Science Unleashed – Rigor Retained

On the Shape of Brainscores for Large Language Models (LLMs)

Jingkai Li[1]

[1]OpenSci.World
Montréal, Québec H4R 2R9, Canada

Research Interests: Computational Brain/Neural Science &
its Cross-disciplinary Topics with AI/ML
Research Cooperation Opportunities are Welcome!

jingkai.li@opensci.world OR
jkli898@126.com

ICLR AGI Workshop, May 11, 2024

Table of Contents

Table of Contents

Introduction - Background

- Artificial Neural Networks (ANNs) were inspired by biological neural networks (Russell & Norvig, 2010)
- Connectionism (Bognar, 2022)
- Sparks of AGI (Bubeck et al., 2023)
- Two-way efforts: LMs to investigate brain/neural science & neuroimaging data to improve NLP models (Karamolegkou et al., 2023)

Introduction - Motivation

- Preceding the realization of AGI, a crucial imperative arises: we want to, and **need** to know the extent of human-likeness inherent in the LLMs under development.
- *The extent to which ANNs have diverged from biological neural networks* remains largely unknown and vastly underexplored, thereby motivating our study.

Introduction - Brainscores

- The Pearson Correlation/Brainscore metric (Schrimpf et al., 2018), as surveyed in Karamolegkou et al. (2023), stands as the predominant method in mapping brains with LMs.
- Other studies rely on the Brainscore metric: Caucheteux et al. (2023) and Oota et al. (2023) etc.
- The brainscores that we try to interpret are based on Caucheteux et al. (2023).

Introduction - Research Questions

- In addition to those efforts to calculate "brainscores", our study seeks to address the fundamental questions: *What is the meaning of the score? Can we derive features to interpret it?*
- Our study extends the concept of "Shape" by systematically comparing the distinctions in the "shapes" between fMRI and LLMs embeddings.

Introduction - Research Questions

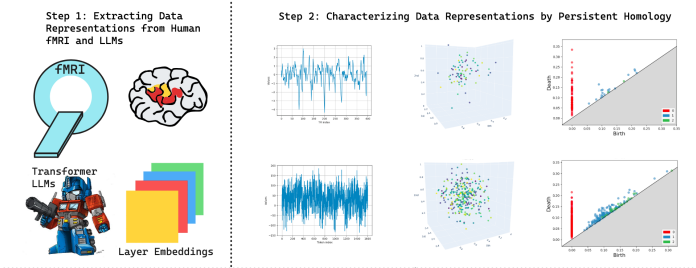
- Leveraging the Topological Data Analysis (TDA) tool Persistent Homology (PH), we characterize data representations in both realms and construct features by computing q-Wasserstein Distances between pairs of their persistence diagrams.
- Subsequently, we learn Linear Regression Models to fit the existing "brainscores", followed by rigorous statistical analyses to identify reliable and valid features among our constructed ones.

Introduction - Contributions

- Ontologically, distinct feature combinations to facilitate the interpretation of existing brainscores.
- Epistemologically, new perspectives (PH + q-Wasserstein distances) to construct features.
- Overall, we take small steps in trying to address the fundamental inquiries: *"In what sense are the LLMs we create human-like?"*

Table of Contents

Methods - Overview



Methods - An Intuitive Introduction to fMRI

- Functional Magnetic Resonance Imaging - fMRI
- Blood Oxygenation Level Dependent - BOLD
- The fMRI data utilized in our investigation originate from the Narratives dataset (Nastase et al., 2021)

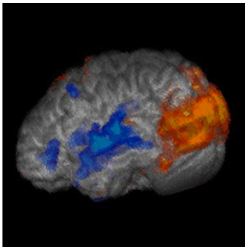


Figure: Original images from https://www.ndcn.ox.ac.uk/divisions/fmrib/what-is-fmri/introduction-to-fmri

Methods - An Intuitive Introduction to PH

- In our investigation, we utilized the powerful tool Persistent Homology (PH) to characterize the extracted data representations obtained from fMRI and LLMs embeddings.

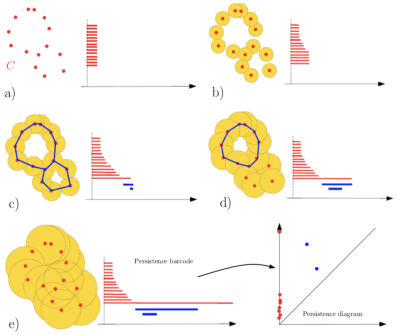


Figure: Original figures from https://github.com/GUDHI/TDA-tutorial/blob/master/Tuto-GUDHI-persistence-diagrams.ipynb

Methods - Intuition in q-Wasserstein Distance

- In our study, series of q-Wasserstein Distances are calculated with the aim of deriving features that subsequently aid in the interpretation of the corresponding brainscores.

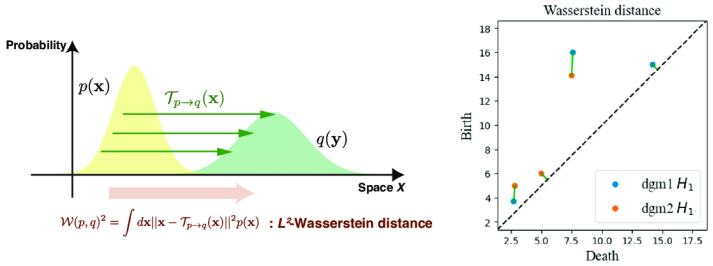


Figure: Original figures: left (Nakazato & Ito 2021), right (Zhang et al. 2021)

Methods - Implementation Details and Experiments

- **Extracting Data Representations from Human fMRI.**
- We systematically derived data representations for *each task, subject, hemisphere, and Region of Interest (ROI)* from the Narratives dataset (Nastase et al., 2021).
- 3 Tasks: "Pie Man", "Shapes", "It's not the Fall that Gets You"
- 8 ROIs: angular gyrus (AG), anterior temporal lobe (ATL), posterior temporal lobe (PTL), inferior frontal gyrus (IFG), middle frontal gyrus (MFG), inferior frontal gyrus orbital (IFGOrb), posterior cingulate cortex (PCC), dorsal medial pre-frontal cortex (dmPFC) + whole brain mask

Methods - Implementation Details and Experiments

- **Extracting Data Representations from Human fMRI.**
- Summary for our three selected tasks from Narratives dataset (Nastase et al., 2021).

Story	Duration	TRs	Words	Subjects	Onset	Valid TRs	Valid Subjects
"Pie Man"	07:02	300	957	82	0	300	75
"Shapes"	06:45	313	910	59	3	310	58
"It's Not the Fall That Gets You"	09:07	400	1,601	56	3	397	54

Methods - Implementation Details and Experiments

- **Extracting Data Representations from Human fMRI.**

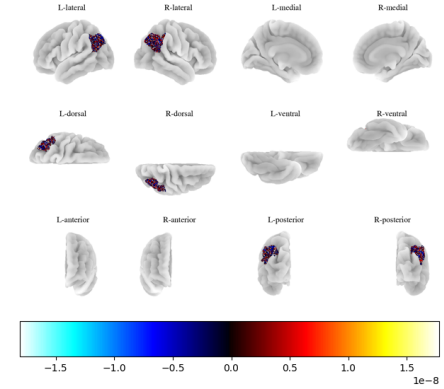


Figure: ROI: AG for: "Pie Man"

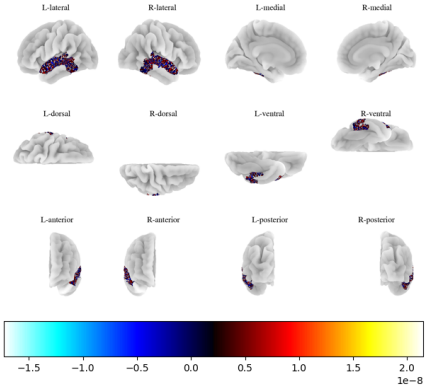


Figure: ROI: PTL for: "Pie Man"

Methods - Implementation Details and Experiments

- **Extracting Data Representations from Human fMRI.**

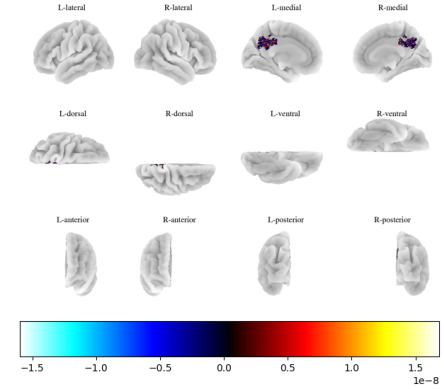


Figure: ROI: PCC for: "Pie Man"

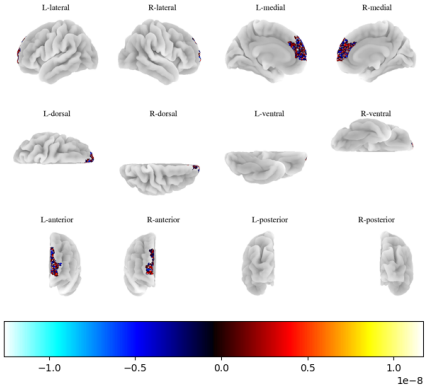


Figure: ROI: dmPFC for: "Pie Man"

Methods - Implementation Details and Experiments

- **Extracting Data Representations from Human fMRI.**
- We derived 3,366 matrices in total, each shape of those being like #TR $\times$ 40,962 voxels (then reduced to 75 voxels).
- Similar to Caucheteux et al. (2023), formally, we denote:
- Y as the fMRI recordings elicited by a subject listening to a task, of size $T \times V$, with T as the number of fMRI time samples (TRs) and V as the number of voxels.

Methods - Implementation Details and Experiments

- **Extracting Data Representations from Human fMRI.**
- Summary for fMRI data representations for our selected 3 tasks from Narratives dataset (Nastase et al., 2021).

Story	# Subjects	# Hemispheres	# ROIs	Total
"Pie Man"	75	2	8+1	$75 \times 2 \times (8+1) = 1350$
"Shapes"	58	2	8+1	$58 \times 2 \times (8+1) = 1044$
"It's Not the Fall That Gets You"	54	2	8+1	$54 \times 2 \times (8+1) = 972$
Total				3,366

Methods - Implementation Details and Experiments

- **Extracting Data Representations from LLMs.**
- As a large-scale study, we examined and analyzed 39 LLMs [1], spanning from albert-base to Llama-70B (quantized), plus their untrained counterparts [2]. To be in line with Schrimpf et al. (2021), we didn't make a distinction between masked language models and causal language models, but limit the architecture of our interest as Transformer (Vaswani et al., 2017) based one.

[1] All the LLMs analysed are publicly accessible via https://huggingface.co/.

[2] Untrained versions were unavailable for LLMs: Llama-70B (quantized), Llama-13B (quantized) and Llama-7B (quantized), thus yielding a total of 36 untrained LLMs.

Methods - Implementation Details and Experiments

- **Extracting Data Representations from LLMs.**
- We sampled and extracted embeddings for *each LLM across each task*.
- This process yielded a total of 225 tensors, encompassing 39 trained LLMs and 36 untrained ones, each multiplied by 3 tasks.
- The tensor is three-dimensional, representing # LLM layers, # tokens in the task, and the LLM embedding dimension.

Methods - Implementation Details and Experiments

- **Extracting Data Representations from LLMs.**
- Similar to Caucheteux et al. (2023), we denote:
- w as a sequence of M words for each task.
- X as the embeddings of a LLM model input with w, of size $M \times U$, with U as the dimensionality of the embeddings (for a layer of i.e. GPT-2, $U = 768$). We explicitly denote X_k as the embeddings extracted from layer k.

Methods - Implementation Details and Experiments

- **Characterizing the fMRI Data Representations by PH.**
- Average across all valid subjects for *each task*, *hemisphere*, *and ROI*.
- Aggregate along the TR dimension $\rightarrow$ 1-D time-series vector.
- len(vector) = # TR for each task.

Methods - Implementation Details and Experiments

- **Characterizing the fMRI Data Representations by PH.**

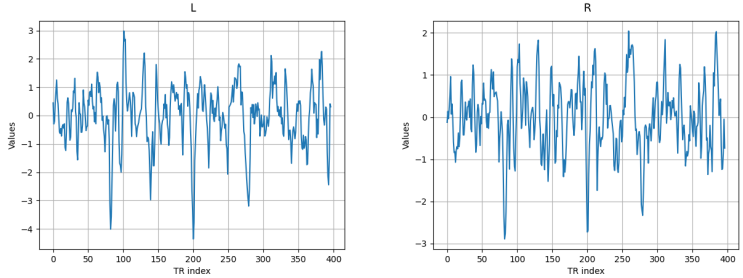


Figure: This figure depicts the time-series of the fMRI BOLD values for both the (L)eft and (R)ight hemispheres and the region of interest (ROI): PCC. These values are averaged across 54 out of 56 subjects while they listened to the task "It's Not the Fall that Gets You," and subsequently aggregated along the TR dimension.

Methods - Implementation Details and Experiments

- **Characterizing the fMRI Data Representations by PH.**
- The TDA Toolkit Giotto-TDA (Tauzin et al., 2021) was employed to project the aforementioned time-series signals into a 3-D space as a point cloud.
- The shape of our projected point cloud is interpreted as optimal time delay $\tau\times$ restricted embedding dimension $d = 3$.
- All values of the point cloud were normalized to fall within the range of $[0, 1]$.
- Finally, we employed the "VietorisRipsPersistence" transformer from Giotto-TDA to execute the computation for persistence.

Methods - Implementation Details and Experiments

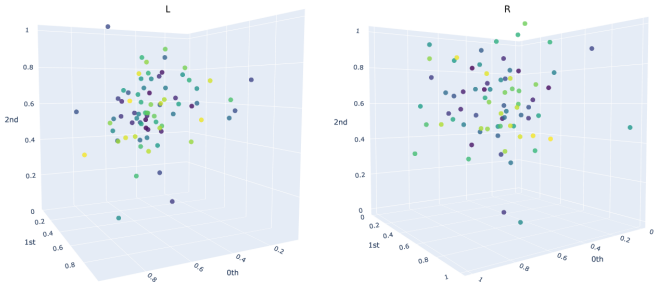


Figure: This figure presents the projected point cloud in 3-D space for the time-series signals. In this specific instance, the embedding dimension is constrained to 3, and the optimal time delay is determined to be $\tau = 78$ for the (L)eft hemisphere and $\tau = 76$ for the (R)ight one. This result pertains to the task "It's Not the Fall that Gets You", which comprised a total of 397 valid TRs.

Methods - Implementation Details and Experiments

- **Characterizing the fMRI Data Representations by PH.**

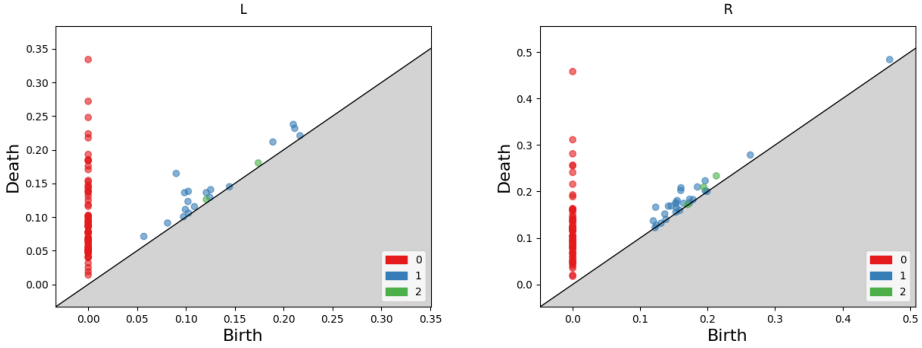


Figure: This figure depicts the persistence diagram derived from the Persistent Homology analysis conducted on the point cloud.

Methods - Implementation Details and Experiments

- **Characterizing the fMRI Data Representations by PH.**

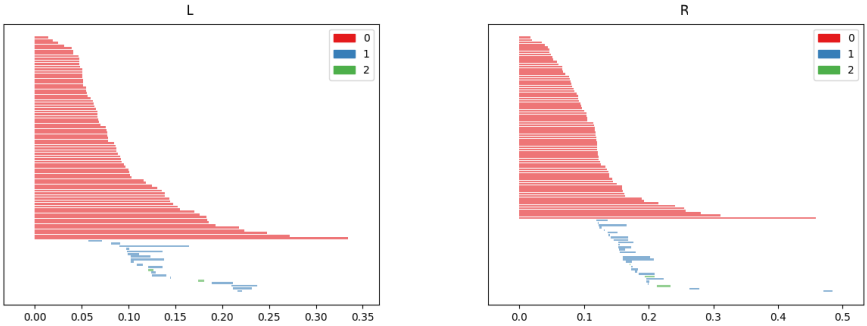


Figure: This figure depicts the persistence barcode derived from the Persistent Homology analysis conducted on the point cloud.

Methods - Implementation Details and Experiments

- **Characterizing the LLMs Embeddings by PH.**
- Aggregate along the Token dimension $\rightarrow$ 1-D time-series vector, for each specific LLM layer [3].
- len(vector) = # Tokens for each task.

[3] we did not process every individual layer of each LLM. Instead, we uniformly sampled eight layers evenly spaced from the first one to the last (inclusive at both ends), plus the intermediate-to-deep layer ($l = \frac{2}{3} n_{layers}$) in line with Caucheteux et al. (2023).

Methods - Implementation Details and Experiments

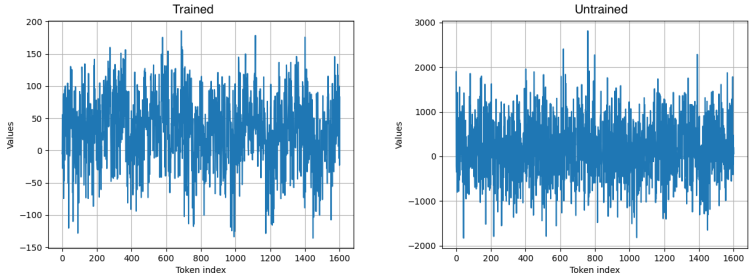


Figure: This figure presents the time-series for the embedding from Layer 21 of the LLM: meta-llama/Llama-2-7b-hf (https://huggingface.co/meta-llama/Llama-2-7b-hf), both trained and untrained, when provided with the task "It's not the Fall that Gets You". The embeddings are then aggregated along the token dimension. Layer 21 from meta-llama/Llama-2-7b-hf corresponds to the intermediate-to-deep layer of the LLM ($l = \frac{2}{3} n_{\text{layers}}$), as defined by Caucheteux et al. (2023).

Methods - Implementation Details and Experiments

- **Characterizing the LLMs Embeddings by PH.**
- The TDA Toolkit Giotto-TDA (Tauzin et al., 2021) was employed to project the aforementioned time-series signals into a 3-D space as a point cloud.
- The shape of our projected point cloud is interpreted as optimal time delay $\tau\times$ restricted embedding dimension $d = 3$.
- All values of the point cloud were normalized to fall within the range of $[0, 1]$.
- Finally, we employed the "VietorisRipsPersistence" transformer from Giotto-TDA to execute the computation for persistence.

Methods - Implementation Details and Experiments

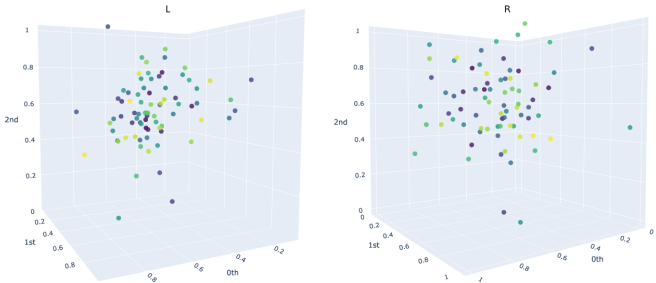


Figure: This figure depicts the projected point cloud in 3-D space for the time-series signals. In this specific instance, the embedding dimension is constrained to 3, with the optimal time delay searched being $\tau = 320$ for the trained model and $\tau = 318$ for the untrained one. The total number of tokens for the task "It's not the Fall that Gets You" is 1601.

Methods - Implementation Details and Experiments

- **Characterizing the LLMs Embeddings by PH.**

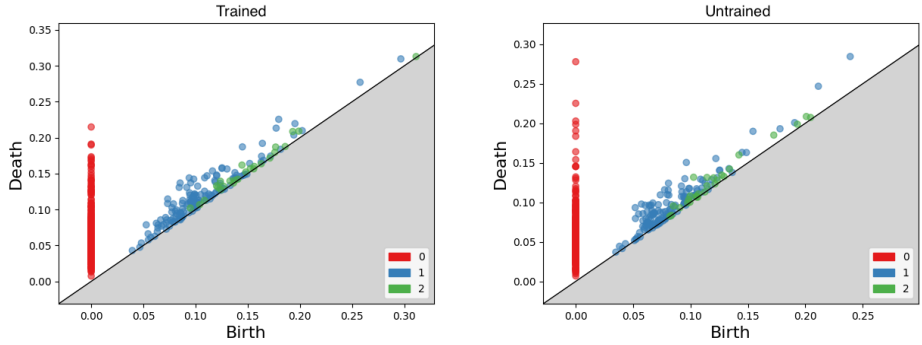


Figure: This figure presents the persistence diagram summarized from the Persistent Homology computed from the point cloud.

Methods - Implementation Details and Experiments

- **Characterizing the fMRI Data Representations by PH.**

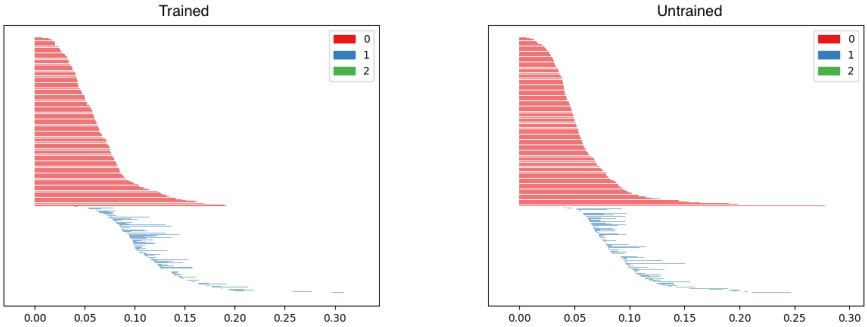


Figure: This figure presents the persistence barcode summarized from the Persistent Homology computed from the point cloud.

Methods - Implementation Details and Experiments

- **Computing q-Wasserstein Distances between Persistence Diagrams.**
- A total of 34,416 combination pairs were generated encompassing both fMRI data and LLMs Embeddings.
- The identification of each pair involved a combination of a *particular ROI, hemisphere, LLM layer, and training status*, consuming the same *task*.
- The ensuing step required the computation of q-Wasserstein Distances for each unique combination [4].

[4] We utilized the GUDHI library (Maria et al., 2014) to generate summary representations, i.e. persistence diagrams and persistence barcodes, from our computed PH. This library was also employed in computing q-Wasserstein Distances.

Methods - Implementation Details and Experiments

- **Learning from q-Wasserstein Distances against Existing Brainscores.**
- The training data was partitioned to learn Linear Regression Models for *each ROI across each hemisphere* in two: "Only Trained" LLMs and "Trained plus Untrained" LLMs.
- A total of 36 models were learned, covering (8+1) ROIs $\times$ 2 hemispheres $\times$ 2 training groups.
- Each of the 34,416 combination pairs is characterized by 903 features, representing the q-Wasserstein Distances, where $q = p$, ranging from 1 to 300, along with the special case $q = p = \infty$, across 3 PH dimensions.

Methods - Implementation Details and Experiments

- **Learning from q-Wasserstein Distances against Existing Brainscores.**
- Exploratory Data Analysis (EDA) on our constructed features.
- A consistent "long tail" distribution followed by the q-Wasserstein Distances spanning from $q = p = 1$ to $q = p = 300$ and $q = p = \infty$ for each pair in the 34,416 combinations, with minor fluctuations in the tail.

Methods - Implementation Details and Experiments

- **Learning from q-Wasserstein Distances against Existing Brainscores.**
- The following figure illustrates the q-Wasserstein Distances computed between the persistence diagrams for fMRI data from ROI: PCC, hemisphere: L, and embeddings from LLM: Llama-2-7b-hf, training status: trained, layer: 21, under the same task: "It's not the Fall That Gets You". The figure displays the results for all three persistent homology dimensions: 0, 1, and 2, respectively.

Methods - Implementation Details and Experiments

- **Learning from q-Wasserstein Distances against Existing Brainscores.**
- Across the q-Wasserstein Distances, we observe "long tail" distributions, where the values converge to approximately 0.30 for persistent homology dimension 0, and approximately 0.14 and 0.07 for persistent homology dimensions 1 and 2, respectively.

Methods - Implementation Details and Experiments

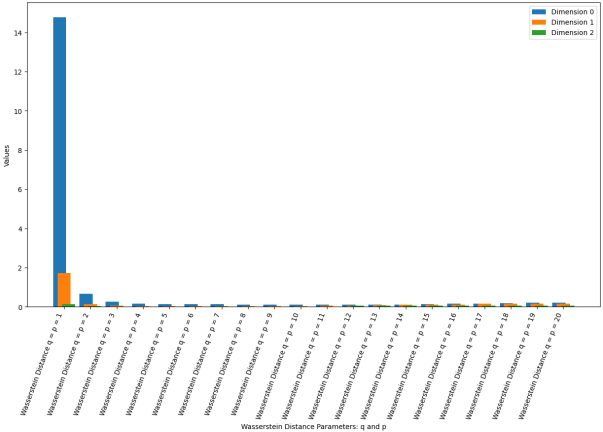


Figure: The figure depicts the q-Wasserstein Distances with parameters q and p spanning from $q = p = 1$ to $q = p = 20$.

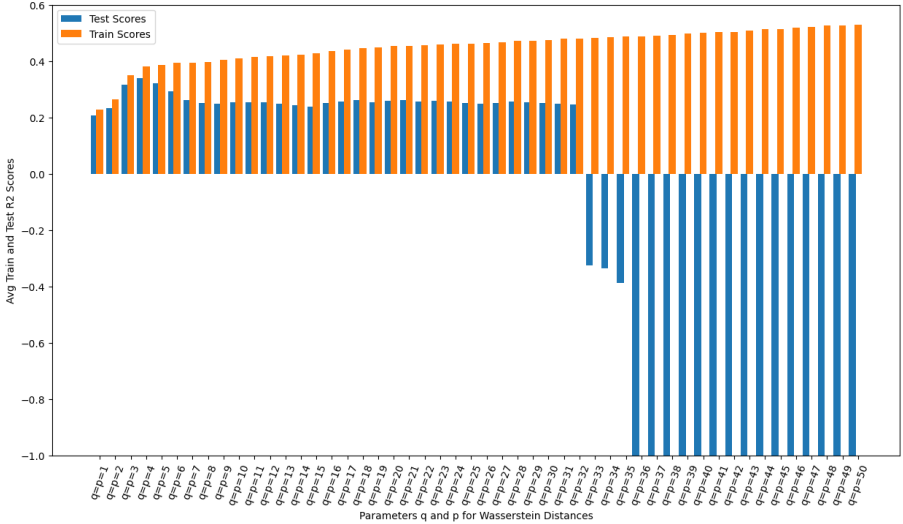
Methods - Implementation Details and Experiments

- **Learning from q-Wasserstein Distances against Existing Brainscores.**
- In the **First Pass**, we incrementally introduced in features from $q = p = 1$ to $q = p = 300$.
- At each iteration, we implemented a 5-time repeated 5-fold cross-validation (CV) pipeline to train the Linear Regression Model.
- We assessed the model's performance using the averaged test R^2 score across the 25 outcomes.
- We employed a stopping criterion triggered by encountering a negative averaged test R^2 score.

Methods - Implementation Details and Experiments

- **Learning from q-Wasserstein Distances against Existing Brainscores.**
- The following figure depicts the averaged train and test R^2 scores acquired during the learning process of the Linear Regression model to correlate with the respective brainscores using "Only Trained" LLMs. The train scores exhibit a gradual increase as additional q-Wasserstein Distances are incorporated progressively. Conversely, the test scores deviate from the pattern, reaching a peak at $q = p = 4$ before sharply declining, with a significant downturn observed beyond $q = p = 35$ (we just hide those very large negative test scores below -1.

Methods - Implementation Details and Experiments



Methods - Implementation Details and Experiments

- **Learning from q-Wasserstein Distances against Existing Brainscores.**
- Formally, we established a mapping $\phi : q \to s$ where $q \in [1, 300] \cup \{+\infty\}$ denotes the search space, and $s \in [0, 1]$ represents the corresponding averaged test R^2 score for each iteration. The q value is determined as $q_{\max} = \arg\max \phi$.

Methods - Implementation Details and Experiments

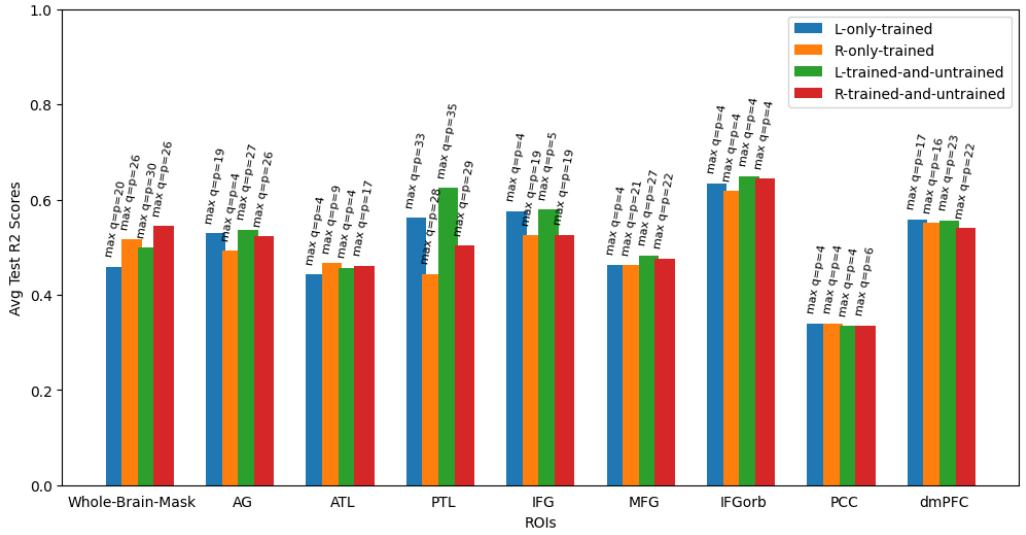
- **Learning from q-Wasserstein Distances against Existing Brainscores.**
- In the **Second Pass**, we designated the p-value with a threshold of $p < 5\%$ to selectively retain desired features given the determined q value.
- Specifically, we computed the p-value for each iteration, averaged the 25 p-values corresponding to each feature, and ultimately retained those with $p < 5\%$.

Table of Contents

Results - First Pass

- The following figure illustrates the maximum q for each ROI and each hemisphere that achieves the highest averaged test R^2 score as our result for the **First Pass**.
- Additionally, it employs distinct colors to denote whether the training dataset encompasses untrained LLMs.

Results - First Pass



Results - Second Pass

- Each table encapsulates the features that ultimately passed through the filtering process for each ROI (including the whole brain mask) and hemisphere (L and R) with each feature's Weight (plus both Lower and Upper Bound of 95% Confidence Intervals), the Standard Error (SE) across all 25 runs, the Feature Importance t-statistic value $|t|$, and the corresponding p value.
- Each table is then further accompanied with two figures in its following, illustrating each feature's Weight across both Lower and Upper Bound of 95% Confidence Intervals, and its Coefficient Importance and Variability respectively.

Results - Second Pass

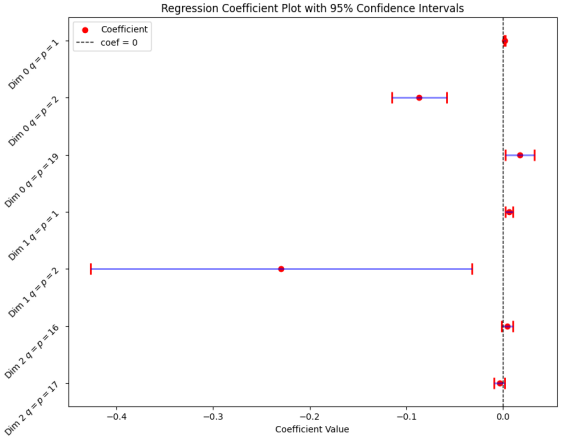
- This table summarizes the filtered-in features generated from the **Second Pass** for the ROI: AG and the hemisphere: L(eft) where the training data contains "Only Trained" LLMs.

Feature	Weight	95% CI Lower	95% CI Upper	SE	$\lvert t \rvert$	p
(Intercept)	$7.7569e{-3}$	$5.2319e{-3}$	$1.0282e{-2}$	$8.9245e{-4}$	$8.0471e0$	$0.0000e0$
Dim 0 $q = p = 1$	$1.7418e{-3}$	$1.3126e{-3}$	$2.1710e{-3}$	$1.9325e{-4}$	$9.1925e0$	$0.0000e0$
Dim 0 $q = p = 2$	$-8.6721e{-2}$	$-1.1524e{-1}$	$-5.8203e{-2}$	$1.9219e{-2}$	$4.9603e0$	$0.0000e0$
Dim 0 $q = p = 19$	$1.7606e{-2}$	$2.5294e{-3}$	$3.2683e{-2}$	$4.8392e{-3}$	$5.4937e0$	$2.6400e{-3}$
Dim 1 $q = p = 1$	$6.3380e{-3}$	$2.4140e{-3}$	$1.0262e{-2}$	$1.4778e{-3}$	$4.2500e0$	$1.0000e{-2}$
Dim 1 $q = p = 2$	$-2.2954e{-1}$	$-4.2704e{-1}$	$-3.2041e{-2}$	$7.8173e{-2}$	$3.3021e0$	$3.3440e{-2}$
Dim 2 $q = p = 16$	$4.3482e{-3}$	$-1.7606e{-3}$	$1.0457e{-2}$	$1.1942e{-3}$	$6.2958e0$	$2.5680e{-2}$
Dim 2 $q = p = 17$	$-3.7191e{-3}$	$-9.1522e{-3}$	$1.7140e{-3}$	$1.2206e{-3}$	$4.9698e0$	$4.4200e{-2}$

Results - Second Pass

- The following figure illustrates the filtered-in features generated from the **Second Pass** for the ROI: AG and the hemisphere: L(eft) where the training data contains "Only Trained" LLMs. On the left sub-figure, each feature's Weight is presented along with both the Lower and Upper Bounds of the 95% Confidence Intervals, while the right sub-figure illustrates its Coefficient Importance and Variability.

Results - Second Pass



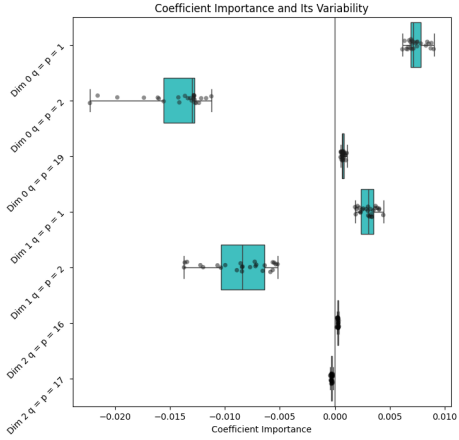


Table of Contents

Discussions and Limitations

- The distinct feature combinations facilitate the interpretation of the current brainscores associated with different ROIs and hemispheres.
- It is notable that while the brainscore serves as a metric denoting the similarity between fMRI and LLM data representations, the q-Wasserstein Distance quantifies the dissimilarities between two distributions. Therefore, particular attention should be directed towards features bearing negative weights.

Discussions and Limitations

- Our learned Linear Regression Models and the subsequently filtered-in features are not yet in their definitive optimal state, as evidenced from the **First Pass**, where all the highest averaged test R^2 scores remain below the ideal threshold to varying degrees.
- Potential avenues for improvement includes dealing with outliers on persistence diagrams.
- The exploration of different q and p values for constructing features through q-Wasserstein Distances.
- The utilization of diverse Linear Regression Models, i.e. Ridge vs. OLS.

Discussions and Limitations

- Moreover, there has been limited comparison of the structural properties between human brain/neural systems and LLMs. We thus propose it as another promising avenue for research.

Table of Contents

Conclusions

- We devoted our efforts in mining the meaning of the novel metric brainscores through the construction of topological features derived from both human fMRI data, encompassing 190 subjects, and 39 LLMs along with their untrained counterparts.
- Subsequently, we trained a total of 36 Linear Regression Models and conducted thorough statistical analyses to discern reliable and valid features from our constructed ones.
- Our findings reveal distinctive feature combinations conducive to interpreting existing brainscores across various ROIs and hemispheres, thereby contributing significantly to advancing iML studies.

Conclusions

- To our knowledge, this study represents the first attempt to comprehend the novel metric brainscore within this interdisciplinary domain.
- Overall, we take incremental steps toward addressing fundamental inquiries: *"In what sense are the LLMs we create human-like?"*.

Thank you for your time and attention!

Jingkai Li[1]

[1]OpenSci.World
Montréal, Québec H4R 2R9, Canada

Research Interests: Computational Brain/Neural Science &
its Cross-disciplinary Topics with AI/ML
Research Cooperation Opportunities are Welcome!

jingkai.li@opensci.world OR
jkli898@126.com

ICLR AGI Workshop, May 11, 2024

Science Unleashed – Rigor Retained