Universitat Pompeu Fabra Barcelona

Treball de fi de màster de Recerca

Scrutinizing the Predictive Power of Large Language Models for Brain Function

Nom i Cognoms Ni Yang

Màster: Lingüística Teòrica i Aplicada

Edició: 2023-2024

Directors: Dr. Wolfram Hinzen

Any de defensa: 2024

Col·lecció: Treballs de fi de màster

Departament de Traducció i Ciències del Llenguatge

Abstract

Based on predictive coding and hierarchical processing as a commonality between large language models (LLMs) and the brain, many studies have linked the two by regressing brain activity on LLMs' representations. However, increasing evidence has revealed problems in this new line of research. To address this issue, we attempted to replicate a pioneering study (Kumar et al., 2022) on an independent fMRI dataset with several methodological adaptations. Results showed overall low correlation scores and sparse predictions across the cortex. Contrary to the reference study, representation's performances across most ROIs did not differ significantly. However, in areas where significant differences were observed, fastText consistently outperformed BERT. Additionally, the layer-wise performance of embeddings and transformations showed no consistent patterns. Our findings challenge the existing assumptions regarding the predictive power of LLMs for brain function and highlight potential issues in the current methodologies to map predictions from LLMs onto brain activations.

Keywords: LLMs, Bertology, computational linguistics, neurolinguistics, natural language processing, Transformers, predictive coding

Table of Contents

1. Introduction

Large language models (LLMs) have made striking advances in recent years, showing outstanding performance on multiple benchmarks, including information retrieval, commonsense reasoning, question answering, and human-like text generation. Promising results have also foregrounded a convergence between brain activations and LLM representations (Anderson et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2021; Pasquiou et al., 2022; Schrimpf et al., 2021), leading to optimistic views that LLMs could predict and illuminate human brain responses to language (Tuckute et al., 2024). However, a number of recent studies have identified problems concerning the theoretical foundations, the interpretation of the previous results, and methodological shortcomings in the earlier research (Bever et al., 2023; Bowers et al., 2023; Feghhi et al., 2024; Zhang et al., 2024). It has become increasingly evident that replication studies supporting such a predictive power are still required and that a gap still exists between current next-token predictors and our biological brains in language comprehension.

Language comprehension is a quintessentially human ability. With recent progress in cognitive science, a considerable amount of theories have emerged to model the process of natural language comprehension in the human brain. At a theoretical level, a link between LLMs and human language comprehension would, in particular, be expected based on the predicting coding theory as a general framework for understanding brain function. It posits that the brain proactively generates anticipatory predictions about upcoming words, sentences, and events during the perception of natural language stimuli (Huang & Rao, 2011; Van Berkum et al., 2005). As evidenced by studies of electrophysiological signals associated with syntactically or semantically surprising words, the brain then uses these predictions to minimize the error between its expectations and the actual language input, a process often referred to as minimizing “prediction error” (Heilbron et al., 2022; Schmitt et al., 2021; Shain et al., 2020).

This predictive mechanism of the human brain has been revealed to be layered and integrated across multi-level linguistic representations, exemplifying the predictive encoding in a hierarchical structure (Heilbron et al., 2021, 2022). Studies in the functional neuroanatomy of language have also supported such a hierarchy via cognitive models of spoken language comprehension, which assume language processing is hierarchically organized in our brain, with increasingly larger abstraction from the acoustic properties of speech at higher processing regions. Specifically, speech processing initially starts from the primary auditory cortex (Heschl's gyrus) in the superior temporal lobe to receive acoustic-phonological signals, eventually spanning to an array of higher-level function regions related to semantic memory (comprising the left anterior temporal lobe (ATL) in particular) and syntactic integration (involving the left inferior frontal gyrus (IFG) among other superior and middle temporal regions), to eventually form a fuller, contextual understanding of language encompassing brain areas like the bilateral middle temporal gyrus (MTG) and left angular gyrus (AG) (Grodzinsky & Friederici, 2006; Hagoort & Indefrey, 2014; Hickok & Poeppel, 2007; Matchin & Hickok, 2020; Van Berkum et al., 2005)

Inspired by biological computation, the next-sentence prediction (NSP), as a fundamental pretraining task of LLMs, has recently been shown to significantly improve their prediction of brain activity in the right hemisphere and the multiple demand network in discourse-level comprehension (Yu et al., 2024). This finding underscores the efficacy of NSP as a training objective, enabling LLMs to capture human comprehension and encode contextual information more accurately. As an embodiment of predictive encoding, LLMs also optimize their "comprehension" by minimizing the loss function and subsequently adjusting the parameters through an iterative process to enhance their predictions during self-supervised training (Radford et al., 2019; Vaswani et al., 2023). As one of the most influential LLMs, Transformers process natural language input as tokens and build up a hierarchical

understanding by aggregating the integrated contextual information through layers with each head attending to particular linguistic properties (Clark et al., 2019; Tenney et al., 2019; Vaswani et al., 2023). Predictive encoding assigns the task of predicting words or sentences to Transformers, relying not solely on explicit textual cues but also on implicit contextual information. Technically, such context-aware encoding of the natural language is achieved by its revolutionizing attention mechanism. Initially, input embeddings are transformed into three separate matrices for queries, keys, and values. By calculating the dot product of each query-key pair, the attention scores are then scaled and normalized to produce attention weights. These weights are used to compute a weighted sum of the value vectors, resulting in aggregated information for each word in the context of the entire sequence. The normalized output is passed through a feed-forward neural network, enabling the model to capture more complex representations. During this process, each attention head performs the above steps independently while working in parallel with other heads, allowing the model to capture different types of dependencies in the input. Eventually, the outputs from all heads are concatenated and then passed through a final linear transformation to combine the information into an integrated single output.

BERT (Bidirectional Encoder Representations from Transformers), functioning as an encoder-centric model with the Transformer architecture, specializes in language understanding (Devlin et al., 2019). Its layer-head specialization potentially contributes to its biological plausibility, facilitating the alignment of hierarchical computations in LLMs with the neural processes of the brain. It has been suggested that the intermediate layers of BERT together compose a rich hierarchy of linguistic information, spanning from surface features at the bottom, syntactic features in the middle, and semantic features at the top (Jawahar et al., 2019). Progressively, more complex tasks are focused on later layers (De Vries et al., 2020). Transformations, in the previous research, have also displayed a head-wise varying reflection

in the processing of linguistic properties (Clark et al., 2019). These architectural features of BERT models particularly raise the question of whether they might mirror how the brain processes linguistic information across different levels of linguistic organization.

Given the commonalities between artificial and biological neural networks in predictive coding principles, hierarchical language processing, and linguistic task-handling abilities, it becomes reasonable to construct a direct linkage between the two. Under the hypothesis that the shared computational principal could be assessed by evaluating predictions by LLMs through a voxel-wise encoding model, a large body of recent studies has attempted to show that Transformers have outperformed earlier methods in explaining brain activity elicited by language stimuli (Anderson et al., 2021; Antonello et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2021; Jain & Huth, 2018; Schrimpf et al., 2021; Toneva & Wehbe, 2019). These findings suggest that the transformer models exhibit remarkable correlations with fMRI data during speech comprehension, oftentimes referred to as their "brain score". In support of the predictive encoding theory, it has been specifically suggested that the depth of Transformers' predictive representations was organized hierarchically in the brain: low-level predictions most effectively modeled the superior temporal sulcus and gyrus, whereas high-level predictions were best at modeling the middle temporal, parietal, and frontal regions (Caucheteux et al., 2023). While the majority of the existing literature treated transformer models in their entirety, one distinct study has conducted a detailed examination of the intricate inner architecture of the transformer model, BERT (Kumar et al., 2022). They focused not only on the widely studied layer-wise embeddings but also the largely overlooked "transformations"—the computations performed by the attention heads. Using fMRI data obtained while participants listened to naturalistic storytelling, these transformations were found to encapsulate a hierarchy of linguistic computations across the cortex. Specifically, transformations occurring at later layers corresponded to higher-level language areas within the brain (e.g. angular gyrus and

inferior frontal gyrus). Their findings lay a new groundwork for leveraging the internal structure of LLMs to unravel the cascade of cortical computations in natural language comprehension, and they provided an essential reference for our study here.

Nevertheless, the newest studies have engendered skepticism regarding the predictive power of LLMs. The first skepticism concerning LLMs' "brain scores" questions whether LLMs trained on bare text could simulate our brain activations that incorporate and integrate information across multiple sensorimotor modalities. From a functional neuroanatomy perspective, human language comprehension involves the participation of modality-specific sensory, motor, and emotion systems, as well as a large number of brain regions that are not modality-specific (multi- or supra-modal, while integrating information from unimodal cortices; Binder & Desai, 2011). If LLMs are trained on language without sharing the embodied reality on which language is based, can they still be successfully synced to human language processing? This skepticism has been supported by a recent study comparing the predictive performance of LLMs and the "psychologically plausible" models in language comprehension at both word and discourse levels. Results suggested that the latter outperformed the LLMs in the prediction of brain activity, in particular within semantic-related areas (Zhang et al., 2024). Moreover, another study expressed concerns regarding the appropriateness of using predictive coding as the foundational theoretical framework in this field (Antonello & Huth, 2024). The authors demonstrated that representations that are best at predicting future words were strictly worse brain models than other representations. Based on this, the study argues that the strongest evidence in favor of predictive coding from encoding model research would likely remain valid even without predictive coding. In addition, another view cautions against some problematic procedures in the methodology and the issue of over-interpretating LLMs' "brain-scores". Replication was conducted by Feghhi et al. (2024) re-analyzing the same datasets as in Schrimpf et al.' study (2021). A problem with the shuffled split-test set was detected, along

with the unexpectedly high performance of superficial linguistic features and non-contextual embeddings. Their results raise questions about the current method and challenge the belief that LLMs and the biological brain share similar computational principles with natural language processing.

With these mixed views and evidence on the predictive power of LLMs in brain functions, exploring the generalizability of the current results and method of mapping LLMs representations onto brain activations becomes essential.

In this study, we aimed to replicate a previous study to confirm the predictive power of LLMs in brain functions and the generalizability of the results and methodology on a different dataset. Given the important role that LLMs' internal computational architecture plays in elucidating the relationship between artificial and biological computation, we specifically chose the study by Kumar et al. (2022) for replication due to its pioneering comparison of transformations with layer-wise embeddings. We sought to replicate this study using an independent fMRI dataset, which involved neurotypical individuals listening to naturalistic storytelling. In the workflow of the reference study, the authors extracted three BERT features: layer-wise embedding, head-wise transformations, and transformation magnitudes. Banded ridge regressions were employed as encoding models to evaluate the performances of Glove, traditional linguistic features, and Transformer components in predicting fMRI data acquired during speech comprehension. Results revealed that both BERT embeddings and transformations outperformed the non-contextual embeddings of GloVe and traditional linguistic features in most language-related regions. In addition, both BERT layer-wise embeddings and transformations displayed different layer-wise preferences, showing layer specificity in the prediction of brain language processing. Specifically, transformations captured a hierarchy of linguistic computations across the cortex, with later-layer transformations mapping onto higher-level language regions in the brain. Based on these results,

in our replication attempt, we focused on four major questions: Can BERT-based representations predict brain functions? To what extent and with what pattern do the predictions match the brain's hierarchical processing of speech? Does BERT's layer-vs-head specificity display a difference as discovered in the previous study? Do NSP-trained and context-sensitive LLMs predict better than non-predictive-encoding-based and static models?

We adopted the same model-based encoding framework as in the reference study (Naselaris et al., 2011; Richards et al., 2019; Yamins & DiCarlo, 2016). Given that we are employing LLMs to predict brain activations during language comprehension, the biological plausibility of these models and the methodological details are extremely crucial. A few improvements were therefore considered appropriate and timely: First, we replaced GloVe with fastText to circumvent the out-of-vocabulary issue, since it breaks down words into character n-grams, which can generate more accurate embeddings for words that were not seen during the training. Secondly, to construct a more realistic context that better fits the actual training of the models, we took a complete utterance as context and generated embeddings based on individual words instead of averaging the words that occurred in the same time repetition (TR) and taking the preceding 20 TRs as context (which doesn't reflect the context as complete utterances). Finally, cumulative word embeddings were added to the comparisons as a way of potentially tracking the buildup of phrasal complexity in the brain—perhaps the most essential linguistic process, which we expected to correspond to the cortical processing differences that we would hope our computational models to pick up.

Based on the findings of our reference study, firstly, we predicted that the average intersubject correlation (ISC) scores would be higher in the primary auditory cortex and lower in high-level language regions. Secondly, the overall above-zero performances and a hierarchy of linguistic computations across the cortex were anticipated, with later-layer transformations yielding better predictions in higher-level language areas in the brain. Thirdly, we expected that

the cumulative word embeddings would display different predictions from the isolated word embeddings, and exhibit an overall better performance in language-related regions of interest (ROIs). Furthermore, the performances of layer-wise embeddings and transformations were anticipated to peak in different layers with localization of the language-related ROIs. Lastly, we predicted that NSP-trained and context-sensitive BERT features would outperform the non-predictive-encoding-based static fastText embeddings in ROIs related to language processing.

2. Methods

Dataset

The fMRI dataset was obtained from the publicly available "The Alice Dataset" (Bhattasali et al., 2020) at https://openneuro.org/datasets/ds002322/versions/1.0.4. Twenty-six subjects participated in the fMRI session (15 females and 11 males, aged 18–24). All qualified as right-handed and self-identified as native English speakers. The audio stimuli were Kristen McQuillan's reading of the first chapter of Lewis Carroll's *Alice in Wonderland,* which contains 4202 words in total and lasted 12.4 minutes. The fMRI data was acquired using a 3T MRI scanner (Discovery MR750, GE Healthcare, Milwaukee, WI) with a 32-channel head coil at the Cornell MRI Facility. Thirteen participants were scanned using a T2-weighted echo planar imaging (EPI) sequence with a repetition time of 2000 ms, echo time of 27 ms, flip angle of 77, image acceleration of 2X, field of view of 216 216 mm, and a matrix size of 72 72. Sixteen participants were scanned with a three-echo EPI sequence where the field of view was 240 240 mm resulting in 33 slices with an in-plane resolution of 3.75 mm2 and thickness 3.8 mm. Data from this second group were from the second EPI echo, where the echo time was 27.5 ms.

fMRI preprocessing

Images were preprocessed using fMRIPrep 21.0.1 (Esteban et al., 2019), based on Nipype 1.6.1 (Gorgolewski et al., 2011). The functional images were spatially standardized with

MNI152NLin2009cAsym space, and all confounds were regressed based on the fMRIprep output (Wang et al., 2024). Confound regression followed the standard pipeline as defined by Wang et al., where we regressed out the effect of head motion, white matter, and cerebrospinal fluids, and added discrete cosines transformation basis regressors to handle low-frequency signal drifts. All functional data were mapped to a 1000-parcel cortical parcellation derived from intrinsic functional connectivity (Schaefer et al., 2018). In addition, the study extracted time series from thirteen ROIs whose functions are related to language processing in the brain, using the Havard-Oxford probabilistic parcellation (max probability of 50%). These regions include twelve cortical regions and one subcortical region, spanning from middle frontal gyrus (MFG), inferior frontal gyrus, triangular part (IFGtri), inferior frontal gyrus, opercular part (IFGope), temporal pole (TP), posterior superior temporal gyrus (pSTG), posterior middle temporal gyrus (pMTG), supramarginal gyrus (SMG), angular gyrus (AG), medial prefrontal cortex (mPFC), precuneus (Prec), parahippocampal gyrus (PHG), Heschel's gyrus (HG) to hippocampus (Hipp). All the time series for the selected ROIs above were averaged between the left and right hemispheres. In particular, the ones for SMG and PHG were averaged between the anterior and posterior parts.

Utterance division

The text for the audio presentation was downloaded from https://www.cs.cmu.edu/~rgs/alice-I.html, accessed on Oct. 23rd, 2023. Words without TR (e.g. 's) were mixed with the previous words, as they are typically abbreviations. This study defines an utterance as a syntactically independent unit that provides new semantic information. For instance, the following sentence "Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do:/ once or twice she had peeped into the book her sister was reading, /'but it had no pictures or conversations in it, and what is the use of a book,' thought Alice 'without pictures or

conversation?" was divided into three independent segments. The entire text contains 304 utterances after the annotation.

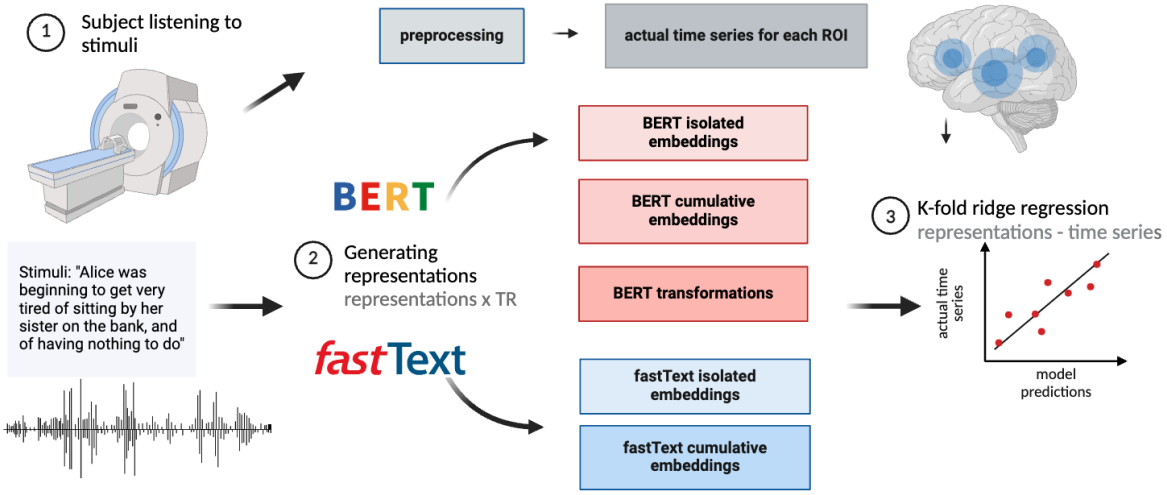

Figure 1: General workflow. (1) After preprocessing the fMRIs of subjects listening to fMRIs, time series in each ROI were extracted. The stimuli used for the storytelling were divided into different utterances. (2) Five representations were used to predict parcelwise fMRI time series. Apart from the static embeddings (fastText features. Blue), we also extracted the contextual embeddings and transformations from BERT (red). (3) The encoding model was estimated from a training subset using ridge regression and evaluated on a left-out test segment using K-fold cross-validation.

FastText and BERT embeddings

The general workflow can be seen in Figure 1. Firstly, two different models were employed for generating the static and context-sensitive embeddings, respectively. A pre-trained fastText word vector model (cc.en.300.bin) served as a model for non-contextual, global word embeddings (Bojanowski et al., 2017). Regarding contextual representations, we took an off-the-shelf model (BERT-base-uncased) from the HuggingFace library (Wolf et al., 2020) to generate word embeddings considering specific contexts. The fastText model represents words as continuous vectors based on their overall usage patterns in the training corpus. Hence, it is

characterized as providing static embeddings that capture the conceptual meaning of words, which is effective for encoding broad and generic semantic relationships but not for differentiating between different senses within specific contexts. This means that "bank" will have the same vector regardless of whether it appears in "I went to the bank to deposit some money" or "There are some willows on the river bank." Unlike fastText, BERT uses a deep transformer model that processes text bi-directionally. Namely, it examines the entire context of a word by considering both the preceding and succeeding words in a sentence. Therefore, BERT generates contextualized embeddings that capture the referential meaning of words by considering their surrounding vocabularies within the attention window. In this case, BERT can generate different vectors for "bank" in the context of a financial institution and a riverbank. This characteristic allows it to understand referential meanings, which was hypothesized to yield higher correlations to brain activations during language comprehension.

In retrieving word embeddings, the fastText model processes each utterance as a sequence of individual words, generating a static 300-dimensional isolated word embedding for each word. To proceed with our pre-trained BERT model, the annotated text was at first segmented into lists of utterances, enabling the entire utterance to serve as the context for the given word. The isolated layer-wise embeddings were obtained by passing the tokenized input through BERT, which returns hidden states from all layers. Each tensor has the shape of (number of layers, number of words, 768), which comprises a list of layer embeddings with 768 dimensions. These hidden states were extracted, excluding the initial encoding embedding layer. Each layer's embeddings were then aligned with the original tokens by averaging the embeddings of the corresponding sub-tokens.

Cumulative embeddings

While the isolated word method takes an individual word as an isolated unit (similarly to the reference study), in the cumulative method, embeddings of the current word and preceding

words in the same utterance were aggregated and averaged to track the buildup of syntactic (phrasal) information and enhance the biological plausibility in language comprehension. For instance, in the utterance "Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do", this cumulative feature will generate embeddings based on word chunks "Alice", "Alice was", "Alice was beginning" and "Alice was beginning to." Each chunk will yield a single embedding acquired by averaging embeddings of all the given words as one unit. In this approach, the fastText model first reads each cumulative unit, and then aggregates the embeddings and computes the mean to represent the given phrase. Analogous to the cumulative method applied to the fastText model, BERT cumulative word embeddings were also obtained by taking the average of the isolated embeddings in its current word chunk.

Transformations

While the layer-wise embeddings encapsulated the contextual linguistic content of each word or phrase, the head-wise transformations served as a unique component of the circuit that allows information to flow between words, which may represent information compression in the process (Cheng et al., 2023). To understand how the Transformer's inner structure maps onto brain activations during speech comprehension, it is important to consider both the layer-wise embeddings and the head-wise transformations. In BERT, each attention head has its unique set of query, key, and value matrices. After computing the attention scores, the outputs of all heads were concatenated and linearly transformed. This combined output was then added back to the input of the layer via a residual connection, allowing the model to focus on different parts of the input sequence (Vaswani et al., 2023). From this process, transformations were captured by extracting the combined output before it passed through the feedforward layer (MLP). In this study, alongside the twelve layer-wise embeddings extracted from the hidden layers, we computed eleven transformations between the twelve layers by acquiring the attention outputs for the given layer embeddings across the attention heads. The attention

output tensor, shaped (number of heads, sequence length, sequence length), reveals how much focus each word in the sequence has on every other word across all attention heads. The tensor's first dimension corresponds to the different attention heads, while the following dimensions form a matrix that represents the attention scores between these words. Eventually, we extracted eleven transformations between the twelve layers and aligned the matrix to the original tokens, which were stored for the following ridge regression.

Predictions and evaluations of representations

Representations' predictions were estimated using ridge regression. Ridge regression is a type of linear regression that includes a regularization term in the loss function that helps to prevent overfitting by shrinking the coefficients. Its efficiency in handling multicollinearity and managing high-dimensional data makes it suitable for encoding the embedding vectors and fMRI time series (Nunez-Elizalde et al., 2019). To estimate representations' predictions, both word-level embeddings and the fMRI time series were aligned to the corresponding word offsets using the canonical hemodynamic response function (HRF) as the encoding method (Pasquiou et al., 2022). To yield a more bias-free evaluation, we applied the K-fold cross-validation scheme. In this approach, during each iteration of the cross-validation process, one fold of the data was set aside for testing while the remaining two folds were used for training the model. This process was repeated three times, ensuring that each part of the data was serving as a test set exactly once, yielding a correlation value for each test set of the outer cross-validation loop for each parcel and subject. These raw cross-validation scores for all five representations were then saved respectively, allowing for the following statistical evaluations of model performances.

Computing intersubject correlation

To facilitate the interpretation and better compare the replicated results, we adopted the percent of noise ceiling (PNC) approach from the reference study (Kumar et al., 2022). Firstly, the

noise ceilings were calculated based on the inter-subject correlation (ISC, Nastase et al., 2019). To calculate the ISC, the average time series data across all subjects, excluding the current subject, were computed to serve as the held-out average. In order to match the raw scores from the regression, we set up the same K-fold cross-validation where in each iteration, one part was used as the test set, and the remaining two parts were used for training. For each partition, the Pearson correlation coefficient between the subject's test-set time series and the held-out average time series was calculated as the "noise ceiling". This coefficient measures the linear relationship between the individual and the average group time series data—that is, the theoretically "optimal" response for the given subject, providing a metric of how deviant this subject's brain activations were from the average. Later, the correlation scores for each ROI were averaged across the three splits, resulting in a single ISC value as the optimal performance in each parcel for the current subject.

Percent noise ceiling estimation

Leveraging the ISC, the PNC was calculated by dividing the Fisher transformed raw scores (RS) by the corresponding noise ceiling for each subject, which denotes the proportion of explained variance relative to the total variance available. To achieve a more intuitive presentation, these proportions were usually visualized and mentioned in the study without "%". Extreme values (>100 or < -100) were truncated into 100 or -100 for better interpretability.

Statistical assessment

Given the small size of the Alice dataset, we also reapplied the two non-parametric methods from Kumar et al.'s paper (2022). After the ridge regression, bootstrapping tests were conducted to compute *p* values for each ROI, followed by false discovery rate (FDR) correction. Initially, to stabilize the variance of correlation coefficients, the raw scores were normalized using the Fisher transformation and then averaged as the observed mean. Next, to center the data around zero, this observed mean was subtracted from each data point. The function then

created 1000 bootstrap samples by selecting values from the centered data with replacement and calculating the mean of each sample. Lately, the *p* value was determined by calculating the proportion of these bootstrap sample means that are greater than or equal to the observed mean. To guard against false positives, Benjamini-Hochberg *p*-value correction was applied to the *p*-values resulting from the bootstraps. We calculated the step-up values using the total number of hypotheses divided into the ranks of sorted *p* values in ascending order. Later, the adjusted *p* values were obtained by computing the product of the step-up values and the sorted *p* values, then taking the cumulative minimum of these products. Each value was then capped at a maximum of 1. Finally, the adjusted *p* values were reordered to match their original sequence and then returned.

In the model-performance comparison, the pair-wise permutation test was applied for significance testing between the different prediction scores. Firstly, the original test statistic was calculated as the absolute value of the median difference between the two sets of representation results. To generate 1000 permutations, each element in the differences array was randomly multiplied by either 1 or -1 with replacement, and the median of these permuted differences was computed. We then evaluated whether the original test statistic exceeded all the permuted test statistics. If the original test statistic was greater than all permuted values, a *p* value of 0.001 was assigned, indicating a highly significant result. Otherwise, the *p* value was computed as the proportion of permuted test statistics that are greater than or equal to the original test statistic, divided by the total number of permutations. Again, we applied the same FDR procedure as in bootstrapping to adjust the p-values derived from the permutation test.

3. Results

Group ISC distribution

The ISC results are displayed in Figure . As seen there, the primary auditory cortex yielded the highest ISC scores, approximately 0.459. In addition, in the high-level language function areas, notable but lower correlations were observed along the STG (anterior and posterior), MTG, IFG, AG, and medially in the precuneus. Overall, the ISC effectively mapped the brain regions involved in speech processing and comprehension, confirming our first prediction.

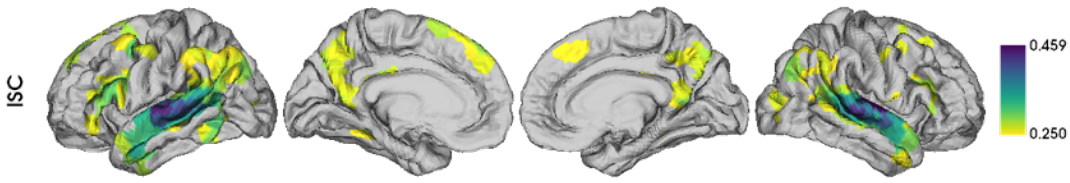

Figure 2: Average ISC across all subjects. The ISCs were aggregated and averaged across all 26 subjects. Dense color indicates higher correlations.

Performance of layer-wise BERT features and fastText in the whole brain

We next compared the whole-brain performance of fastText and of BERT features across its separate layers (Figure). Two aspects were investigated here: the extent to which LLMs display predictive power regarding brain activations, and whether the specificity of layers and heads within the transformer structure could offer a detailed localization of cortical linguistic processing. We found that, across all models, the LLMs' PNC reached its upper limit at 38, which was reflected in the correlation coefficient as a top score of 0.076. While it was confirmed that the PNC roughly matched the result from Kumar et al.'s research, the raw correlation coefficients revealed a low performance across all five representations. On top of that, contrary to the expectation based on the reference study (Kumar et al., 2022), significant predictions were distributed very scarcely and sparsely across the brain. This shows no overall interpretable pattern, whether in cumulative and isolated word embeddings from the fastText

model (Figure A and 3B), or in the case of the three BERT features (Figure A, 3B, and 3C). Within the 1000 ROIs, as indicated by Figure A and 3B, subareas of some high-level language-processing ROIs like MTG (isolated word embeddings, 3rd, 4th, and final layers), AG (cumulative word embeddings, last layer), and IFG (cumulative word embeddings, 2nd layer) were localized by BERT, while STG was predicted by mainly second and third layer in transformations (Figure C). When looking at the layer-wise performance, the isolated word embeddings were sparsely distributed with no clear tendency, which contradicts the monotonic progression pattern reported in the reference study. In the case of transformations, instead of peaking at intermediate layers with reduced performance for the final layers as discovered in the previous study, slightly more concentrated predictions were found in both early and later layers. Together, the whole-brain PNC and RS provide no strong evidence to support that the layer-wise performance of these features cleanly maps onto the cortical hierarchy for language comprehension.

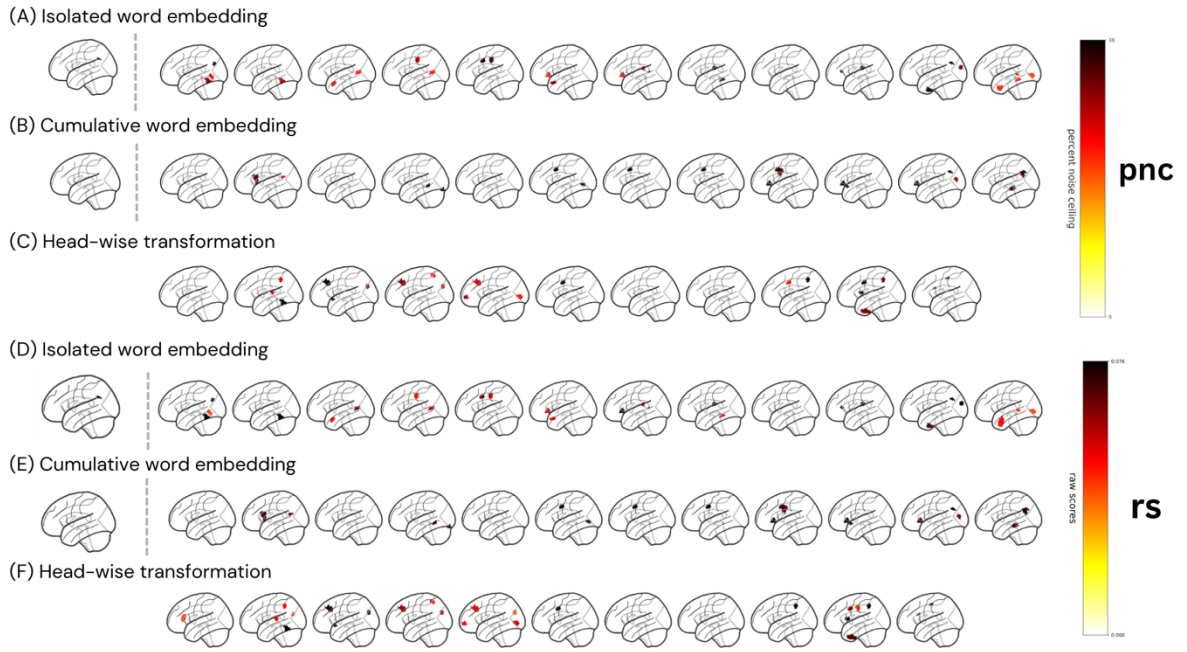

Figure 3: Layer-wise performance of the five representations across all cortical parcels. Percent noise ceiling (PNC) is noted as "pnc", while raw scores (RS) are noted as "rs" on the right side. PNC scores that are lower than 0 were truncated for better interpretability. (A) Layer-

wise PNC of isolated word embeddings (fastText embeddings are the first graph before the dotted line, and BERT embeddings of 12 layers are displayed in the section after the dotted line). (B) Layer-wise PNC of cumulative word embeddings (the fastText-BERT layout is the same as in A). (C) Layer-wise PNC of BERT transformations. There are only 11 transformations because they were computed between layers. (D) Layer-wise RS of isolated word embeddings (fastText before the dotted line, and BERT embeddings of 12 layers after it). (E) Layer-wise RS of cumulative word embeddings (the fastText-BERT layout is the same as in D). (F) Layer-wise RS of BERT transformations. Both cortical maps display the left hemisphere and were thresholded to display only ROIs with statistically significant model performance (non-parametric bootstrap hypothesis test; FDR controlled at $p < .05$).

Additionally, while PNC was employed as a metric to enhance interpretability and reduce potential noise from individual-subject differences, the RSs were incorporated to exhibit a more direct correlation and complementary view of model performances. When comparing PNC and RS, the performance of the fastText representations and BERT isolated word embeddings remained consistent across both methods (Figure A, 3B, 3D, and 3E). Nevertheless, in BERT cumulative word embeddings, performance varied in several regions (Figure B and 3E). Notably, in the transformations, there were remarkable discrepancies in scores between the two methods in both early and later layers, with RS revealing more predicted regions (Figure C and 3F).

Performance of fastText and BERT layer-wise features in selected ROIs

To further probe into the models' performance and layer-to-brain localization, we examined the PNC of all the models in pre-selected regions associated with speech processing (Figure). Considerable variance within all model performances was observed, with the scores ranging predominantly from -20 to 20, without any consistently high performance for any of the five representations. Overall, no clear patterns or layer progression could be concluded across the

thirteen selected ROIs for all three BERT features, which aligns with our results from the performance of layer-wise BERT features in the whole brain.

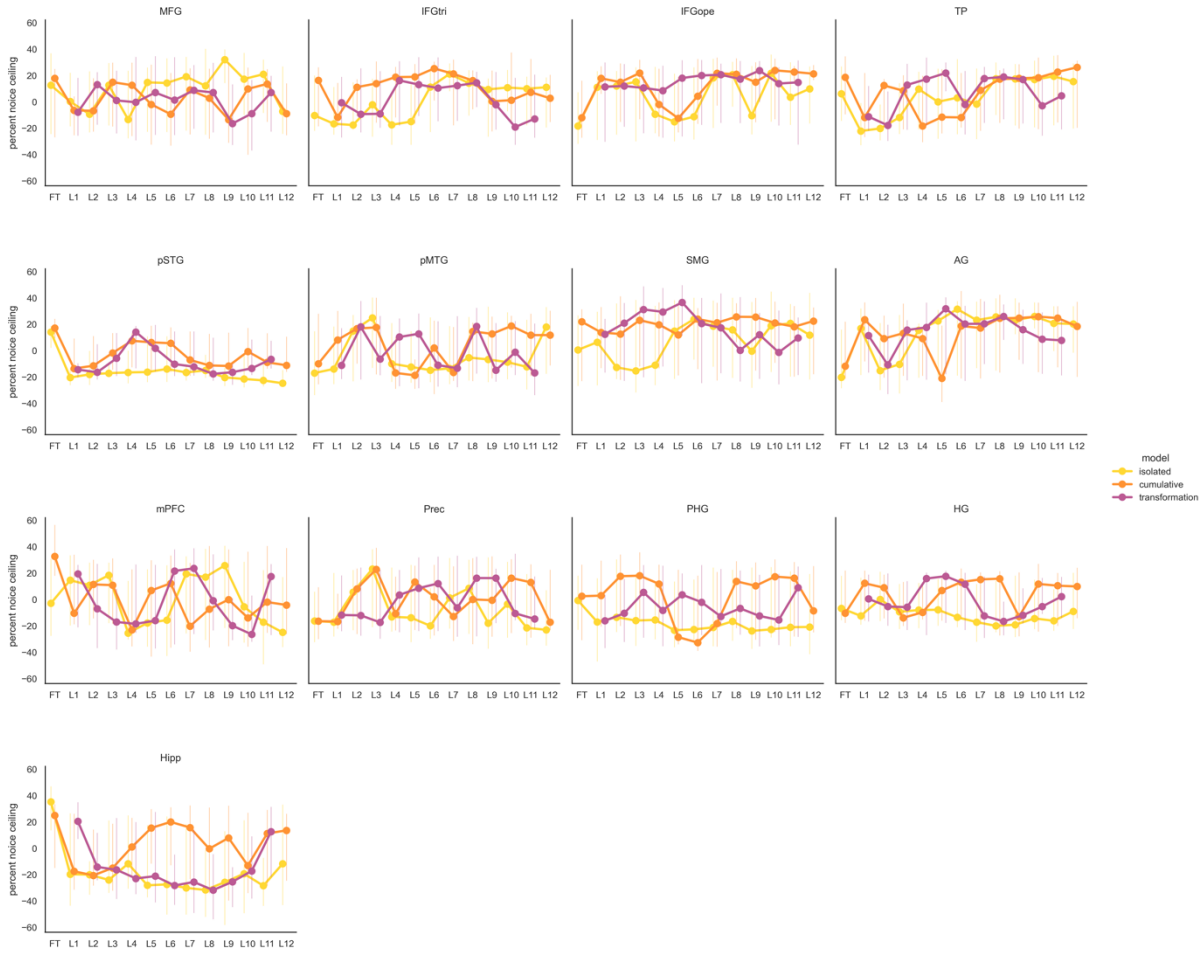

Figure 4: Layer-wise model performance for isolated word embeddings (yellow), cumulative word embeddings(orange), and transformations (purple) in thirteen ROIs. FastText model's two representations were plotted as the first two dots to facilitate comparison (FT on the x-axis). Performance of BERT layers and transformations were plotted layer-wise (L1-L12 on the x-axis for embeddings, and L1-L11 for transformations). Markers indicate median performance and error bars indicate 95% bootstrap confidence intervals.

BERT layer preference in the whole brain

Pursuing the question of BERT layers and their brain predictions, we visualized layer preference across all ROIs. As shown in Figure A, isolated word embeddings revealed a

progression starting from early layers in the primary auditory cortex and extending to later layers toward higher language regions, including STG, MTG, IFG, inferior parietal cortex (IPC), as well as precuneus, more prominently on the left. In juxtaposition, transformations followed a similar pattern. However, the cumulative word embeddings exhibited a partially reversed order of the two, which produced a preference for late layers in the primary auditory cortex and for early layers in IFG, MTG, and AG.

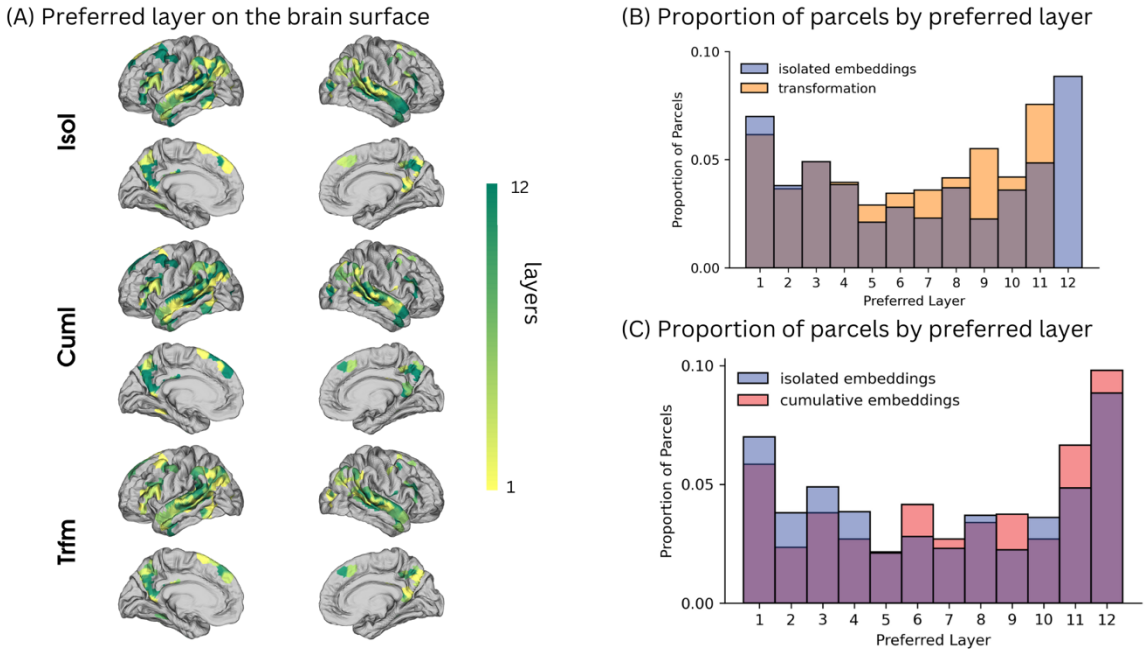

Figure 5: Layer preferences for BERT embeddings and transformations. (A) Layer preferences on the cortical surface for BERT isolated embeddings (Isol), cumulative (Cuml), and transformations (Trfm). Only ROIs with ISC greater than 25% of the PNC for both embeddings and transformations are included for visualization purposes. For each ROI, the layer with the higher PNC was plotted as the preferred layer. Lighter colors represent early layers while darker colors indicate the later ones. (B) Histogram of the preferred layer across 1000 ROIs. The proportion of parcels depicted in the histogram corresponds to the percentage of ROIs where this specific layer achieved the best prediction performance. Both performances for BERT transformations (yellow) and isolated word embeddings (blue) show a U-shaped curve. (C)

Histogram of the preferred layer across 1000 ROIs. Performances for isolated (yellow) and cumulative word embeddings (red) also display a similar U-shaped distribution.

Next, to analyze layer performance specificity, we determined which layer performs best across the brain by computing the proportion of parcels where each layer attained its optimal prediction. The isolated embeddings and transformations depicted a peak performance at the first layer and the last layer (Figure B), which yielded different results from embeddings being preferred in the later layers and transformations in the intermediate layers from the reference study (Kumar et al., 2022). Meanwhile, the cumulative word embeddings displayed a similar tendency as the other two, despite having the strongest preference for the last layer and more variance in the intermediate layers (Figure C).

Model comparison across selected ROIs

Finally, to assess the overall performance of the five different representations, a comparison was conducted between the performances of the two fastText representations and the averaged PNC of the three BERT features in the selected language-related ROIs. Results are presented descriptively in Figure . Throughout all the ROIs, results broadly indicate low-performance levels; however, it is notable that in TP and SMG, all five representations yielded a positive PNC, perhaps reflecting the mapping of brain function in terms of semantic memory and phonological processing, respectively (Figure).

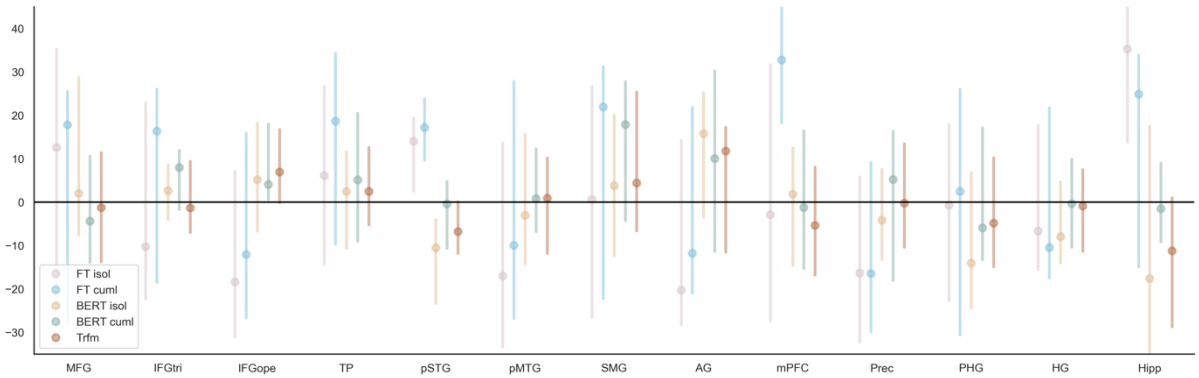

Figure 6: Representation performance across selected ROIs. We use PNC estimated using intersubject correlation to evaluate the prediction performances of five representations: fastText isolated word embeddings (FT isol), fastText cumulative word embeddings (FT cuml), BERT isolated word embeddings (BERT isol), BERT cumulative word embeddings (BERT cuml), and BERT transformations (Trfm). To summarize the overall performance of these different representations, the BERT features across all heads and layers were averaged. Markers indicate median performance and error bars indicate a 95% bootstrap confidence interval.

When comparing the representations' performance pair-wise, a deviation from the reference study and prior works (Schrimpf et al., 2021; Goldstein et al., 2022; Kumar et al., n.d.) becomes apparent once more. BERT features, which encoded contextual meanings in contexts, did not differ significantly from the static fastText representations in most language-related regions ($p > .05$ in MFG, IFGtri, IFGope, TP, pMTG, SMG, Prec, PHG, and HG; permutation test; FDR corrected; Figure). Significantly, both fastText isolated embeddings and cumulative embeddings outperformed BERT isolated embeddings in pSTG (Figure). Additionally, the performance of fastText isolated embeddings also substantially exceeded transformations in this region. In the Hippocampus, the prediction of fastText isolated embeddings topped all three Transformer features, remarkably the BERT isolated embeddings, with a significant median difference of 52.86. Likewise, in mPFC, the fastText cumulative embeddings outperformed all three BERT representations. Of note, the only region where BERT isolated embeddings and transformations significantly outperformed the fastText isolated embeddings was AG. Looking within the BERT features, no significant differences emerged between isolated word embeddings, cumulative word embeddings, and transformations.

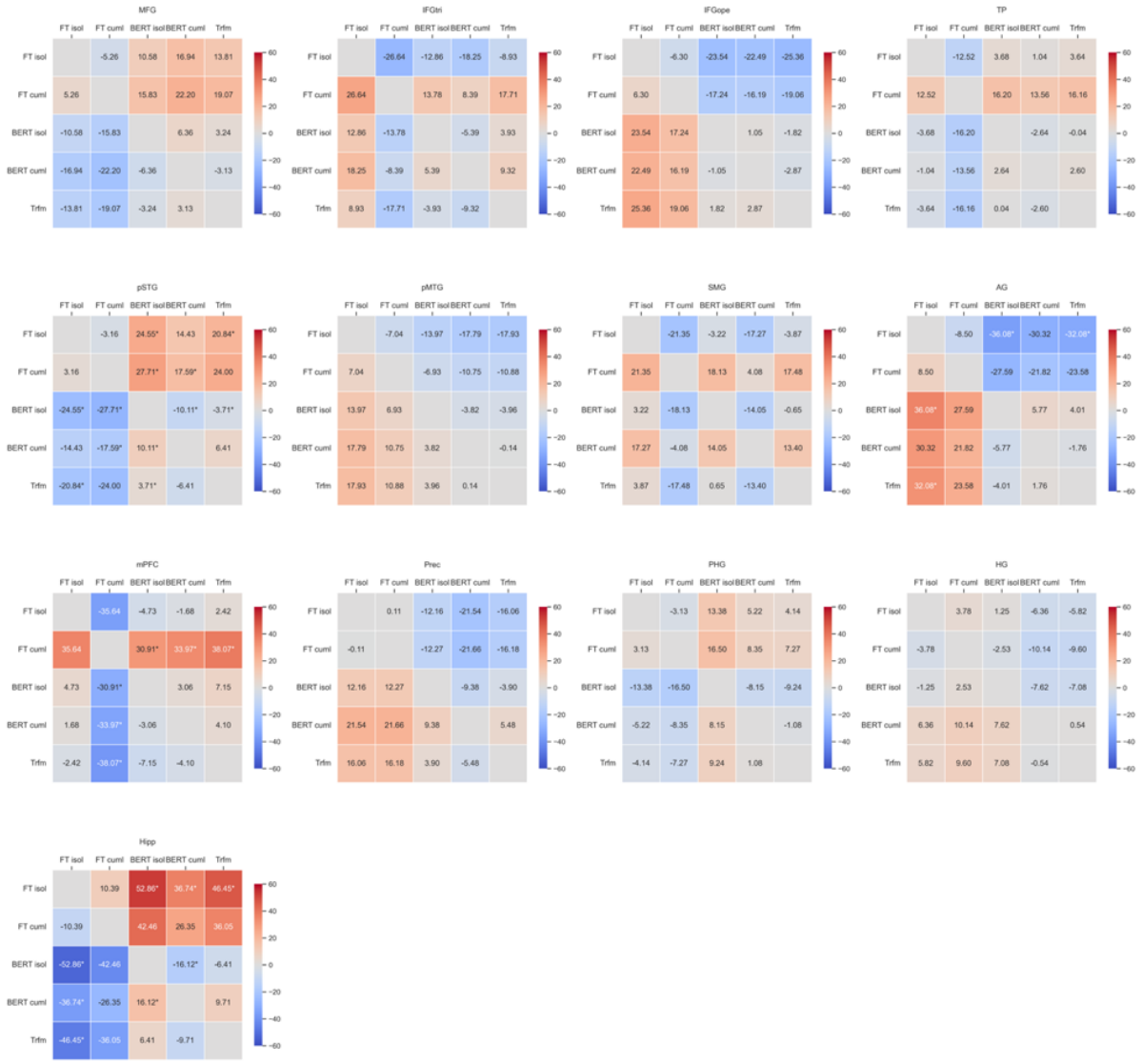

Figure 7: Heat map of significant ROIs. Row representations are compared with column ones. Median differences in PNC are illustrated in the middle, with the significant results marked with *. Red squares indicate positive differences while the blue squares indicate negative differences. The larger the median difference, the denser the color. (Permutation test, FDR $p<0.05$).

4. Discussion

Reflecting on the initial questions that we posed: Can BERT-based representations predict brain functions? To what extent and with what pattern do the predictions match the brain's

hierarchical speech processing? Does BERT's layer-vs-head specificity display a difference as discovered in the previous study? Do NSP-trained and context-sensitive LLMs predict better than non-predictive-encoding-based and static models? Our findings provided nuanced insights. Firstly, in contrast to the reference study, the overall BERT feature performances exhibited a scarce and scattered distribution across the brain and a high variance in language-related ROIs**.** On top of that, the highest raw correlation coefficient was less than 0.08, which raises questions about the BERT's predictive power of brain function. Secondly, although the whole-brain prediction distribution did not align with the brain's speech-processing hierarchy, early layers of BERT's isolated word embeddings and transformations were preferred by the primary auditory cortex, while later layers were favored by the STG, MTG, IFG, and IPC, suggesting a rough mapping to the cortical speech processing hierarchy. Moreover, against our prediction, although results showed that cumulative word embeddings had partially reversed prediction orders compared to isolated word embeddings, there were no significant differences in performance or layer-specificity between the two. Furthermore, opposite to what was anticipated, the performance of layer-wise embeddings and transformations exhibited similar layer preferences and demonstrated no layer-specificity in the localization of language-related ROIs. Unexpectedly, the NSP-trained and context-sensitive BERT features didn't outperform the non-predictive-encoding-based static fastText embeddings in most language-related ROIs.

There are several potential explanations for the lower brain scores and sparser distribution observed compared to the reference study. Firstly, the reference study had 63 subjects, while our study used a smaller dataset of 26 subjects. One issue that arose from using a smaller dataset was the appropriateness of the bootstrapping method. It has been evidenced from the previous study that standard bootstrapping cannot overcome the intrinsic challenge of underestimating the mean in small samples with high log variance (Mostofian & Zuckerman, 2019). While bootstrap-based methods perform well with relatively large samples, they exhibit

biases in small sample settings, with the directions and sizes of these biases varying inconsistently (Iba et al., 2021). That being said, certain biases derived from the test may change the outcomes due to a smaller data size. Furthermore, our results showed that negative correlation coefficients have also exceeded the threshold, indicating a significant negative relationship between model predictions and actual brain responses, which contradicts the purpose of employing such a method. These two flaws call into question the methodological appropriateness of the statistical approaches used in the reference research.

Moreover, in the ridge regression, we implemented the hemodynamic function from Pasquiou et al. (2022) to handle the hemodynamic lagging. In the approach used by Kumar et al., confound variables such as the silence indicator vector, word count and phoneme count vectors, and the phoneme indicator matrix were assigned as “bands” in the banded ridge regression. To capture the delayed response, specific time points (1.5, 3.0, 4.5, and 6.0 seconds) were selected. They then made multiple copies of each feature in the data, including BERT features and the confound variables to represent the feature values at these various lag times. These lagged features were then concatenated to form an extended feature set that contained both the original and all lagged versions of each feature. Next, the ridge regression model was fitted using these extended feature sets, selecting lags that provided the best predictions based on the highest scores. However, since hemodynamic lag is a physiological characteristic inherent to each brain region, adjusting it to fit the data raises concerns about the validity of its predictions. Consequently, selecting the lag with the best scores could be another reason that accounts for the better model performances in the reference study.

In addition, another methodological issue involves the PNC. Being computed by the ISC and RS, the proportion, as a metric for enhancing interpretability, could sometimes present a distorted view of reality. For instance, if the RS and ISC are both negative values, the PNC could produce a positive value, which incorrectly reflects a misleading positive correlation

between the model prediction and brain activation. This could potentially explain the discrepancy between the RS and PNC from our results. Hence, we maintain a cautious stance on the reliability of this metric and the potential for false discoveries in Kumar et al.'s study.

Procedure-wise, we improved in setting the BERT's context window based on the reference work. Initially, word embeddings were averaged for each TR, using the preceding 20 TRs as context. However, this method risks splitting words from the same utterance across different segments, which is biologically unrealistic and contradicts the NSP task that BERT was trained on. BERT, having learned from complete utterances, is optimized to work with entire sentences, which allows it to fully utilize its design and generate embeddings that capture the complexities of natural language. Despite these improvements, our predictions were still less significant than those in Kumar et al.'s study, again, raising concerns about potential false discoveries in their findings.

Looking into the performances of BERT's internal components, only the layer preference in the brain surface reflects the same hierarchical localization as in human language processing. Notably, although the cumulative word embeddings displayed a partially reversed layer preference order compared to the other two BERT features and a slightly stronger preference for later layers, no significant overall better performances. This mismatch could potentially indicate that either the brain does not distinguish between words and phrases, or LLMs are not effectively capturing these patterns, which we speculated are more likely to be true. When comparing transformations and isolated embeddings, we found no significant differences in layer-wise preferences across both the whole brain and selected language areas. This lack of distinction raises questions about whether the functional specificity and accumulation of contextual information in these features resemble similar processes in the brain.

When comparing the overall performance between models across thirteen language-related ROIs, significant differences were observed in only four, suggesting no general significant difference in predictive power between fastText and BERT. Surprisingly, BERT features didn't outperform the static embeddings significantly in HG as in the reference study. Within these four ROIs, fastText isolated and cumulative embeddings both outperformed BERT isolated embeddings in the pSTG, which is crucial for recognizing and interpreting various auditory stimuli. We speculate that since fastText generates embeddings based on the full words as well as subwords or n-grams of characters, this method enables the model to capture more nuanced word meanings based on morphological similarities. This feature might be more effectively aligned with how auditory language processing occurs in the pSTG, especially in distinguishing between similar-sounding words with different meanings. While fastText unexpectedly surpassed BERT in the mPFC without clear reasons, in the memory-related hippocampus, fastText isolated embeddings remarkably outperformed all three Transformer features. This could be due to that fastText, trained on complete narratives, aligns more closely with the hippocampus's role in processing non-contextual memories, whereas BERT is more related to episodic memory through its training on masked words and sentences. Different from others, AG was the only region where BERT isolated embeddings and transformations significantly outperformed the fastText isolated embeddings. One possible explanation is that AG is involved in referential meaning processing, which gave an advantage to BERT's context-sensitive features rather than fastText. Overall, these findings did not provide convincing support to the notion that BERT's features effectively reflect the hierarchical processing of language.

Similar to the results from Feghhi et al. (2024) but opposite to Kumar et al.'s (2022), the NSP-pre-trained contextual embedding didn't outperform the non-predictive coding and static fastText embedding. This finding doesn't confirm the prevalent predictive coding theory,

nor provide any evidence to suggest the BERT's contextual architecture contributes to a better predictive power. While predictive encoding theory has been a popular hypothesis for many studies, it was not specifically validated by our findings. We therefore proposed several theories that might provide a better explanation for the limited predictive power of LLMs. One such alternative, the shallow brain hypothesis, suggested that the predictive coding models have largely overlooked the neurobiological evidence showing that all hierarchical cortical areas exchange signals with subcortical areas in parallel (Suzuki et al., 2023). Highlighting the computational capabilities of cortical microcircuits and thalamocortical loops, the shallow architecture challenges the dominant cortex-focused, hierarchical architectures in predictive coding networks as a rising new theory. Secondly, addressing the modality issues mentioned earlier, human language connects embodied reality with sentient actions and sensory information. Hence, linguists argue that LLMs, which are trained solely on text, model the brain's language organization poorly due to their lack of interaction with infant biology, personal experiences, unique principles, and natural laws that shape language development and use (Bever et al., 2023). Meanwhile, as evidenced by the comparative analysis conducted by Zhang et al. (2024), the psychological-plausible models with multi-modality outperformed the LLMs in predicting fMRI data in many language areas, which confirmed this fundamental shortcoming of current LLMs. Thus, our results underscored the need to rethink the grounding theory and LLM frameworks to incorporate more biological plausibility and multi-modal factors to better simulate human language processing in the brain.

The study has one significant limitation as mentioned by the reference work: although BERT displays a remarkable ability to understand text, its bidirectional design does not closely mimic the unidirectional nature of how humans typically process speech, which casts doubt on its biological plausibility. As a suggestion for future lines of research, apart from the implementation of more suitable non-parametric statistical tests, exploring multi-modal and

more biologically-inspired LLMs could better bridge the gap between the human brain and LLMs (Bever et al., 2023), and provide an understanding of not only the functional specialization but also the functional integration across specialized LLMs' components. In regard to theoretical implications, greater attention to subcortical areas should be addressed to encompass a more comprehensive understanding of brain functions (Suzuki et al., 2023). Building on our current findings, a comparative analysis exploring how LLM-generated word embeddings and sentence embeddings differently predict brain activations could be another interesting line of research.

Although our study highlights several methodological issues with using LLMs representations and does not produce solid evidence to support the predictive coding theory, mapping the internal structure of LLMs to cortical language processing provides a direct approach to enhance our mechanistic understanding of human language processing. This approach may ultimately offer valuable insights into understanding functional specialization in both LLMs and the brain.

References

Anderson, A. J., Kiela, D., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Raizada, R. D. S., Grimm, S., & Lalor, E. C. (2021). Deep Artificial Neural Networks Reveal a Distributed Cortical Network Encoding Propositional Sentence-Level Meaning. *Journal of Neuroscience*, *41*(18), 4100–4119. https://doi.org/10.1523/JNEUROSCI.1152-20.2021

Antonello, R., & Huth, A. (2024). Predictive Coding or Just Feature Discovery? An Alternative Account of Why Language Models Fit Brain Data. *Neurobiology of Language*, *5*(1), 64–79. https://doi.org/10.1162/nol_a_00087

Antonello, R., Turek, J. S., Vo, V., & Huth, A. (2021). Low-dimensional Structure in the Space of Language Representations is Reflected in Brain Responses. *Advances in Neural Information Processing Systems*, *34*, 8332–8344. https://proceedings.neurips.cc/paper_files/paper/2021/hash/464074179972cbbd75a39abc6954cd12-Abstract.html

Bever, T. G., Chomsky, N., Fong, S., & Piattelli-Palmarini, M. (2023). Even deeper problems with neural network models of language. *The Behavioral and Brain Sciences*, *46*, e387. https://doi.org/10.1017/S0140525X23001619

Bhattasali, S., Brennan, J., Luh, W.-M., Franzluebbers, B., & Hale, J. (2020). The Alice Datasets: fMRI & EEG Observations of Natural Language Comprehension. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 120–125). European Language Resources Association. https://aclanthology.org/2020.lrec-1.15

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). *Enriching Word Vectors with Subword Information* (arXiv:1607.04606). arXiv. https://doi.org/10.48550/arXiv.1607.04606

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385. https://doi.org/10.1017/S0140525X22002813

Caucheteux, C., Gramfort, A., & King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, *7*(3), 430–441. https://doi.org/10.1038/s41562-022-01516-2

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134. https://doi.org/10.1038/s42003-022-03036-1

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019a). What Does BERT Look at? An Analysis of BERT's Attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 276–286. https://doi.org/10.18653/v1/W19-4828

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019b). What Does BERT Look at? An Analysis of BERT's Attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 276–286. https://doi.org/10.18653/v1/W19-4828

De Vries, W., Van Cranenburgh, A., & Nissim, M. (2020). What's so special about BERT's layers? A closer look at the NLP pipeline in monolingual and multilingual models. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4339–4350. https://doi.org/10.18653/v1/2020.findings-emnlp.389

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Feghhi, E., Hadidi, N., Song, B., Blank, I., & Kao, J. (2024). *What Are Large Language Models Mapping to in the Brain? A Case Against Over-Reliance on Brain Scores*.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, S. C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2021). *Thinking ahead: Spontaneous prediction in context as a keystone of language in humans and machines* (p. 2020.12.02.403477). bioRxiv. https://doi.org/10.1101/2020.12.02.403477

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, *5*. https://doi.org/10.3389/fninf.2011.00013

Grodzinsky, Y., & Friederici, A. D. (2006). Neuroimaging of syntax and syntactic processing. *Current Opinion in Neurobiology*, *16*(2), 240–246. https://doi.org/10.1016/j.conb.2006.03.007

Hagoort, P., & Indefrey, P. (2014). The neurobiology of language beyond single words. *Annual Review of Neuroscience*, *37*, 347–362. https://doi.org/10.1146/annurev-neuro-071013-013847

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2021). *A hierarchy of linguistic predictions during natural language comprehension*.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119. https://doi.org/10.1073/pnas.2201968119

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113

Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *Wiley Interdisciplinary Reviews. Cognitive Science*, *2*(5), 580–593. https://doi.org/10.1002/wcs.142

Iba, K., Shinozaki, T., Maruo, K., & Noma, H. (2021). Re-evaluation of the comparative effectiveness of bootstrap-based optimism correction methods in the development of multivariable clinical prediction models. *BMC Medical Research Methodology*, *21*(1), 9. https://doi.org/10.1186/s12874-020-01201-w

Jain, S., & Huth, A. (2018). Incorporating Context into Language Encoding Models for fMRI. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper_files/paper/2018/hash/f471223d1a1614b58a7dc45c9d01df19-Abstract.html

Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of Language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. https://doi.org/10.18653/v1/P19-1356

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (n.d.). *Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model*.

Matchin, W., & Hickok, G. (2020). The Cortical Organization of Syntax. *Cerebral Cortex (New York, N.Y.: 1991)*, *30*(3), 1481–1498. https://doi.org/10.1093/cercor/bhz180

Mostofian, B., & Zuckerman, D. M. (2019). Statistical Uncertainty Analysis for Small-Sample, High Log-Variance Data: Cautions for Bootstrapping and Bayesian Bootstrapping. *Journal of Chemical Theory and Computation*, *15*(6), 3499–3509. https://doi.org/10.1021/acs.jctc.9b00015

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Nastase, S. A., Gazzola, V., Hasson, U., & Keysers, C. (2019). Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, *14*(6), 667–685. https://doi.org/10.1093/scan/nsz037

Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage*, *197*, 482–492. https://doi.org/10.1016/j.neuroimage.2019.04.012

Pasquiou, A., Lakretz, Y., Hale, J., Thirion, B., & Pallier, C. (2022). *Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps* (arXiv:2207.03380). arXiv. https://doi.org/10.48550/arXiv.2207.03380

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., …

Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. https://doi.org/10.1038/s41593-019-0520-2

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex (New York, N.Y.: 1991)*, *28*(9), 3095–3114. https://doi.org/10.1093/cercor/bhx179

Schmitt, L.-M., Erb, J., Tune, S., Rysop, A. U., Hartwigsen, G., & Obleser, J. (2021). Predicting speech from a cortical hierarchy of event-based time scales. *Science Advances*, *7*(49), eabi6070. https://doi.org/10.1126/sciadv.abi6070

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 of the United States of America*, *118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307

Suzuki, M., Pennartz, C. M. A., & Aru, J. (2023). How deep is the brain? The shallow brain hypothesis. *Nature Reviews Neuroscience*, *24*(12), 778–791. https://doi.org/10.1038/s41583-023-00756-z

Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the Classical NLP Pipeline. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4593–4601). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1452

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/749a8e6c231831ef7756db230b4359c8-Abstract.html

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., Kay, K., & Fedorenko, E. (2024). Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, *8*(3), 544–561. https://doi.org/10.1038/s41562-023-01783-7

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(3), 443–467. https://doi.org/10.1037/0278-7393.31.3.443

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Wang, H.-T., Meisler, S. L., Sharmarke, H., Clarke, N., Gensollen, N., Markiewicz, C. J., Paugam, F., Thirion, B., & Bellec, P. (2024). Continuous evaluation of denoising strategies in resting-state fMRI connectivity using fMRIPrep and Nilearn. *PLOS Computational Biology*, *20*(3), e1011942. https://doi.org/10.1371/journal.pcbi.1011942

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Yu, S., Gu, C., Huang, K., & Li, P. (2024). Predicting the next sentence (not word) in large language models: What model-brain alignment tells us about discourse comprehension. *Science Advances*, *10*(21), eadn7744. https://doi.org/10.1126/sciadv.adn7744

Zhang, Y., Wang, S., Dong, X., Yu, J., & Zong, C. (2024). *Navigating Brain Language Representations: A Comparative Analysis of Neural Language Models and Psychologically Plausible Models* (arXiv:2404.19364). arXiv. http://arxiv.org/abs/2404.19364