMs' ability to predict brain activation does not strongly differ between language and non-language-related brain areas. The set of features that drive prediction performance across areas is not entirely the same, but there is a considerable correlation between the features that language-related and non-language-related regions rely on for brain predictions. Hence, we suggest that more research is needed to understand the nature of the information that drives brain predictions in LLMs.

Keywords: neural encoding; LLMs; GPT-2; feature analysis; language network

Introduction

In recent years, large language models (LLMs) have gained significant prominence. These models demonstrate remarkable natural language processing (NLP) skills while also effectively predicting brain activity associated with language processing in humans. Their success in predicting brain activity has been taken to indicate that they provide good models of language processing in the human brain (Caucheteux et al., 2022; Schrimpf et al., 2021; Toneva & Wehbe, 2019). However, the reasons behind this high brain predictivity are still not fully understood.

In this study, we investigate whether the models' ability to predict brain activity in humans is limited to language-related regions, and whether the predictions in functionally distinct brain regions are influenced by the same or different features in LLMs' internal representations.

Methods

Since our focus is on investigating whether prediction success depends on brain activation related to language processing, we chose to examine GPT-2 (Radford et al., 2019), one of the most successful models in prior research (Schrimpf et al., 2021), instead of comparing multiple models. For the model's predictions, we turn to the fMRI dataset from Pereira et al. (2018), a dataset that has been central in previous studies exploring LLMs' ability to predict brain activity in the language network.

Ridge Regression Analysis

We used the GPT-2 model's hidden embeddings to predict the voxel activations of the Pereira et al. (2018) dataset using Ridge Regression with k-fold cross validation (k=5). The alpha parameter of the Ridge Regression model is tuned for each voxel separately during the training. Then we correlated the predicted voxel activations with real activation from the dataset. The brain-area-wise results are obtained by averaging the voxel correlations within each ROI.

SHapley Additive exPlanations (SHAP)

To determine how much the same features from the hidden layers of the GPT-2 model contributed to the predictions of language-related and non-language-related brain regions, we employed the SHapley Additive exPlanations (SHAP) analysis (Lundberg and Lee, 2017). SHAP analysis evaluates each feature's influence by taking into account every potential feature combination and how each feature contributes to the prediction.

Results

Our results showed that the embeddings of the 11th layer of the GPT-2 model provided the highest predictive performance. Therefore, we focus on this layer while presenting the results. Although the predictive performance for most of the language-related areas seems to be higher than for most highly correlated non-language-related regions, the differences are small. Some non-language areas seem to be predicted as well as (and partly even better than) some language-related regions. This leads us to consider whether these predictions are driven by shared or distinct features within the hidden layers.

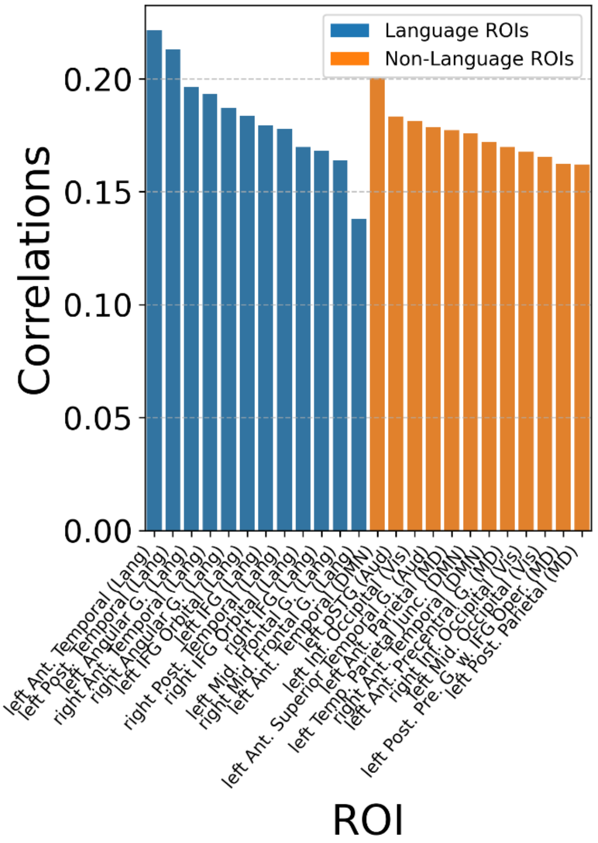

Figure 1: Correlation of predictions with observed activations of both language-related and non-language-related areas (only for the 11th layer, which showed overall the best performance).

It is crucial to answer this question because if the same features influence the predictivity of language-related and non-language-related processes in the brain, it may raise the question of whether the information that drives brain predictivity in LLMs is related to language. To address this issue, we calculated the SHAP values for each feature in GPT-2's 11th hidden layer, averaged the values for each region of interest (ROI), and then correlated language-related and non-language-related areas with each other. We divided the correlations into three categories—SHAP value correlations between language-related regions, between non-language-related regions, and between language-related and non-language-related regions—to increase clarity.

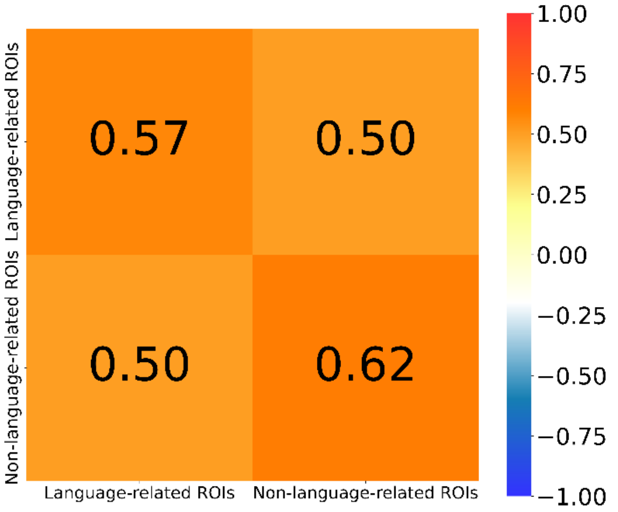

Figure 2: Average correlation of both language-related and non-language-related areas' SHAP values of the 11th layer.

The results suggest that, on average, SHAP values of language-related regions exhibit stronger correlations with each other than those of non-language-related regions. However, it is important to note that correlations are not drastically different. This finding may suggest that brain prediction by LLMs is at least somewhat influenced by information that is not related to language, and/or that the brain has a shared base for information processing, even across functionally separate areas.

Conclusion

In this work, we investigated how functionally distinct brain activations are predicted by an LLM model and whether the same or different features of the models' hidden layers drive the predictions. Results suggest that although not exactly the same features drive the GPT-2's brain prediction, there seems to be a common basis for the prediction of language-related and non-language-related brain areas. This suggests that more research is required to understand the nature of the information that drives brain predictions in LLMs.

References

Caucheteux, C., Gramfort, A., & King, J.-R. (2022). Deep language algorithms predict semantic comprehension from brain activity. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-20460-9

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting ModelPredictions. In 31st Conference on Neural Information Processing Systems (NIPS2017). Long Beach, CA.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. Nature Communications, 9(1). https://doi.org/10.1038/s41467-018-03068-4

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The Neural Architecture of language: Integrative modeling converges on Predictive Processing. Proceedings of the National Academy of Sciences, 118(45). https://doi.org/10.1073/pnas.2105646118

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Vancouver, Canada.