Dynamics of Human-LLM Alignment

Aleksandra Bakalova
Saarland University
alba00013@stud.uni-saarlnd.de

Abstract

In this work we evaluated the dynamics of the emergence of the alignment between word representations obtained from a language model and the human brain. We tracked the emergence of the human-model alignment by examining several model checkpoints saved throughout training and evaluated the role of eight representative linguistic properties in this process.
All the code is available on github [1]

1 Introduction and related work

There are several recent works [8, 6, 1] that explore the correlations between the representations of language units taken from the neural language model and the human brain. Surprisingly, these papers show that the neural representations appear to be highly predictive of the human brain activity [6]. In this work, we aim to further examine the human-model alignment. While most of the previous works in this domain focused on evaluating the performance of the model's final training checkpoint, we examine the emergence of alignment during model training by evaluating several checkpoints saved at different training steps. In order to find the possible correlates of the emergence of alignment, we also evaluate the role of eight representative linguistic properties in this process.

In the closest work to ours, the authors evaluated the performance of GPT-2 in predicting language-responsive voxels' activations in the Pereira2018 fMRI benchmark [3]. They found that the Pearson correlation between model and brain representations increases throughout the training process, and slightly decreases when training for more than 100% of training steps. The perplexity of GPT-2 follows the same pattern: it increases throughout the training process and decreases if training for more than 100% of training steps.

In the other related work [1] the authors evaluate the performance in brain encoding of the models from the OPT family depending on the size of the model. Their results show that the more parameters the model has, the better it predicts brain activity.

In [4] the authors found that removing some of the linguistic properties from the model representations significantly decreases the correlation between model and brain representations of the stimuli. They claim that those linguistic properties are important for accurately encoding brain activations. In this work we build on the methodology proposed in [4] and explore the impact of each of the linguistic properties on brain alignment scores throughout the training.

2 Model and dataset

2.1 Language model

We chose the 160M-parameter model from Pythia family [2], as this model has many available checkpoints saved throughout training and was designed specifically for analysing language models,

[1] https://github.com/CUPalex/alignment

Bridging Language in Machines and Language in the Brain.

which may be beneficial for future research. It also resembles GPT-2 model in terms of the number of parameters, which may help in comparing the results of this work with others.

We choose to take the checkpoints on steps 0, 1000, 15000, 43000, 72000, 100000, 143000 which roughly indicate 0%, 1%, 10%, 30%, 50%, 70% and 100% of training.

We take measurements from every second layer of the model due to the resource constraints.

The representations for each word are taken as the representations of the last token of that word. Each word has the context of 50 previous tokens. We do not consider the words for which the available context is smaller than 50 tokens.

We include the perplexity measures of the model on our dataset in figure 2.1. The best-performing checkpoint resembles 70% of training steps, which slightly overperforms the checkpoints resembling 50% and 100% of training.

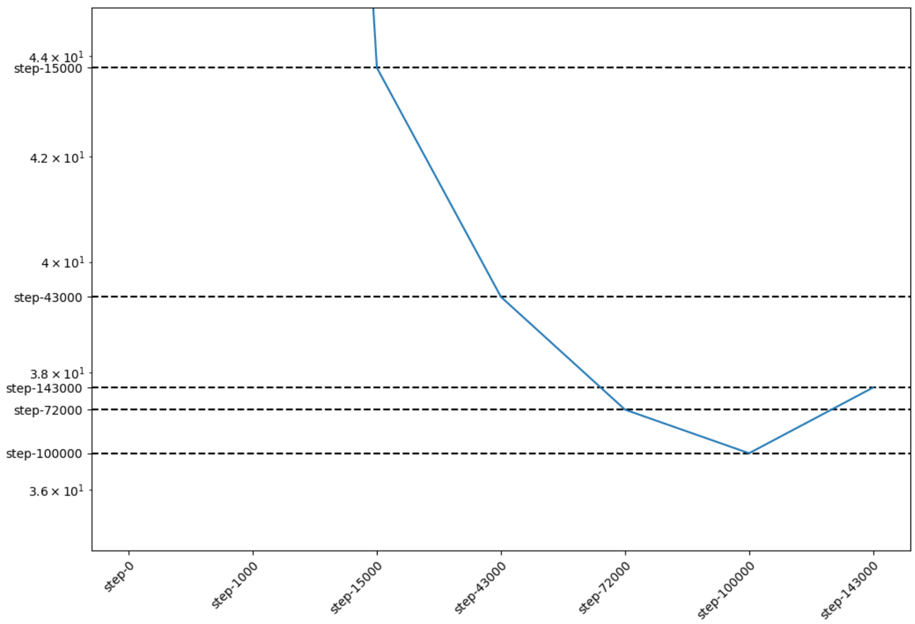

Figure 1: Perplexity of the model measured throughout training on the Harry Potter dataset with the same context window as used for retrieving representations.

As a baseline, we also add a random representation, which does not depend on the input data and is a collection of random normally distributed values of the same shape as the layer activations.

2.2 Dataset

We use the publically available dataset [9] which consists of fMRI recordings of 8 participants reading chapter 9 of the book Harry Potter and the Sorcerer's Stone. This dataset is one of the largest publicly available datasets in terms of samples per participant, which improves the reliability of the brain alignment estimate. We take the already preprocessed dataset made available by [8].

Due to the resource constraints, we take the data only of three participants, who are denoted as F, M and H.

3 Methodology

Overall, we take several checkpoints of the model and measure the performance of every second layer of each checkpoint in encoding brain activations. Apart from that, to get a more fine-grained view of the dynamics of brain alignment, we choose several linguistic properties and measure how much each of them affects the encoding performance. We do this also for every second layer of each checkpoint.

3.1 Linguistic properties

3.1.1 Choice of linguistic tasks

In terms of the linguistic properties that we use to get more insight into the process of brain alignment improvement, we follow [4] and choose the following linguistic tasks: *Sentence Length* (length of the sentence in words), *TreeDepth* (depth of syntactic tree), *TopConstituents* (sequence of top-level constituents of the syntactic constituent tree), *Tense* (tense of the verb in main clause), *Subject Number* (number of the subject in main clause), *Object Number* (number of the object in main clause). We use Stanford core-NLP stanza [5] library to get annotations for our dataset. Each word of the dataset is annotated with the same label as the sentence to which it belongs.

Moreover, as we compute the alignment per-word, we also find it beneficial to evaluate the impact of per-word linguistic tasks. Thus, we use two more tasks, *POS* (part-of-speech tags) and *Smallest Phrase Constituent* (the smallest phrase constituent above a word), introduced by [7].

As a baseline, we add a random binary classification task. In this task we randomly assign each word with one of two classes.

3.1.2 Removal of properties

To measure the effect of each property on brain alignment scores we use a ridge regression method following [4]. For each of the linguistic tasks we train a regression with a loss function $\min_\theta \|\mathbf{W} - \mathbf{T}\theta\|_F^2 + \lambda\|\theta\|_F^2$, where $\mathbf{W} \in \mathcal{R}^{N\times \text{embedding\_dim}}$ is the matrix with word features for N words obtained from the model and $\mathbf{T} \in \mathcal{R}^{(N+1)\times 1}$ are the labels for the given linguistic task corresponding to the same words stacked together with the intercept. Then we can get the residual word representations by removing the specific linguistic property: $r(\mathbf{T}) = \mathbf{W} - \mathbf{T}\theta$. We train a separate regression for every layer, every checkpoint and every linguistic property. We choose the λ parameter by 4-fold cross-validation from $\lambda \in \{10^k | k \in [-5 \ldots 5], k \in \mathbb{Z}\}$.

3.1.3 Balancing classes

The linguistic tasks we choose are classification tasks, and classes in these tasks have different number of words that belong to the class. We balance the distribution by choosing data-specific division on classes. The distribution of classes for sentence-level tasks used in this work can be seen in the figure 3.1.3, while the original distribution is reported in the figure A. We choose the thresholds to distribute the data between classes more evenly and view the *Sentence Length* task as 3-class classification on sentences of length from 1 to 7, from 8 to 16, and more than 16, and the *Tree Depth* task as the 2-class classification task on the sentences with the syntactic tree depth less or equal than 3 and more than 3. We make the *Top Constituents* task a binary classification task on whether the sequence of top constituents of the sentence is the most popular (NP, VP) or not. The *Tense* task is the 2-way classification task on whether the verb is in past tense or not, the *Subject Number* is the 2-way classification task on whether the subject is singular or not, and the *Object Number* is a 2-way classification task on whether the object is present in the sentence or not.

For the word-level tasks we perform the same analysis. The distribution and classes used in this work can be found in the figure 3.1.3, while the original distribution is in the figure A.

3.1.4 Task similarities

The linguistic properties we chose can be correlated, and thus the results we get for one task may be affected by other tasks as well. However, we observe that most all the tasks, except for *Tree Depth* and *Sentence Length*, have relatively low correlations. We report the Pearson correlations between class labels in table A.

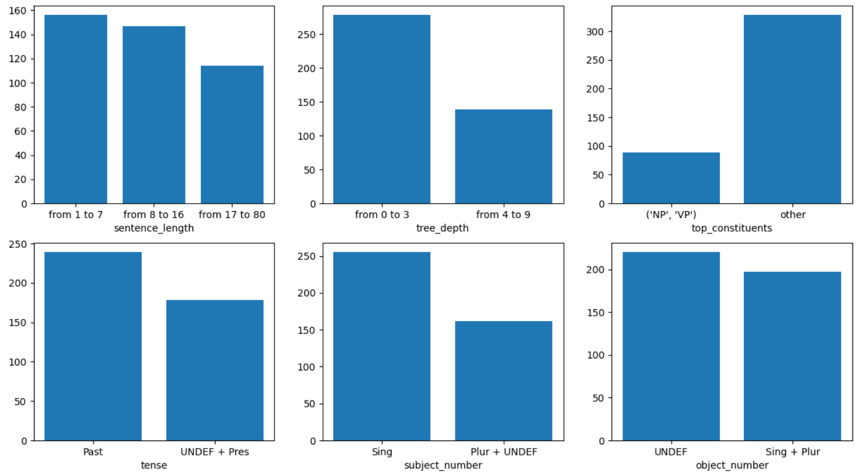

Figure 2: Distribution of classes for sentence-level linguistic tasks used in this work. On y axis the number of sentences which have the corresponding property is reported. UNDEF states for the sentences where the value of the property is unclear or not applicable. For example, Object Number task is not applicable to the sentences with no direct object.

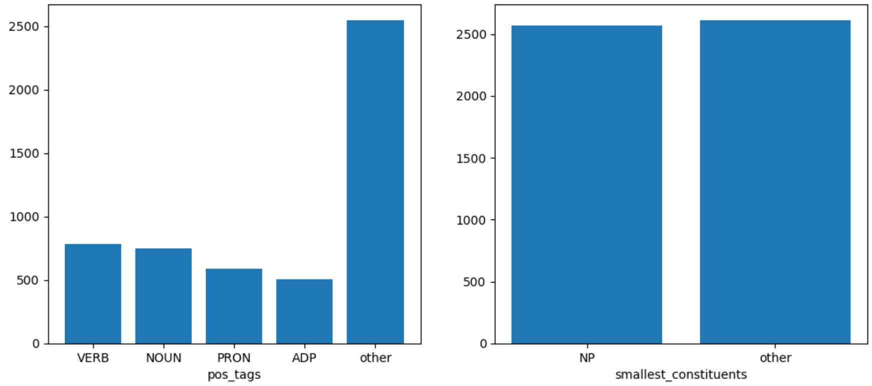

Figure 3: Distribution of classes for word-level linguistic tasks used in this work. On y axis the number of words which have the corresponding property is reported.

3.1.5 Success of property removal

To ensure that the removal of the property is successful we report the probing tasks accuracies on retrieving class labels from model representations before and after feature removal in table 3.1.5. We use logistic regression with l^2 regularization as probing classifier and choose regularization parameter using 4-fold cross-validation from values $C \in \{10^k | k \in [-3 \ldots 1], k \in \mathbb{Z}\}$. Not all of the linguistic properties were successfully removed with this method, thus we work only with the tasks that were removed correctly, namely *Tree Depth, Tense* and *Smallest Constituents*. We say that the task was removed successfully if the probing performance before removal is higher than chance, and is lower than chance after removal for all layers, model checkpoints and participants.

		Random	SentenceLength	TreeDepth	TopConst	Tense
last step	layer-0	0.52 | 0.52	0.55 | 0.55	**0.63 | 0.59**	0.79 | 0.79	**0.69 | 0.67**
	layer-2	0.49 | 0.49	0.59 | 0.54	**0.68 | 0.59**	0.79 | 0.79	**0.72 | 0.67**
	layer-4	**0.51 | 0.49**	0.60 | 0.52	**0.68 | 0.57**	0.79 | 0.79	**0.71 | 0.67**
	layer-6	0.49 | 0.50	0.59 | 0.50	**0.66 | 0.52**	0.79 | 0.79	**0.72 | 0.67**
	layer-8	0.50 | 0.50	0.58 | 0.45	**0.67 | 0.45**	0.79 | 0.79	**0.71 | 0.67**
	layer-10	0.50 | 0.50	0.58 | 0.41	**0.67 | 0.38**	0.79 | 0.79	**0.71 | 0.65**
	layer-12	0.50 | 0.50	0.57 | 0.35	**0.64 | 0.30**	0.78 | 0.78	**0.70 | 0.54**
last layer	random	**0.52 | 0.49**	0.50 | 0.50	0.54 | 0.54	0.79 | 0.79	0.65 | 0.65
	step-0	**0.50 | 0.50**	0.54 | 0.44	**0.62 | 0.50**	0.78 | 0.78	0.67 | 0.65
	step-1000	0.49 | 0.48	0.59 | 0.41	**0.67 | 0.33**	0.78 | 0.79	**0.73 | 0.64**
	step-15000	**0.52 | 0.49**	**0.57 | 0.32**	**0.66 | 0.27**	0.79 | 0.79	**0.70 | 0.52**
	step-43000	0.50 | 0.49	**0.56 | 0.31**	**0.67 | 0.38**	0.78 | 0.78	**0.68 | 0.49**
	step-72000	0.50 | 0.50	**0.56 | 0.28**	**0.66 | 0.35**	0.79 | 0.78	**0.68 | 0.46**
	step-100000	0.51 | 0.50	**0.56 | 0.29**	**0.65 | 0.27**	0.78 | 0.78	**0.68 | 0.46**
	step-143000	0.50 | 0.50	0.57 | 0.35	**0.64 | 0.30**	0.78 | 0.78	**0.70 | 0.54**
	chance	0.50	0.32	0.59	0.79	0.67
		Subject Number	Object Number	POS	Smallest Constituents	
last step	layer-0	**0.71 | 0.71**	0.62 | 0.62	0.84 | 0.77	**0.86 | 0.49**	
	layer-2	0.70 | 0.71	**0.63 | 0.62**	0.90 | 0.92	**0.88 | 0.36**	
	layer-4	0.71 | 0.71	**0.64 | 0.62**	0.90 | 0.92	**0.89 | 0.33**	
	layer-6	**0.72 | 0.71**	**0.64 | 0.62**	0.89 | 0.92	**0.89 | 0.33**	
	layer-8	**0.72 | 0.71**	0.62 | 0.61	0.88 | 0.90	**0.88 | 0.32**	
	layer-10	**0.73 | 0.70**	0.62 | 0.58	0.87 | 0.86	**0.87 | 0.30**	
	layer-12	**0.72 | 0.65**	0.59 | 0.50	0.82 | 0.79	**0.86 | 0.30**	
last layer	random	0.70 | 0.70	0.59 | 0.59	0.48 | 0.48	0.50 | 0.50	
	step-0	0.71 | 0.70	0.61 | 0.57	0.87 | 0.88	**0.75 | 0.35**	
	step-1000	**0.73 | 0.69**	0.62 | 0.55	0.88 | 0.89	**0.89 | 0.31**	
	step-15000	**0.71 | 0.62**	0.61 | 0.48	0.84 | 0.83	**0.87 | 0.30**	
	step-43000	**0.72 | 0.62**	0.61 | 0.47	0.85 | 0.85	**0.86 | 0.31**	
	step-72000	0.71 | 0.59	0.60 | 0.44	0.83 | 0.83	**0.86 | 0.31**	
	step-100000	0.71 | 0.60	0.60 | 0.46	0.83 | 0.82	**0.86 | 0.29**	
	step-143000	**0.72 | 0.65**	0.59 | 0.50	0.82 | 0.79	**0.86 | 0.30**	
	chance	0.71	0.63	0.49	0.50	

Table 1: Probing task accuracies for different features for all the layers in the last model checkpoint and the last layer in all the checkpoints, including random tensor of the same size as model activations tensor. The last row in the table shows the proportion of the biggest class in the task, which we consider chance performance. In bold are results where the accuracy before removal is above chance, and the accuracy after removal is below chance. We consider such results as successful removal. Not all the tasks and checkpoints are present in the table and in the analysis due to the resource constraints.

3.2 Brain alignment

To compute brain alignment scores, we follow [4] and fit ridge regression with word features taken from a model as input and voxel activations as targets. We use 10-fold cross-validation to find the best regularization parameter from values $\lambda \in \{10^k | k \in [-6 \dots 10], k \in \mathbb{Z}\}$. As the times of words exposure in the dataset do not exactly match the times of brain measurements, we take the representation of the brain measurement as the mean of all the word representations exposed in the period before each measurement and after the previous one. To account for the time delays in fMRI recordings, we also add word representations from 4 previous measurements as extra features. We fit regressions separately for each linguistic task, for the features with removed linguistic property and for initial word features, for model layer, checkpoint, participant, and voxel.

4 Results

4.1 Brain alignment with initial model representations

In this section we confirm the results of [3] and observe that the mean Pearson correlation across all voxels throughout training resembles the perplexity curve 4.1.

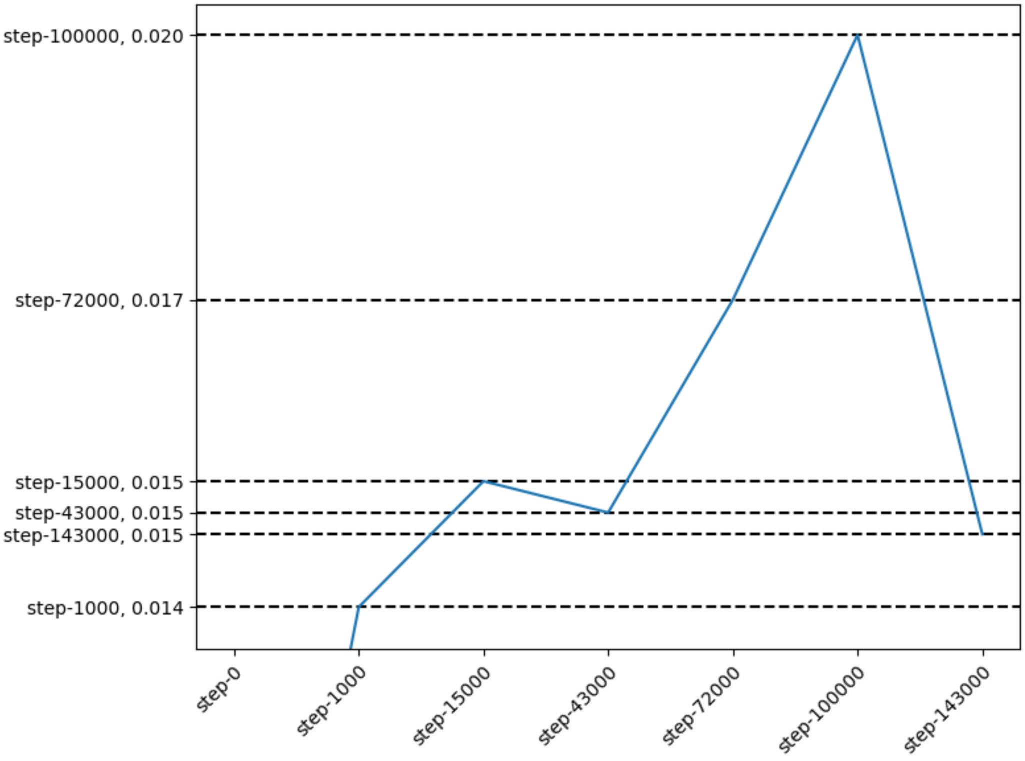

Figure 4: The average Pearson correlations across voxels between predicted and real brain activations. On x axis there are steps of model training, step-0 denotes random initializations, and step-143000 denotes final checkpoint. For each participant and each model step we take the layer with the highest average correlation and average that across participants. A graph where we take the average over layers is presented in A and it shows similar dynamics.

We also report the best performing layers according to mean Pearson correlation across voxels for each participant and each step in the figure 4.1. We find that the best performing layers vary highly between participants, but deep and middle layers perform better than shallow layers on average.

4.2 Brain alignment with residual model representations

In figure 4.2 for each model checkpoint and each linguistic task we report the layers of the model for which the difference in alignment scores between the initial and residual word representations is significantly different for all subjects. To determine the layers where the difference is significant we use a two-tailed t-test on the mean correlation values obtained from residual and initial word representations with p-value 0.05. We find that on average removal of linguistic properties affects the brain alignment more in later steps and middle layers.

Moreover, we report the difference between the overall average alignment scores of the initial model representations and the representations with each linguistic property removed for each model training step in figure 4.2. We find that the linguistic properties show different trends throughout training. First, the representations with removed random property in all the steps have approximately the same alignment scores as the initial representations, which validates our approach. Second, the removal of *Tense* property decreases the alignment score in all the considered training steps. Third, the removal

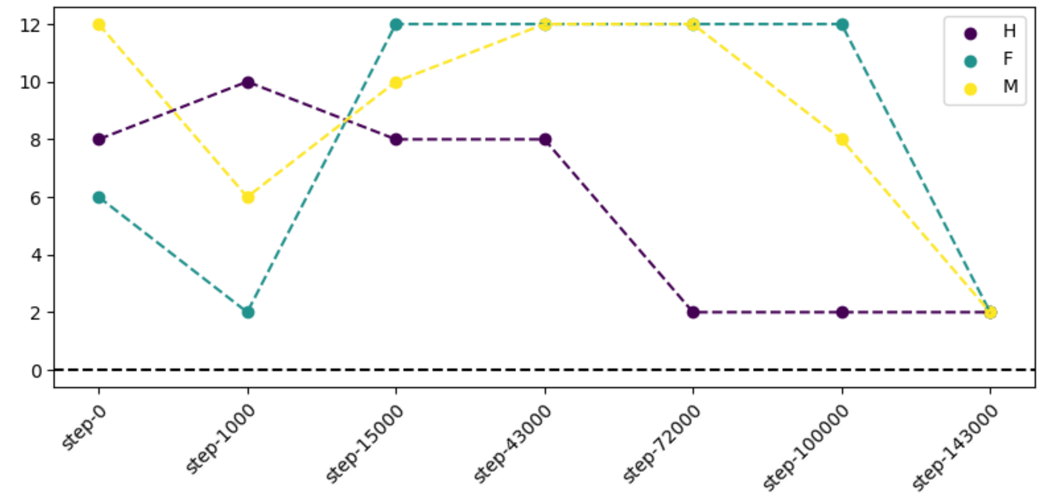

Figure 5: The best performing layers according to average Pearson correlation across voxels. On x axis there are steps of model training, step-0 denotes random initializations, and step-143000 denotes final checkpoint.

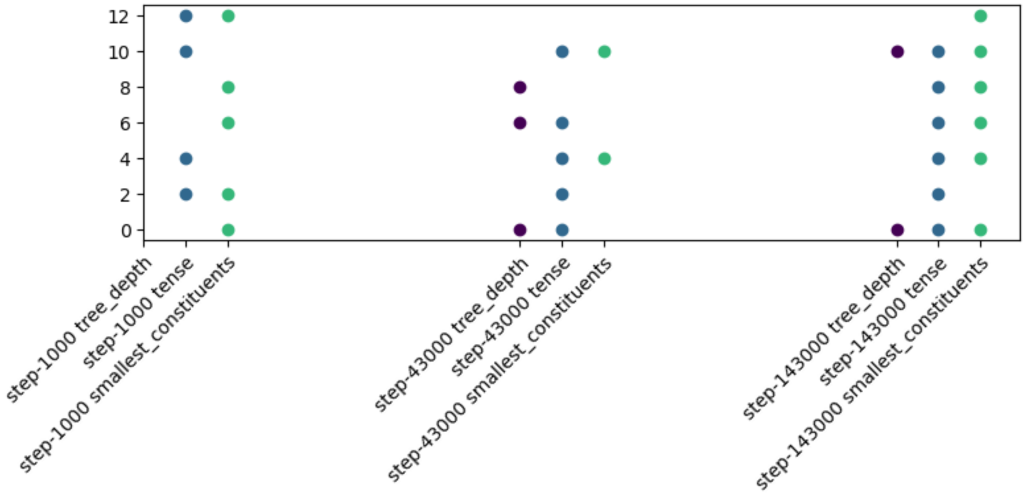

Figure 6: The layers of the models for which the removal of a linguistic property made a significant impact on alignment scores. The layers are represented in the y axis, while model training steps and linguistic properties are on x axis. The presence of a dot in the intersection of a feature, layer and model training step means that the brain alignment scores for this step and layer between initial and residual model representations are different with 95% confidence.

of *Tree Depth* property decreases the alignment scores starting from 30% of training. Lastly, the removal of *Smallest Constituents* property increases the alignment scores throughout training.

Lastly, we report average alignment scores for each layer in each step for different residual representations. We observe that the trends across layers are very similar for all the residual and initial representations, but diffrent for different model steps.

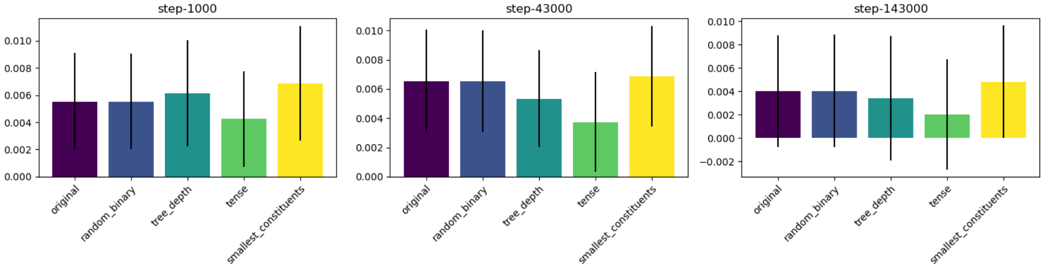

Figure 7: The difference between the average alignment scores of the initial model representations and the representations with each linguistic property removed. The average is computed across voxels, participants and layers, and the error bars are drawn for the averages computed across voxels and participants only.

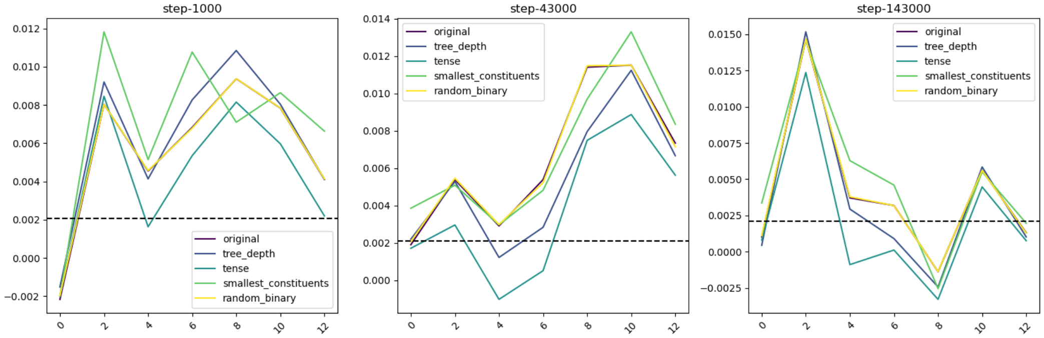

Figure 8: Alignment scores averaged across participants and voxels. The scores are given for each step, each layer (x axis) and the residual representations for each linguistic property (color). The original line stands for initial model representations without property removal. Black dotted line denotes the alignment score of the random tensor.

5 Conclusion

In this work we explored the emergence of alignment between word representations in language models and in the brain. We confirmed that the model representations alignment curve resembles the model perplexity curve and that middle layers and further steps of training produce better alignment results. We also observed that removal of some linguistic properties from model representations can affect the alignment differently depending on the training step when the removal happens and the linguistic property itself. Moreover, we find that the layers which better align with brain representations change throughout model training process, but stay similar with removal of linguistic properties.

References

1. Richard Antonello, Aditya Vaidya, and Alexander G Huth. “Scaling laws for language encoding models in fMRI”. In: *arXiv preprint arXiv:2305.11863* (2023).
2. Stella Biderman et al. “Pythia: A suite for analyzing large language models across training and scaling”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 2397–2430.
3. Eghbal A Hosseini et al. “Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training”. In: *BioRxiv* (2022), pp. 2022–10.
4. Subba Reddy Oota, Manish Gupta, and Mariya Toneva. “Joint processing of linguistic properties in brains and language models”. In: *arXiv preprint arXiv:2212.08094* (2022).

5. Peng Qi et al. “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2020. URL: `https://nlp.stanford.edu/pubs/qi2020stanza.pdf`.
6. Martin Schrimpf et al. “The neural architecture of language: Integrative modeling converges on predictive processing”. In: *Proceedings of the National Academy of Sciences* 118.45 (2021), e2105646118.
7. Xing Shi, Inkit Padhi, and Kevin Knight. “Does string-based neural MT learn source syntax?” In: *Proceedings of the 2016 conference on empirical methods in natural language processing*. 2016, pp. 1526–1534.
8. Mariya Toneva and Leila Wehbe. “Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)”. In: *Advances in neural information processing systems* 32 (2019).
9. Leila Wehbe et al. “Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses”. In: *PloS one* 9.11 (2014), e112575.

A Appendix

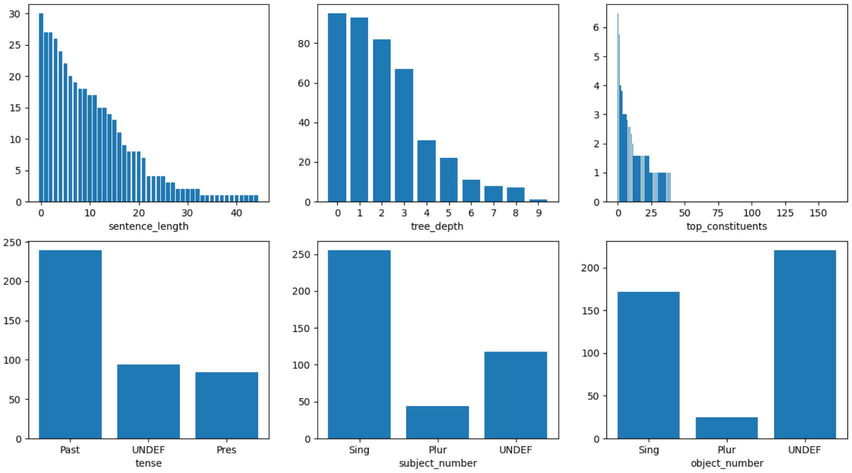

Figure 9: Original distribution of classes for sentence-level linguistic tasks. On y axis the number of sentences which have the corresponding property is reported. UNDEF states for the sentences where the value of the property is unclear or not applicable. For example, Object Number task is not applicable to the sentences with no direct object. The distribution of Top Constituents is given in log scale and the labels for this task are hidden for readability.

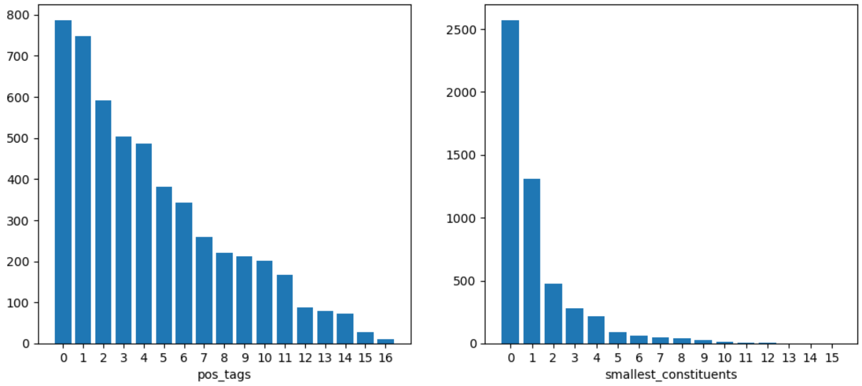

Figure 10: Original distribution of classes for word-level linguistic tasks. The labels are substituted with numbers for readability and sorted by the number of words which have this labels.

	SentLen	TreeDep	TopConst	Tense	SNum	ONum	POS	SmConst
SentLen	1.0	0.7	*-0.0	-0.2	-0.2	-0.0	*-0.0	0.0
TreeDep	0.7	1.0	-0.1	-0.1	-0.2	-0.0	*-0.0	0.0
TopConst	*-0.0	-0.1	1.0	0.1	0.1	-0.1	0.0	*0.0
Tense	-0.2	-0.1	0.1	1.0	0.3	*-0.0	*0.0	*0.0
SNum	-0.2	-0.2	0.1	0.3	1.0	0.1	*0.0	*-0.0
ONum	-0.0	-0.0	-0.1	*-0.0	0.1	1.0	*0.0	-0.0
POS	*-0.0	*-0.0	0.0	*0.0	*0.0	*0.0	1.0	-0.1
SmConst	0.0	0.0	*0.0	*0.0	*-0.0	-0.0	-0.1	1.0

Table 2: Pearson correlations between tasks labels. The correlations with p-value greater or equal to 0.05 are reported with an asterisk.

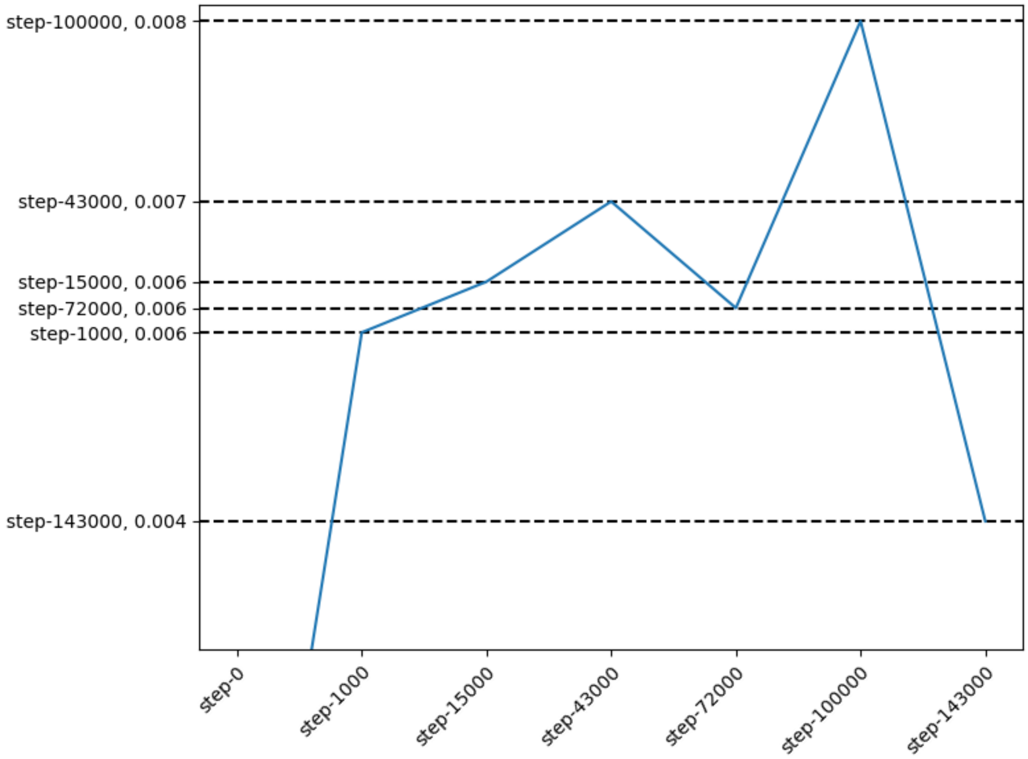

Figure 11: The average Pearson across voxels correlations between predicted and real brain activations. On x axis there are steps of model training, step-0 denotes random initialization, and step-143000 denotes final checkpoint. For each model step we take the average across correlation across voxels, layers and participants.