Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain

Authors: Ariel Goldstein[1,2,7*], Eric Ham[1,7], Samuel A. Nastase[1], Zaid Zada[1], Avigail Grinstein-Dabus[2], Bobbi Aubrey[1,3], Mariano Schain[2], Harshvardhan Gazula[1], Amir Feder[2], Werner Doyle[3], Sasha Devore[3], Patricia Dugan[3], Daniel Friedman[3], Michael Brenner[2,4], Avinatan Hassidim[2], Orrin Devinsky[3], Adeen Flinker[3,5], Omer Levy[6,8], Uri Hasson[1,8]

Affiliations:

[1] Department of Psychology and the Neuroscience Institute, Princeton University, Princeton, NJ

[2] Google Research

[3] New York University Grossman School of Medicine, New York, NY

[4] School of Engineering and Applied Science, Harvard University, Cambridge, MA

[5] New York University Tandon School of Engineering, Brooklyn, NY for me

[6] Blavatnik School of Computer Science, Tel-Aviv University, Israel

[7] Equal first author

[8] Equal senior authors

* Corresponding author. Email: ariel.y.goldstein@gmail.com

Abstract

Deep language models (DLMs) provide a novel computational paradigm for how the brain processes natural language. Unlike symbolic, rule-based models from psycholinguistics, DLMs encode words and their context as continuous numerical vectors. These "embeddings" are constructed by a sequence of layered computations to ultimately capture surprisingly sophisticated representations of linguistic structures. How does this layered hierarchy map onto the human brain during natural language comprehension? In this study, we used ECoG to record neural activity in language areas along the superior temporal gyrus and inferior frontal gyrus while human participants listened to a 30-minute spoken narrative. We supplied this same narrative to a high-performing DLM (GPT2-XL) and extracted the contextual embeddings for each word in the story across all 48 layers of the model. We next trained a set of linear encoding models to predict the temporally-evolving neural activity from the embeddings at each layer. We found a striking correspondence between the layer-by-layer sequence of embeddings from GPT2-XL and the temporal sequence of neural activity in language areas. In addition, we found evidence for the gradual accumulation of recurrent information along the linguistic processing hierarchy. However, we also noticed additional neural processes that took place in the brain, but not in DLMs, during the processing of surprising (unpredictable) words. These findings point to a connection between language processing in humans and DLMs where the layer-by-layer accumulation of contextual information in DLM embeddings matches the temporal dynamics of neural activity in high-order language areas.

Introduction

Deep language models (DLMs) provide an alternative computational framework for how the human brain processes natural language (Caucheteux & King, 2022; Goldstein et al., 2022; Schrimpf et al., 2021; Yang et al., 2019). Classical psycholinguistic models rely on rule-based manipulation of symbolical representations embedded in hierarchical tree structures (Kako & Wagner, 2001; Lees & Chomsky, 1957). In sharp contrast, DLMs encode words and their context as continuous numerical vectors—i.e. embeddings. These embeddings are constructed via a sequence of non-linear transformations across layers to yield the sophisticated representations of linguistic structures needed to produce language (Adiwardana et al., 2020; Brown et al., 2020; Radford et al., 2019; Yang et al., 2019).

Recent research has begun identifying shared computational principles between the way the human brain and DLMs represent and process natural language. In particular, several studies have used contextual embeddings derived from DLMs to successfully model human behavior as well as neural activity measured by fMRI, EEG, MEG, and ECoG during natural speech processing (Antonello et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022; Heilbron et al., 2020; Hollenstein et al., 2021; Schwartz et al., 2019; Toneva & Wehbe, 2019). Furthermore, recent studies have shown that similarly to DLMs, the brain incorporates prior context into the meaning of individual words (Jain & Huth, 2018; Caucheteux et al., 2021a; Schrimpf et al., 2021), spontaneously predicts forthcoming words (Goldstein, Zada et al., 2022), and computes post-word-onset prediction error signals (Donhauser & Baillet, 2020; Willems et al., 2016; Heilbron et al., 2020; Goldstein, Zada et al., 2022).

In this study, we focus on the internal sequence of non-linear transformations of the embeddings across layers within DLMs in relation to the internal processing of words in natural language in the human brain. How do these embeddings change across layers, and how do the layerwise sequence of transformations map onto the processing hierarchy of natural language in the human brain?

Recent work in natural language processing (NLP), has identified certain trends in the properties of embeddings across layers in DLMs (Rogers et al., 2020; Manning et al., 2020; Tenney et al., 2019). Embeddings at early layers most closely resemble the static, non-contextual input embeddings (Ethayarajh, 2019) and best retain the original word order (Liu et al., 2019); in contrast, embeddings are thought to become progressively more context-specific and sensitive to long-range linguistic dependencies among words across layers (Cui et al., 2019; Tenney et al., 2019).

Embeddings at the final layers are typically specialized for the training objective (next-word prediction in the case of GPT2–3) (Brown et al., 2020; Radford et al., 2019). These properties of the embeddings emerge from the conjunction of the architectural specifications of the network, the predictive objective, and the statistical structure of real-world language (Richards et al., 2019; Hasson et al., 2020).

In this study, we investigated how the layered structure of DLM embeddings maps onto the temporal dynamics of neural activity in language areas during natural language comprehension. Naively, we may expect the layerwise embeddings to roughly map onto a cortical hierarchy for language processing (similarly to the mapping observed between convolutional neural networks and the primate ventral visual pathway; Güçlü & van Gerven, 2015; Yamins & DiCarlo, 2016). In such a mapping, early language areas will be better modeled by embedding extracted from early layers of DLMs, whereas higher-order areas will be better modeled by embeddings extracted from later layers of DLMs. Interestingly, studies that examined the layer-by-layer match between DLM embeddings and brain activity using fMRI have observed that intermediate layers tend to provide the best fit across many language ROIs (Toneva & Wehbe, 2019; Caucheteux et al., 2021a; Schrimpf et al., 2021; Kumar, Sumers et al., 2022). These findings do not support the hypothesis that DLMs capture the processing sequence of words in natural language in the human brain.

In contrast, using ECoG recording with superior spatiotemporal resolution, we report that the human brain's internal temporal processing of spoken narrative matches the internal sequence of non-linear layerwise transformations in DLMs. The contextual embedding for each word in the narrative was extracted from all 48 layers in a specific DLM (GPT2-XL; Radford et al., 2019). Next, we compared the internal sequence of embeddings across the layers of GPT2-XL for each word to the sequence of neural responses recorded via ECoG in human participants. We first replicated the finding that intermediate layers best predict cortical activity. However, the improved temporal resolution of our ECoG recordings revealed a remarkable alignment between the layerwise DLM embedding sequence and the temporal dynamics of cortical activity during natural language comprehension. For example, within the inferior frontal gyrus (IFG; i.e. Broca's area) we observed a temporal sequence of encoding where earlier layers yield peak encoding performance earlier in time relative to word onset, and later layers yield peak encoding performance later in time. This finding suggests that the sequence of transformation across layers in DLMs maps onto a temporal accumulation of information in high-level language areas. Furthermore, we found evidence for the gradual accumulation of recurrent information along the linguistic processing hierarchy.

These findings point to a strong connection, with crucial differences, between the way the human brain and DLMs code natural language.

Results

We collected electrocorticographic (ECoG) data from 9 epilepsy patients while they listened to a 30-minute audio podcast ("Monkey in the Middle", NPR 2017). In prior work (Goldstein et al., 2022), we used embeddings from the final hidden layer of GPT2-XL to predict brain activity and found that these contextual embeddings outperform static (i.e. non-contextual) embeddings (see also Caucheteux et al., 2021a; Schrimpf et al., 2021). In this paper, we expand our analysis by modeling the neural responses for each word in the podcast using contextual embeddings extracted from each of the 48 hidden layers in GPT2-XL (Fig. 1A). We focus on four areas along the ventral language processing stream (Hickok & Poeppel, 2004; Karnath, 2001; Poliva, 2016): middle superior temporal gyrus (mSTG, n = 28 electrodes), anterior superior temporal gyrus (aSTG, n = 13), inferior frontal gyrus (IFG; i.e. Broca's area, n = 46), and the temporal pole (TP, n = 6). We selected electrodes previously shown to have significant encoding performance for static (GloVe) embeddings (corrected for multiple tests; Goldstein et al. 2022). Finally, given that prior studies have reported improved encoding results for words that are correctly predicted by DLMs (Caucheteux & King, 2022; Goldstein et al., 2022), we separately model the neural responses for correct predictions (i.e., where GPT2-XL's top-1 next-word predictions were correct; n = 1709) versus incorrect predictions. To ensure that we only analyze incorrect predictions and to match the statistical power across the two analyses, we defined incorrect predictions as cases where all top-5 next-word predictions were incorrect (n = 1808) (see Figs. S1–3 for analyses of all words combined).

For each layer and each lag (25 ms shifts relative to word onset), we fit a linear regression model using 90% of the words and predict brain activity in the remaining 10% of the words (10-fold cross-validation). We evaluate the performance of our model by correlating our predicted neural responses for each word with the actual neural responses (Fig. 1A–B). The analysis is repeated for each lag, ranging from -2000 ms before word onset (0 ms) to +2000 ms after word onset. We color-coded the encoding performance according to the index of the layer from which the embeddings were extracted, ranging from 1 (red) to 48 (blue; Fig. 1A). To better visualize the temporal dynamic across layers, we scaled the encoding performance to peak at 1 (Fig. 1B, right panel). To evaluate our procedure on specific regions of interest (ROIs), we average the encodings for the electrodes in the relevant ROIs before scaling.

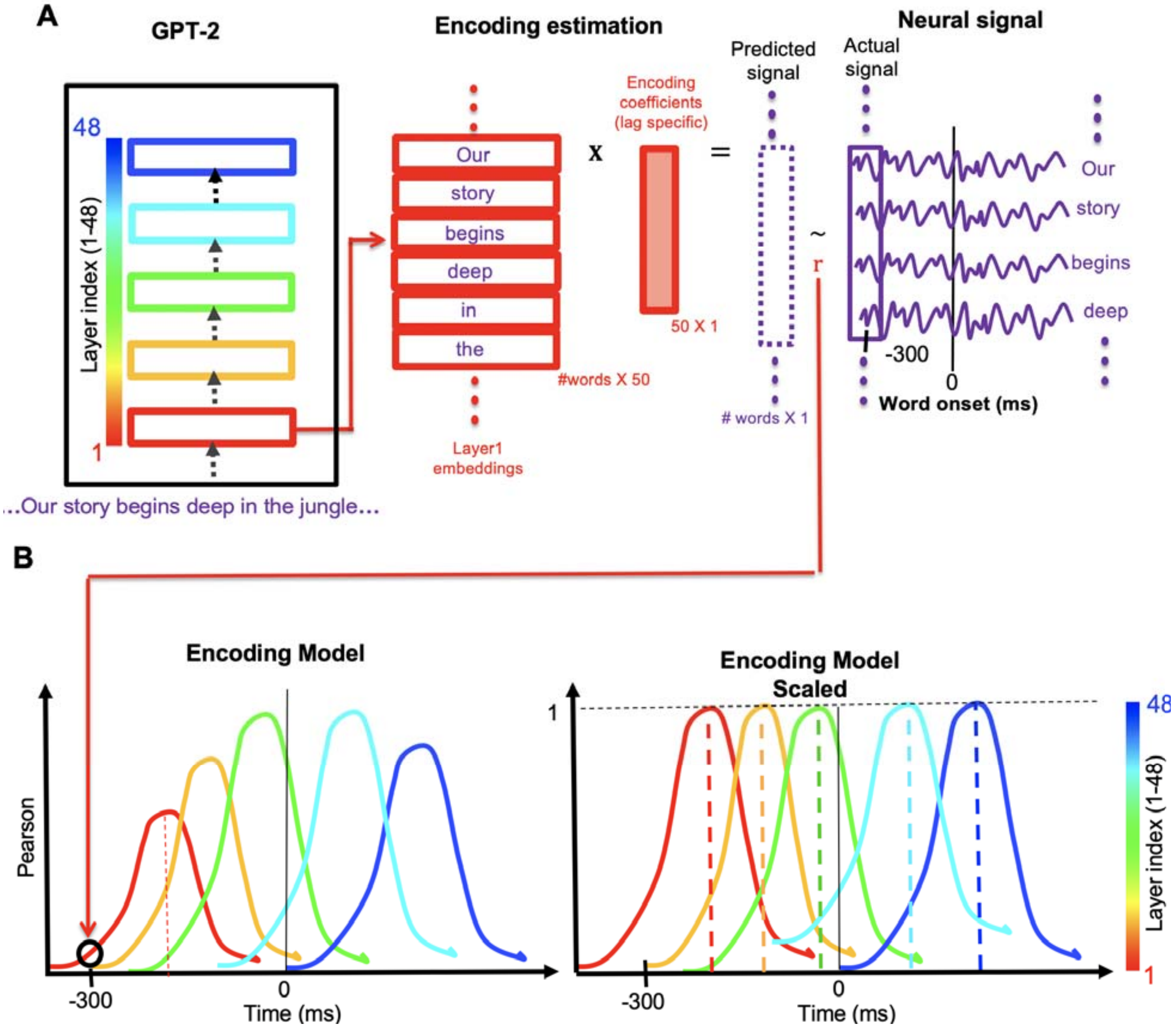


Figure 1. Layerwise encoding models. (A) We extracted the neural signal for each specific electrode before and after each word onset (denoted lag 0). The words and the neural signals were split into training and test sets comprising non-overlapping subsets of words for 10-fold cross-validation. The neural signal is averaged over a 200 ms rolling window with incremental shifts of 25 ms. For each word in the story, a contextual embedding is extracted from each layer of GPT-2 (for example, layer 1, red). The dimensionality of the embeddings is reduced to 50 using PCA. For each lag and each electrode, we used linear regression to estimate an encoding model that predicts the neural signal from the word embeddings. In order to evaluate the linear model, we used the 50-dimensional weight vector estimated from the training set to predict the neural signal of the words in the left-out test set from the corresponding embeddings. We evaluated the performance of the model by computing the correlation between the predicted neural signal and the actual neural signal of the words in the test set. (B) This process was repeated for lags ranging from -2000 ms to +2000 ms relative to word onset using the embeddings from each of the 48 hidden layers of GPT2-XL. We then rescaled the performance of the encoding model for each layer to one by normalizing the peak performance; this allows us to more easily visualize the temporal dynamics of encoding performance across layers.

We start by focusing on neural responses for correctly predicted words in electrodes at the inferior frontal gyrus (IFG; Broca's area; N = 46), a central region for semantic and syntactic linguistic processing (Goldstein et al., 2022; Hagoort, 2005; Hagoort & Indefrey, 2014; Ishkhanyan et al., 2020; LaPointe, 2012; Saur et al., 2008; Yang et al., 2019).

The peak correlation of the encoding models in the IFG was observed for the intermediate layer 22 (Fig. 2B; for other ROIs and predictability conditions see Supp. Fig. 1). This corroborates recent findings from fMRI (Caucheteux et al., 2021; Schrimpf et al., 2021; Toneva & Wehbe, 2019) where encoding performance peaks for intermediate layers, yielding an inverted U-shaped curve across layers (Fig. 2B). This inverted U-shaped pattern holds for all language areas (Fig. S1), suggesting that the layers of the model do not naively correspond to different cortical areas in the brain.

The fine-grained temporal resolution of ECoG recordings, however, suggests a more subtle dynamic pattern. All 48 layers yield robust encoding in the IFG, with encoding performance near zero at the edges of the lag window (-2000 ms and 2000 ms) and increased performance around word onset. This can be seen in the combined plot of all 48 layers (Fig. 2C; for other ROIs and predictability conditions see Supp. Fig. 2) and when we plot individually selected layers (Fig. 2D, layers 5, 25, 45). A closer look at the encoding results over lags (time) for each layer revealed an orderly dynamic in which the peak encoding performance for the early layers (e.g., layer 5, red, in Fig. 2D) tends to precede the peak encoding performance for intermediate layers (e.g., layer 25, green), which are followed by the later layers (e.g., layer 45, blue). To visualize the temporal sequence across lags we normalized the encoding performance for each layer by scaling its peak performance to 1 (Fig. 2E; for other ROIs and predictability conditions see Supp. Fig. 3). The layerwise encoding models in the IFG tend to peak in an orderly sequence over time. To quantitatively test this claim, we correlated the layer index (1–48) with the lag that yielded the peak correlation (Fig. 2F). The analysis yielded a strong significant positive Pearson correlation of 0.85 (p<10e-13; similar results were obtained with Spearman correlation; r = .80). Lastly, we also conducted a non-parametric analysis where we permuted the layer index 100,000 times (keeping the lags that yielded the peak correlations fixed) while correlating the lags with these shuffled layer indices. Using the null distribution we computed the percentile of the actual correlation (r=0.85) an got a significance of p<10e-5. Together, these results suggest that, for correct predictions, the sequence of internal transformations across the layers in GPT2-XL matches the sequence of internal transformations across time within the IFG.

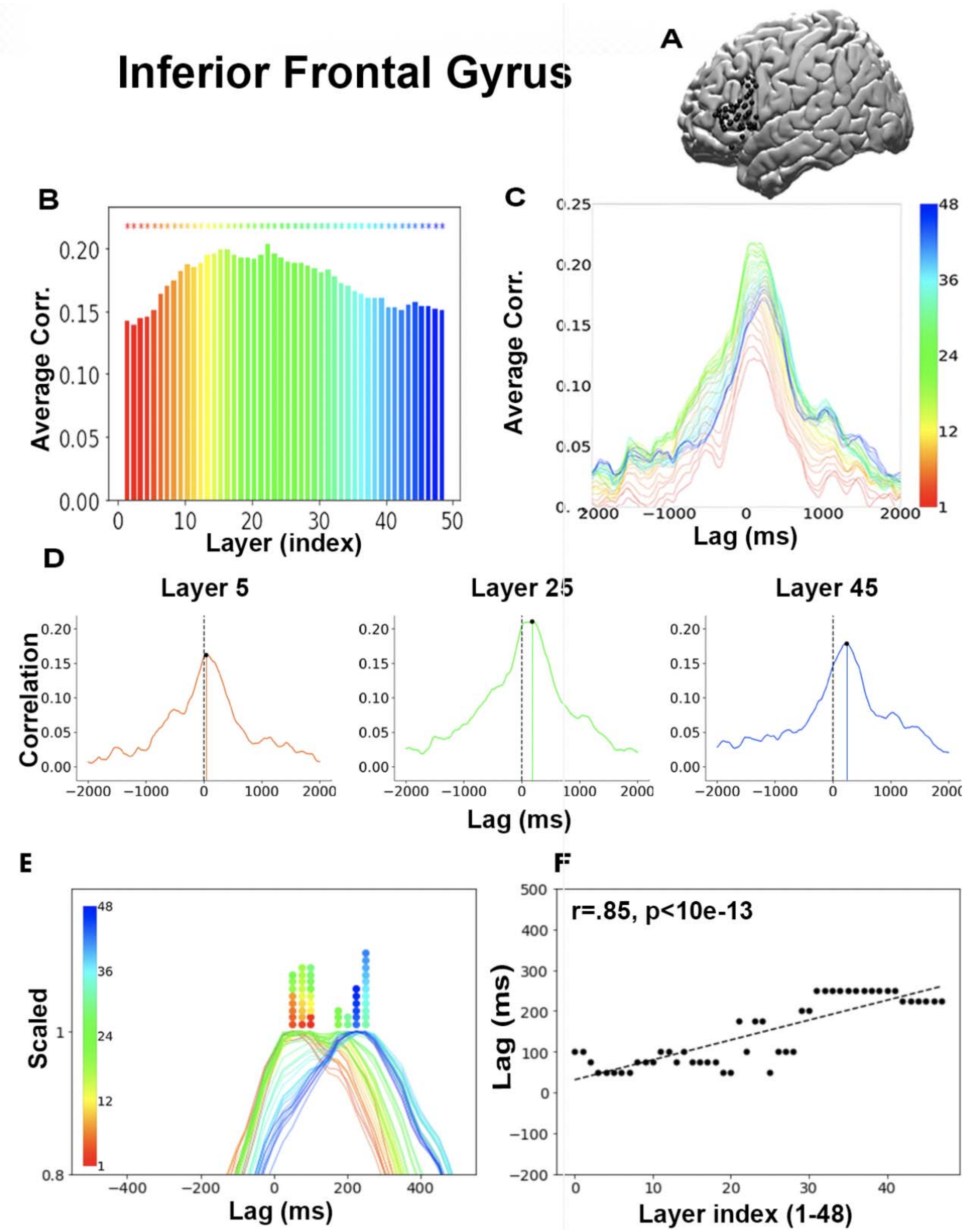


Figure 2. Temporal dynamics of layerwise encoding for correctly predicted words in IFG. (A). We recorded from 46 electrodes in the inferior frontal gyrus (IFG; Broca's area) that show positive encoding for word embeddings (GLoVe). (B) For each electrode in the IFG, we performed an encoding analysis for each hidden layer (1-48) at each lag (-2000 ms to 2000 ms). We then averaged encoding performance across all electrodes in the IFG to get a single mean performance value for each lag and layer. We color-

coded these encoding performance values according to the index of the layers from which the embeddings were extracted (red to blue). The peak encoding performance across lags for each layer at each electrode was averaged across electrodes and color-coded from early layers (red) to late layers (blue). Significance was assessed using bootstrap resampling across electrodes (see Materials and Methods). (C) Average correlation across electrodes for each layer at lags ranging from -2000 ms to +2000 ms relative to word onset (lag 0). (D) Encoding performance for layers 5, 25, and 45 demonstrates the layerwise shift of peak performance across lags. (E) Scaled encodings. Each layer encoding peak was scaled to 1. The colored dots mark the peaked encoding lag for each layer. The results show that the deeper the layer is in the model, the later its encoding model peaks (see the sequence from red to blue along the x-axis). (F) Scatter plot of the lag that yields peak encoding performance as a function of the index of layers.

Next, we compared the temporal encoding sequence across three additional temporal language ROIs (Fig. 3), starting with mSTG (near early auditory cortex) and moving up along the ventral linguistic stream to aSTG and TP. We did not observe obvious evidence for a temporal structure in the mSTG (*r* =-.24). This suggests that the temporal dynamic observed in IFG is regionally specific and does not take place in the early stages of the neural processing hierarchy. In addition to the IFG, we found evidence for the same orderly temporal dynamic in aSTG (*r* = .92, p<10e-20) and TP (*r* = .93, p<10e-22). Similar results were obtained with Spearman correlation (mSTG *r* =-.24, p=.09; aSTG r=.55, p=.9; IFG r=.79, p<10e-11;TP r=.95, p<10e-21), demonstrating that the effect is robust to outliers. Following our procedure for the IFG we conducted permutation tests by shuffling the layers order that yielded the following p-values: p<.02 (mSTG), p<10e-5 (aSTG, IFG). Furthermore, the width of the temporal sequence gradually increases as we proceed along the ventral linguistic hierarchy (see the increase in steepness of the slopes across language areas in Fig. 3). This was tested using Levene's test which yielded significant differences between the standard deviations of lags that yield maximal correlations for the different layers in the mSTG and aSTG (F = 48.1, p<.01), as well as between the aSTG and TP (F = 5.8, p<.02). The largest temporal separation across layer-based encoding models was seen in TP, with more than a 500 ms difference between the peak for layer 1 (around -100 ms) and the peak for layer 48 (around 400 ms).

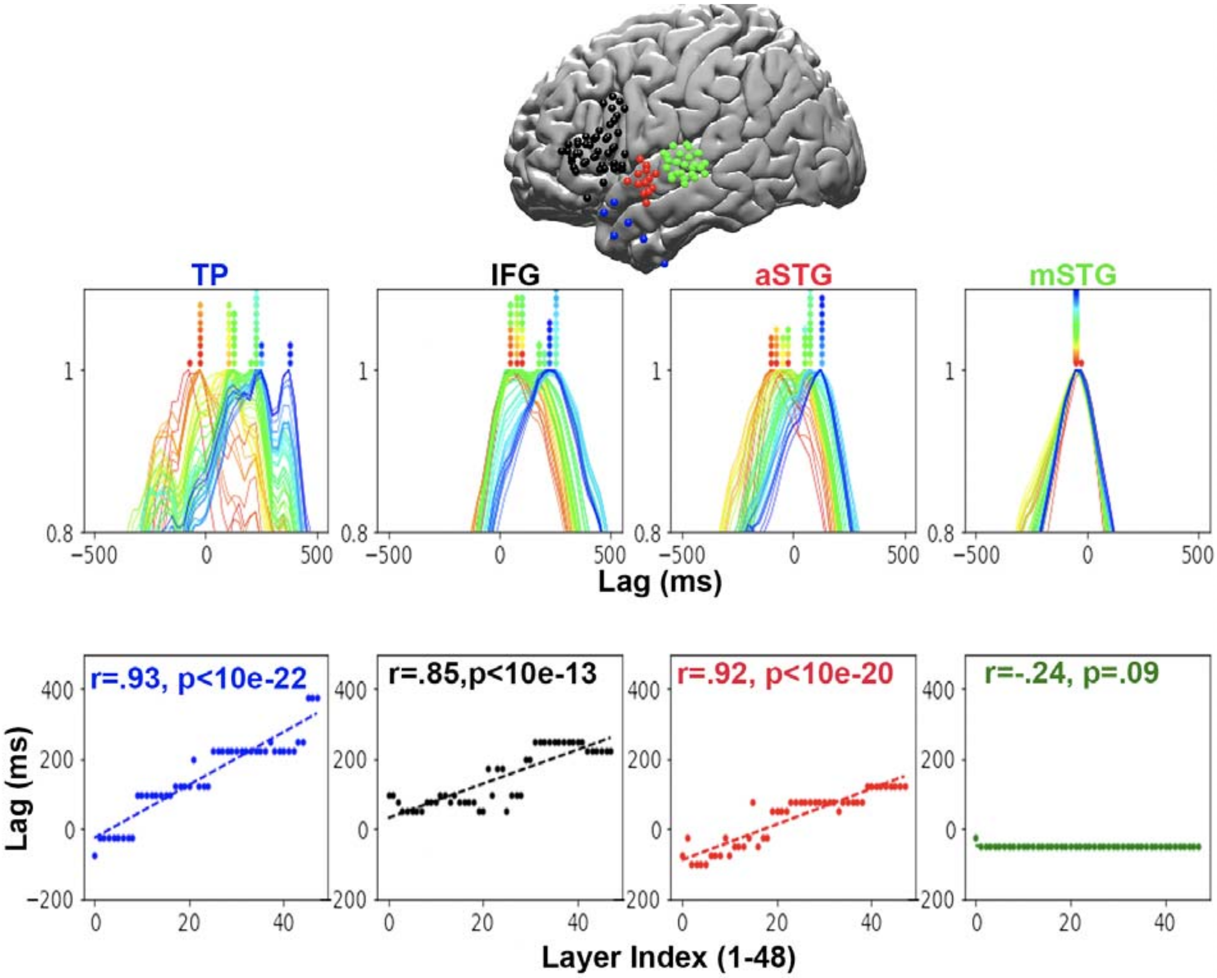


Figure 3. Temporal hierarchy along the ventral language stream for correctly predicted words. Scaled encoding for ROIs along the ventral language processing stream, from the middle superior temporal gyrus (mSTG) to the anterior superior temporal gyrus (aSTG), inferior frontal gyrus (IFG), and the temporal pole (TP). The results reveal a temporal sequence of layer-based encoding in all language areas besides mSTG. Furthermore, the processing timescales (slop of lag-difference across layers) increased along the ventral linguistic hierarchy from mSTG to aSTG to IFG and TP.

The temporal correspondence described so far was observed for words the model accurately predicted; does the same pattern hold for words that were not accurately predicted? We conducted the same layerwise encoding analyses in the same ROIs for unpredictable words—i.e. words for which the probability assigned to the word was not among the top-5 highest probabilities assigned by the model (N = 1808). We still see evidence, albeit slightly weaker, for layer-based encoding sequences in the IFG (*r* = .81, p<10e-11) and TP (*r* = .57, p<10e-4), but not aSTG (*r* = .09, p>.55) or mSTG (*r* = -.10,p>.48). Similar results were obtained with Spearman correlation (mSTG *r* = -.10, p>.48; aSTG r=.02, p>.9; IFG r=.8, p<10e-11; TP r=.72, p<10e-8), demonstrating that the effect is robust to outliers. We conducted permutation tests that yielded the following p-values: p>.24 (mSTG), p>.27 (aSTG),p<10e-5 (TP, IFG). We also noticed a crucial difference between the encoding of the correctly and incorrectly predicted words in the IFG. In the IFG the encoding for early layers (red) shifted from around word onset (lag 0) for correct prediction to later lags (around 300ms) for incorrect predictions. We ran a paired t-test to compare the average lags (across the electrodes in a ROI) that yield the maximal correlations (i.e., peak encoding performance) across predicted and unpredicted words for each layer. The paired t-test indicated that the shift in peak encoding (at the ROI level) was significant for 9 out of the 12 first layers (corrected for multiple comparisons see supp. table 1, q<0.05).

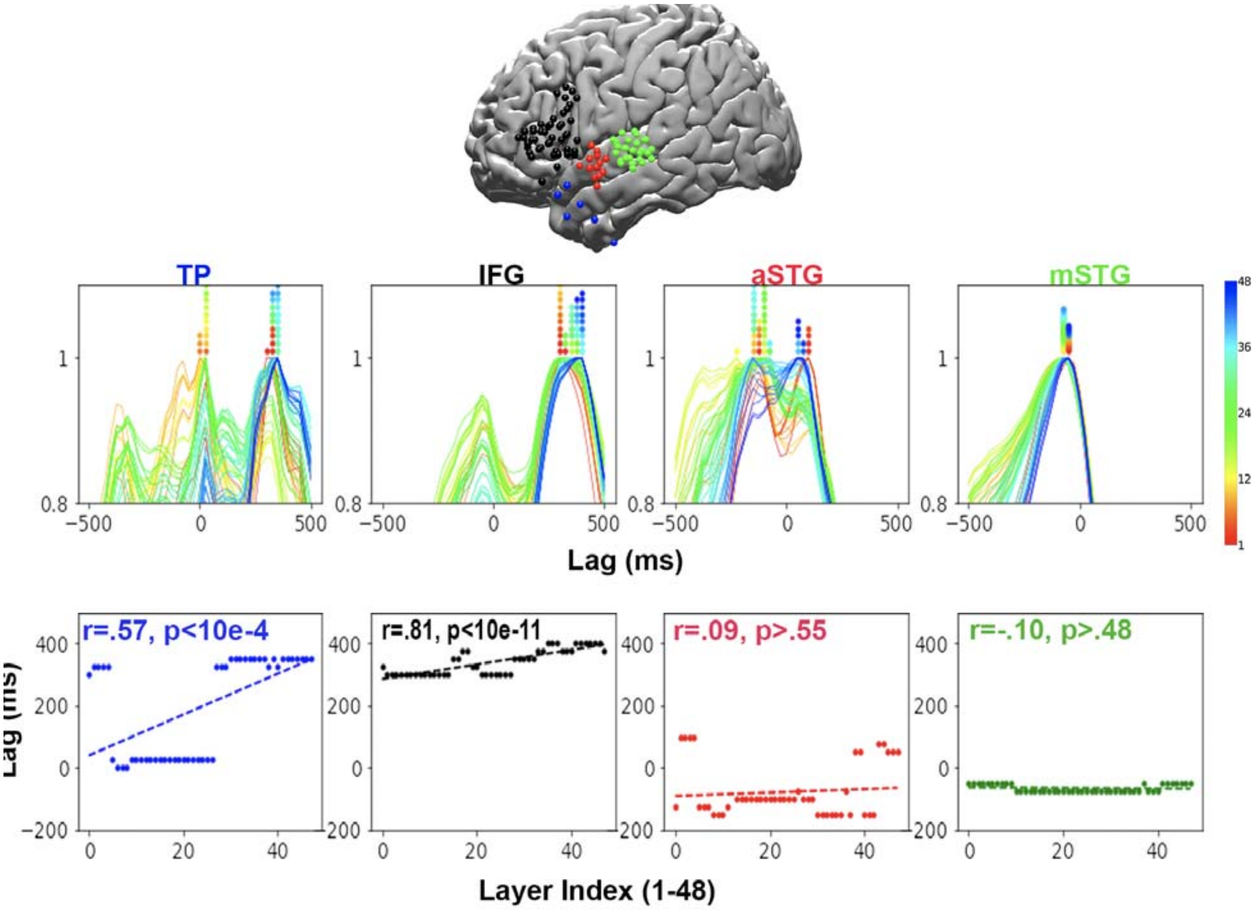


Figure 4. Temporal hierarchy along the ventral language stream for incorrectly predicted words. Scaled encoding performance for separate areas along the ventral language pathway, from the middle superior temporal gyrus (mSTG) to the anterior superior temporal gyrus (aSTG), inferior frontal gyrus (IFG), and the temporal pole (TP). The encoding analysis was performed for words that were incorrectly predicted by the model. A word was classified as incorrectly predicted if it was not among the top 5 most probable words predicted by GPT2-XL given the context.

Discussion

Prior studies reported shared computational principles (e.g., prediction in context and representation using multidimensional embeddings space) between DLMs and the human brain (Schrimpf et al., 2021; Caucheteux et al., 2022; Goldstein et al., 2022). In the current study, we extracted the contextual embeddings for each word in a chosen narrative across all 48 layers and fitted them to the neural responses to each word in our human participants. We found that the *sequence* of layerwise transformations learned by GPT2-XL maps onto the temporal *sequence* of transformations of linguistic input in high-level language areas. This finding reveals a surprising and important link between how DLMs and the brain process language: conversion of discrete input into multidimensional (vectorial) embeddings, which are further transformed via a sequence of non-linear transformations to match the context-based statistical properties of natural language (Manning et al., 2020). These results provide additional evidence for shared computational principles between the way DLMs and the human brain process natural language.

At the same time, our study also points to implementational differences between the internal sequence of computations in transformer-based DLMs and the human brain. GPT2 relies on a “transformer” architecture, a neural network architecture developed to process hundreds to thousands of words in parallel. In other words, transformers are designed to parallelize a task that is largely computed serially, word by word, in the human brain. While transformer-based DLMs process words sequentially over layers, in the human brain we found evidence for similar sequential processing, but over time relative to word onset within a given cortical area. For example, we found that within high-order language areas (such as IFG and TP) the sequence of layerwise processing in DLMs corresponded to a sequence of temporal processing.

What are possible explanations for this result? First, it may be that cortical computations within a given language area are better aligned with recurrent architectures, where the internal computational sequence is deployed over time rather than over layers. In addition, however, we observed evidence for recurrent processing at different time scales across different levels of the linguistic processing hierarchy. That is, the sequence of temporal processing unfolds over longer timescales as we proceed up the processing hierarchy, from aSTG to IFG, and TP. Second, it may be that layered architecture of GPT2 is recapitulated within the local connectivity of a given language area like IFG (rather than across cortical areas). That is, local connectivity within a given cortical area may resemble the layered graph structure of GPT2. Third, it is possible that long-range connectivity *between* cortical areas could yield the temporal sequence of processing observed within a single cortical area. Together, these results hint that a deep language model with stacked recurrent networks may better fit the human brain's neural architecture for processing natural language. Interestingly, there have been several attempts to develop such new architectures, such as universal transformers (Dehghani et al., 2018; Lan et al., 2019) and reservoir computing (Dominey, 2021). Future studies will have to compare how the internal processing of natural language compares between these models and the brain.

Previous results indicate that the ability to encode the neural responses in language areas using DLMs varies with the accuracy of their next-word predictions and is lower for incorrect predictions (Caucheteux & King, 2022; Goldstein et al., 2022). In contrast, we observed that even for unpredicted words, the temporal encoding sequence was maintained in high-order language areas (IFG and TP). However, we do find a difference in the neural responses for unpredictable words in the IFG, in which early layers encoding in IFG shifted from around word-onset for predictable words to around 300-400ms after word-onset for unexpected words (Fig. 4). This finding suggests that the dynamic of neural responses in human language areas is systematically different for predictable and unpredictable words.

Replicating prior studies (Caucheteux et al., 2021a; Schrimpf et al., 2021; Schwartz et al., 2019), we also noticed that intermediate layers best matched neural activity in language areas (Fig. S2). Intermediate layers are thought to best capture the syntactic and semantic structure of the input (Hewitt & Manning, 2019; Jawahar et al., 2019) and generally provide the best generalization to other NLP tasks (Liu et al., 2019). The improved correlation between neural activity and GPT2–XL's intermediate layers suggests that the language areas place additional weight on such intermediate representations. At the same time, each layer's embedding is distinct and represents different linguistic dimension (Rogers et al., 2020), and thus, invoke a unique temporal encoding pattern. Thus, our finding of a gradual sequence of transitions in language areas is complimentary and orthogonal to the level of encoding across layers.

This paper provides strong evidence that DLMs and the brain process language in a similar way. Given the clear circuit-level architectural differences between DLMs and the human brain, the convergence of their internal computational sequences may be surprising. Classical psycholinguistic theories postulated an interpretable rule-based symbolic system for linguistic processing. In contrast, DLMs provide a radically different, statistical learning framework for learning the structure of language by predicting speakers' language use in context. This kind of unexpected mapping (layer sequence to temporal sequence) can point us in novel directions for both understanding the brain and developing neural network architectures that better mimic human language processing. Taken together, this study provides strong evidence for shared internal computations between DLMs and the human brain and calls for a paradigm shift from a symbolic representation of language to a new family of contextual embeddings and statistical learning-based models.

Materials and methods

Data acquisition and preprocessing

The procedure for collecting and preprocessing including the high-gamma-band extraction (70-200 hz) of the neural signal (ECoG) from the participants while they listened to the podcast are described in (Goldstein et al., 2022; Hickok, 2009; Rauschecker, 2012; Saur et al., 2008)

Linguistic embeddings

In order to extract contextual embeddings for the stimulus text, we first tokenized the words for compatibility with GPT2-XL. We then ran the GPT2-XL model implemented in HuggingFace (Wolf et al., 2020) on this tokenized input. To construct the embeddings for a given word, we passed the set of up to 1023 words preceding the word (the context) along with the current word as input to the model. We include the current word for convenience, but the embedding we extract is the output generated for the previous word. This means that the current word is not used to generate its own embedding and its context only includes previous words. We constrain the model in this way because our human participants do not have access to the words in the podcast before they are said during natural language comprehension.

GPT2-XL is structured as a set of blocks that each contain a self-attention sub-block and a subsequent feedforward sub-block. The output of a given block is the summation of the feedforward output and the self-attention output through a residual connection. This output is also known as a “hidden state” of GPT2-XL. We consider this hidden state to be the contextual embedding for the block that precedes it. For convenience, we refer to the blocks as “layers”; that is, the hidden state of output by block 3 is referred to as the contextual embedding for layer 3. In order to generate the contextual embeddings for each layer, we store each layer’s hidden state for each word in the input text. Fortunately, the HuggingFace implementation of GPT2-XL automatically stores these hidden states when a forward pass of the model is conducted. Different models have different numbers of layers and and embeddings of different dimensionality. The model used herein, GPT2-XL, has 48 layers and the embeddings at each layer comprise 1600-dimensional vectors. For a sample of text containing 101 tokens, we would generate an embedding for each layer and each word, excluding the first word as it has no prior context. This results in 48 1600-dimensional embeddings per word and 100 words; 48 * 100 = 4800 total 1600-long embedding vectors. Note that in this example the context length would increase from 1 to 100 as we proceed through the text.

Dimensionality reduction

Before fitting the encoding models, we first reduce the dimensionality of the embeddings by applying principal component analysis (PCA) and retaining the first 50 components. This procedure effectively focuses our subsequent analysis on the 50 orthogonal dimensions in the embedding space that account for the most variance in the stimulus.

Encoding models:

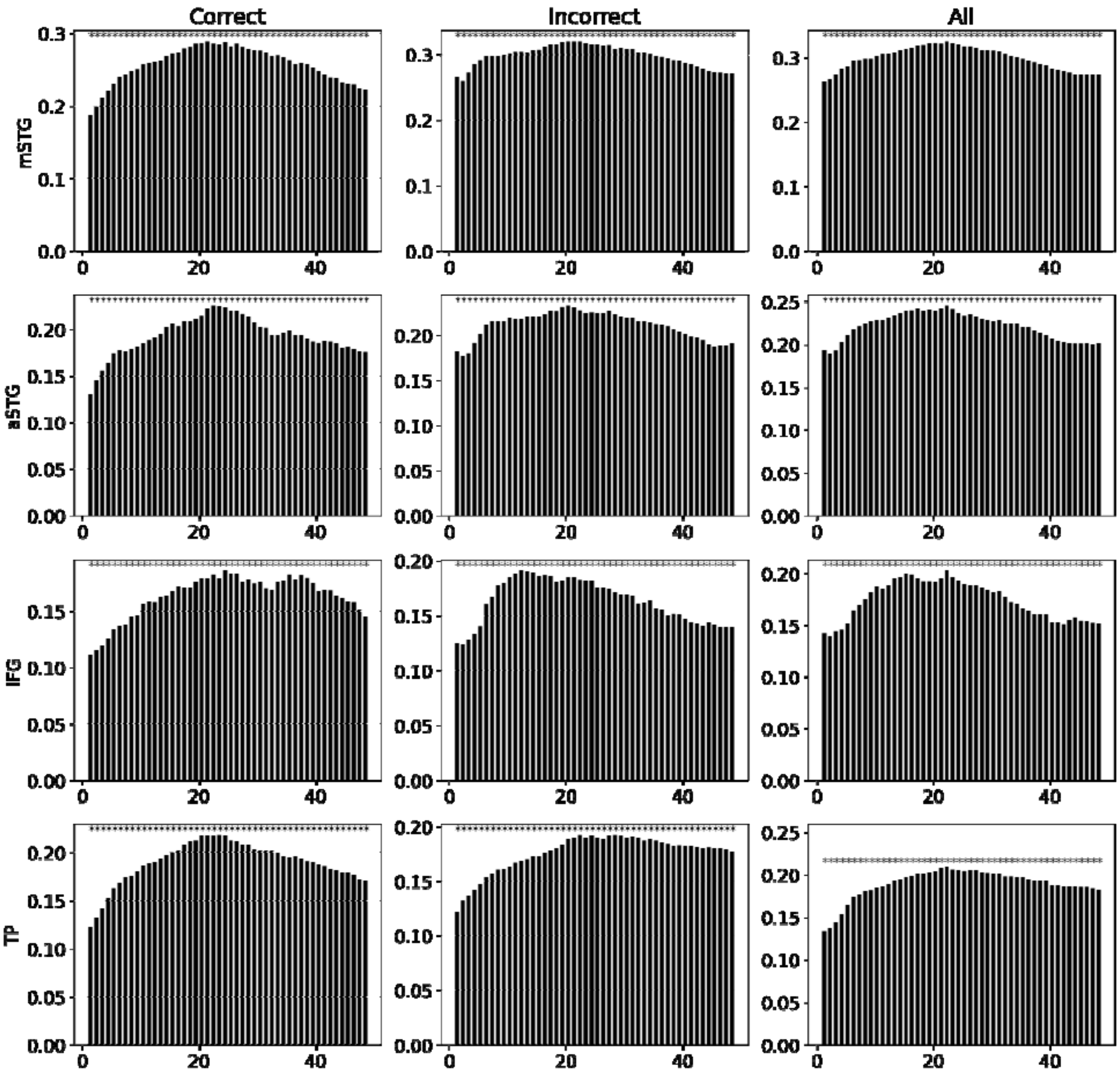
Linear encoding models were estimated at each lag (-2000 ms to 2000 ms in 25-ms increments) relative to word onset (0 ms) to predict the brain activity for each word from the corresponding contextual embedding. Before fitting the encoding model, we smoothed the signal using a rolling 200-ms window. We used a 10-fold cross-validation procedure ensuring that for each cross-validation fold, the model was estimated from a subset of training words and evaluated on a non-overlapping subset of held-out test words: the words and the corresponding brain activity were split into a training set (90% of the words) for model estimation and a test set (10% of the words) for model evaluation. Encoding models were estimated separately for each electrode (and each lag relative to word onset). For each cross-validation fold, we used ordinary least squares (OLS) multiple linear regression to estimate a weight vector (50 coefficients for the 50 PCA components) based on the training words. We then used those weights to predict the neural responses at each electrode for the test words. We evaluated model performance by computing the correlation between the predicted brain activity and the actual brain activity across the held-out test words; we then averaged these correlations across electrodes. This procedure was performed for all the hidden states in GPT2-XL to generate an “encoding” for each layer.

Correct and incorrect predictions

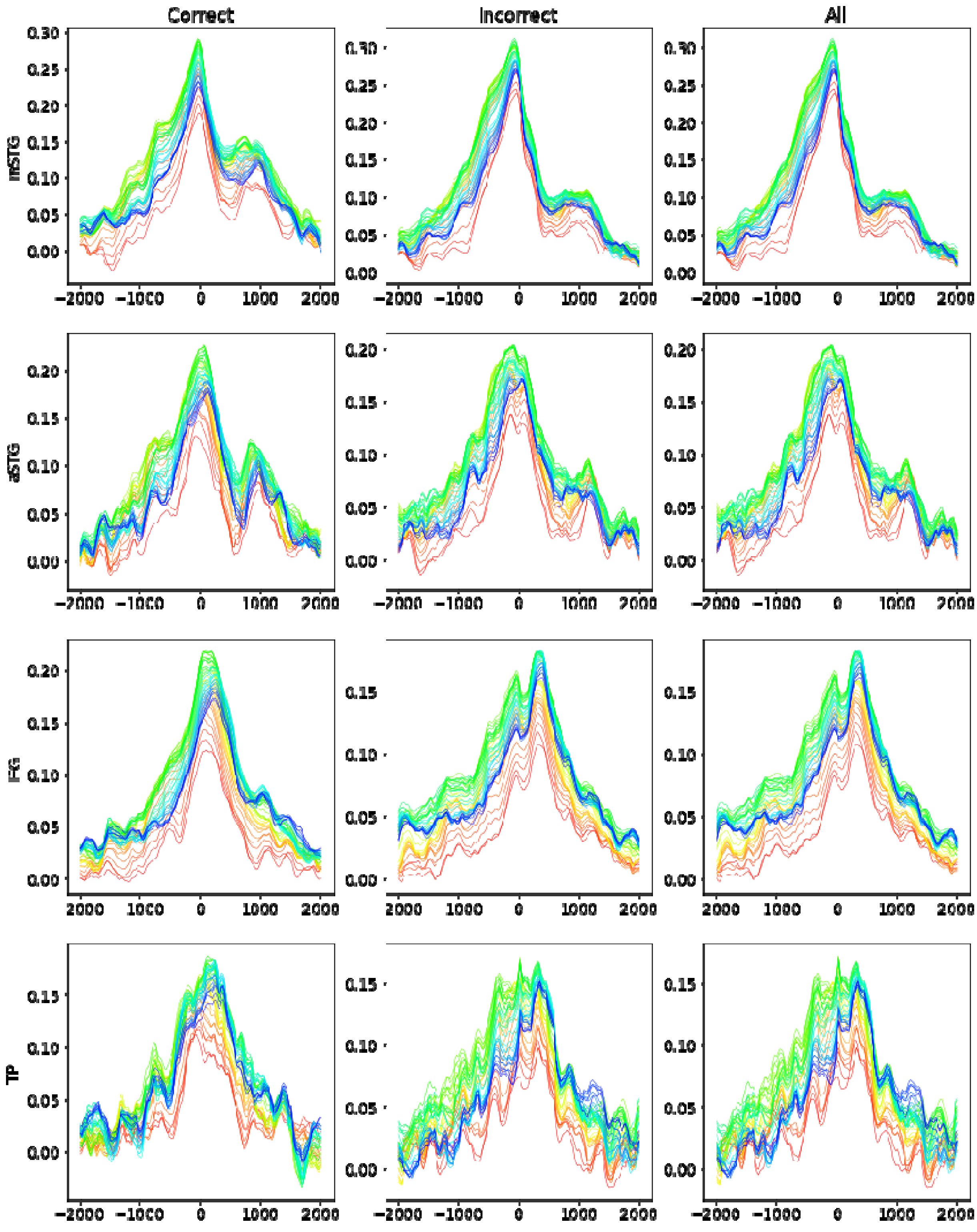
After generating encodings for all words in the podcast transcript, we split the embeddings into two subsets: words that the model predicted correctly and words that the model predicted incorrectly. A word was considered to be predicted correctly if the model assigned that word the highest probability of occurring next among all possible words. We refer to these subsets of embeddings as “top 1 predictable” (1709/4744 = 36%) ” and “top 5 predictable”. To reduce the stringency of top 1 prediction, we also created subsets of “top 5 predictable” (2936/4744 = 62%) and “top 5 unpredictable” words where the criterion for correctness was that the probability for the correct word must be among the highest five probabilities assigned to words by the model. We then trained linear encoding models as outlined above on these subsets of embeddings.

Statistical significance

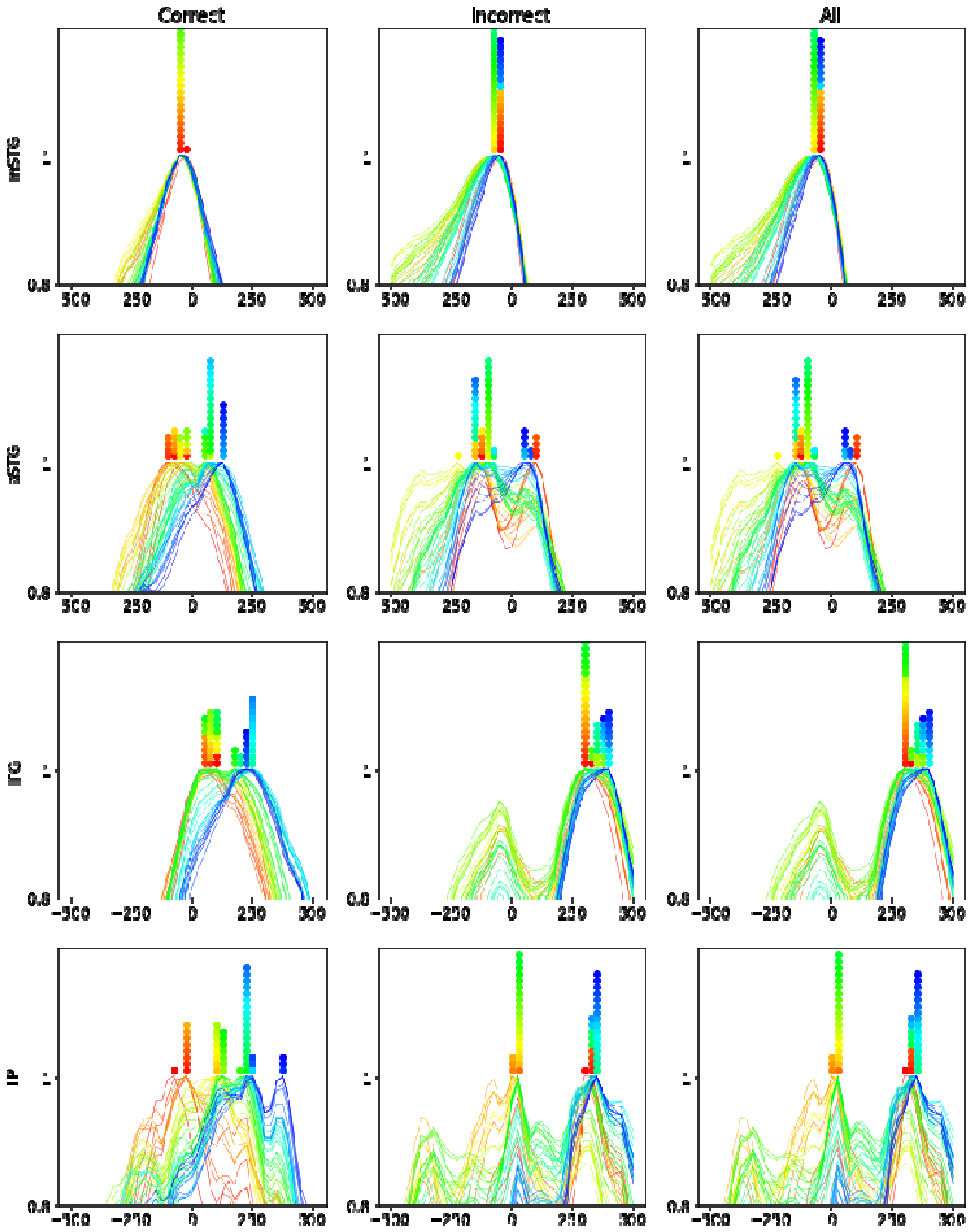
To establish the significance of the bars in Fig. 2B we conducted a bootstrapping analysis for each lag. Given the values of the electrodes in a specific layer and a specific ROI, we sampled the max correlations values with replacement 10^4 samples with the size of the number of electrodes. For each sample we computed the average and generated a distribution (consisting of 10^4 points). We then compared the actual mean for the lag-ROI pair to estimate how significant it is given the generated distributions. The '*' indicates two-tailed significance of $p<0.01$.



Supplementary Figure 1. Peak correlations of electrode-averaged encodings for each combination of layer (1-48) and brain area (mSTG, aSTG, IFG and TP) and word classification (correctly predicted, incorrectly predicted, all words). The significance test is done using bootstrap analysis across the electrodes.



Supplementary Figure 2. Encoding averaged over electrodes for each combination of layer (1-48), brain area (mSTG, aSTG, IFG and TP) and word classification (correctly predicted, incorrectly predicted, all words).



Supplementary Figure 3. Scaled encoding for each combination of layer (1-48), brain area (mSTG, aSTG, IFG and TP) and word classification (correctly predicted, incorrectly predicted, all words). For completion the correlation between the layer index and max-lag for condition 'All': mSTG (r=.56, p<10e-4), aSTG (r=.81, p<10e-11), IFG (r=.89 ,p<10e-16), TP (r=.75 ,p<10e-9)

Layer index	p. value	q-value	Layer index	p-value	q-value
1	0.784826	0.459996	25	0.731631	0.3963
2	0.061834	0.015458	26	0.719935	0.360514
3	0.016337	0.000953	27	0.608419	0.278859
4	0.016337	0.00111	28	0.569881	0.249323
5	0.409457	0.153546	29	0.38613	0.136755
6	0.016337	0.001688	30	1	0.949051
7	0.016337	0.001847	31	0.990003	0.696491
8	0.035199	0.0066	32	1	0.906098
9	0.016522	0.002409	33	1	0.94135
10	0.023182	0.003864	34	1	0.922248
11	0.016337	0.002042	35	1	0.9526
12	0.051168	0.01066	36	0.719935	0.374966
13	0.283744	0.094581	37	0.927577	0.618385
14	0.276009	0.086253	38	1	0.844096
15	0.21497	0.0627	39	0.784826	0.474165
16	0.059726	0.013687	40	0.927577	0.61807
17	0.016337	0.000372	41	0.820915	0.513072
18	0.191238	0.051794	42	1	0.961332
19	0.425111	0.168273	43	1	0.736217
20	0.990003	0.701252	44	1	0.942203
21	1	0.993479	45	0.548249	0.228437
22	0.749748	0.421733	46	1	0.968246
23	0.703456	0.337073	47	1	1
24	1	0.825196	48	1	0.907848

Supplementary Table 1. The p value and FDR-corrected q-value of the paired sampled t-test comparing the lags that achieve maximal correlation in the encoding across the different layers (n=48) of GPT2-XL.

References

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., & Le, Q. V. (2020). Towards a Human-like Open-Domain Chatbot. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2001.09977

Antonello, R., Turek, J., Vo, V., & Huth, A. (2021). Low-Dimensional Structure in the Space of Language Representations is Reflected in Brain Responses. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2106.05426

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & Others. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Caucheteux, C., Gramfort, A., & King, J.-R. (n.d.). *GPT-2's activations predict the degree of semantic comprehension in the human brain*. https://doi.org/10.1101/2021.04.20.440622

Caucheteux, C., Gramfort, A., & King, J. R. (2021). GPT-2's activations predict the degree of semantic comprehension in the human brain. *bioRxiv*. https://www.biorxiv.org/content/10.1101/2021.04.20.440622v2.abstract

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 134.

Donhauser, P. W., & Baillet, S. (2020). Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron*, *105*(2), 385–393.e9.

Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., & Kaiser, Ł. (2018). Universal Transformers. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1807.03819

Dominey, P. F. (2021). Narrative event segmentation in the cortical reservoir. *PLoS Computational Biology*, *17*(10), e1008993.

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, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005-10014.

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences*, *9*(9), 416–423.

Hagoort, P., & Indefrey, P. (2014). The neurobiology of language beyond single words. *Annual Review of Neuroscience*, *37*, 347–362.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron*, *105*(3), 416–434.

Heilbron, M., Armeni, K., Schoffelen, J. M., & Hagoort, P. (2020). A hierarchy of linguistic predictions during natural language comprehension. *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.12.03.410399v1.abstract

Hickok, G. (2009). The functional neuroanatomy of language. In *Physics of Life Reviews*

(Vol. 6, Issue 3, pp. 121–143). https://doi.org/10.1016/j.plrev.2009.06.001

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1-2), 67–99.

Hollenstein, N., Renggli, C., Glaus, B., Barrett, M., Troendle, M., Langer, N., & Zhang, C. (2021). Decoding EEG Brain Activity for Multi-Modal Natural Language Processing. *Frontiers in Human Neuroscience*, *15*, 659410.

Ishkhanyan, B., Michel Lange, V., Boye, K., Mogensen, J., Karabanov, A., Hartwigsen, G., & Siebner, H. R. (2020). Anterior and Posterior Left Inferior Frontal Gyrus Contribute to the Implementation of Grammatical Determiners During Language Production. *Frontiers in Psychology*, *11*, 685.

Kako, E., & Wagner, L. (2001). The semantics of syntactic structures. In *Trends in Cognitive Sciences* (Vol. 5, Issue 3, pp. 102–108). https://doi.org/10.1016/s1364-6613(00)01594-1

Karnath, H. O. (2001). New insights into the functions of the superior temporal cortex. *Nature Reviews. Neuroscience*, *2*(8), 568–576.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1909.11942

LaPointe, L. L. (2012). *Paul Broca and the Origins of Language in the Brain*. Plural Publishing.

Lees, R. B., & Chomsky, N. (1957). Syntactic Structures. In *Language* (Vol. 33, Issue 3, p. 375). https://doi.org/10.2307/411160

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(48), 30046–30054.

Poliva, O. (2016). From Mimicry to Language: A Neuroanatomically Based Evolutionary Model of the Emergence of Vocal Language. *Frontiers in Neuroscience*, *10*, 307.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Rauschecker, J. P. (2012). Ventral and dorsal streams in the evolution of speech and language. *Frontiers in Evolutionary Neuroscience*, *4*, 7.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866.

Saur, D., Kreher, B. W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M.-S., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, W., Rijntjes, M., Hennig, J., & Weiller, C. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(46), 18035–18040.

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). https://doi.org/10.1073/pnas.2105646118

Schwartz, D., Toneva, M., & Wehbe, L. (2019). Inducing brain-relevant bias in natural language processing models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 14123–14133). Curran Associates, Inc.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the Classical NLP Pipeline. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1905.05950

Toneva, M., & Wehbe, L. (2019, May 28). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.* http://arxiv.org/abs/1905.11833

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356-365.

Yang, X., Li, H., Lin, N., Zhang, X., Wang, Y., Zhang, Y., Zhang, Q., Zuo, X., & Yang, Y.

(2019). Uncovering cortical activations of discourse comprehension and their overlaps with common large-scale neural networks. In *NeuroImage* (Vol. 203, p. 116200). https://doi.org/10.1016/j.neuroimage.2019.116200

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf