Reviewed Preprint
v1 • October 22, 2024
Not revised

Reviewed Preprint
v2 • August 4, 2026
Revised by authors

Scale matters: Large language models with billions (rather than millions) of parameters better match neural representations of natural language

Zhuoqiao Hong[1,*], Haocheng Wang[1,*] ✉, Zaid Zada[1], Harshvardhan Gazula[2], David Turner[1], Bobbi Aubrey[1], Leonard Niekerken[1], Werner Doyle[3], Sasha Devore[3], Patricia Dugan[3], Daniel Friedman[3], Orrin Devinsky[3], Adeen Flinker[3], Uri Hasson[1,], Samuel A Nastase[1,**], Ariel Goldstein[4,**]**

[1]Department of Psychology and the Neuroscience Institute, Princeton University, Princeton, United States • [2]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, United States • [3]New York University Grossman School of Medicine, New York, United States • [4]Business School, Data Science Department and Cognitive Science Department, Hebrew University, Jerusalem, Israel

For correspondence: kw1166@princeton.edu

[*] Equal first author, alphabetical order

[**] Equal senior author

Competing interests: No competing interests declared

Funding: See page 22

Reviewing editor: Nai Ding, Zhejiang University, China

eLife Assessment

This **important** study investigates how the size of an LLM may influence its ability to model the human neural response to language recorded by ECoG. Overall, **solid** evidence is provided that larger language models can better predict the human ECoG response. This study will be of interest to both neuroscientists and psychologists who work on language comprehension and computer scientists working on LLMs.

https://doi.org/10.7554/eLife.101204.2.sa4

Abstract

Recent research has used large language models (LLMs) to study the neural basis of naturalistic language processing in the human brain. LLMs have rapidly grown in complexity, leading to improved language processing capabilities. However, neuroscience researchers haven't kept up with the quick progress in LLM development. Here, we utilized several families of transformer-based LLMs to investigate the relationship between model size and their ability to capture linguistic information in the human brain. Crucially, a subset of LLMs were trained on a fixed training set, enabling us to dissociate model size from architecture and training set size. We used electrocorticography (ECoG) to measure neural activity in epilepsy patients while they listened to a 30-minute naturalistic audio story. We fit electrode-wise encoding models using contextual embeddings extracted from each hidden layer of the LLMs to predict word-level neural signals. In line with prior work, we found that larger LLMs better capture the structure of natural language and better predict neural activity. We also found a logarithmic relationship where the encoding performance peaks in relatively earlier layers as model size increases. We also observed variations in the best-performing layer across different brain regions, corresponding to an organized language processing hierarchy.

Introduction

How has the functional architecture of the human brain come to support everyday language processing? Modeling the underlying neural basis that supports natural language processing has proven to be prohibitively challenging for many years. Deep learning has brought about a transformative shift in our ability to model natural language in recent years. Leveraging

s from statistical learning theory and using vast real-world datasets, deep learning algorithms can reproduce complex natural behaviors in visual perception, speech analyses, and even human-like conversations. With the recent emergence of large language models (LLMs), we are finally beginning to see explicit computational models that respect and reproduce the context-rich complexity of natural language and communication. LLMs rely on simple self-supervised objectives (e.g., next-word prediction) to learn to produce context-specific linguistic outputs from real-world corpora—and, in the process, implicitly encode the statistical structure of natural language into a multidimensional embedding space (Linzen & Baroni, 2021; Manning et al., 2020; Pavlick, 2022).

Critically, there appears to be an alignment between the internal activity in LLMs for each word embedded in a natural text and the internal activity in the human brain while processing the same natural text. Indeed, recent studies have revealed that the internal, layer-by-layer representations learned by these models predict human brain activity during natural language processing better than any previous generations of models (Caucheteux & King, 2022; Goldstein et al., 2022, 2024; Kumar et al., 2024; Schrimpf et al., 2021).

LLMs, however, contain millions or billions of parameters, making them highly expressive learning algorithms. Combined with vast training text, these models can encode a rich array of linguistic structures—ranging from low-level morphological and syntactic operations to high-level contextual meaning—in a high-dimensional embedding space. Recent work has argued that the "size" of these models—the number of learnable parameters—is critical, as some linguistic competencies only emerge in larger models with more parameters (Bommasani et al., 2021; Kaplan et al., 2020; Manning et al., 2020; Sutton, 2019; C. Zhang et al., 2021). For instance, in-context learning (Liu et al., 2022; Xie et al., 2021) involves a model acquiring the ability to carry out a task for which it was not initially trained, based on a few-shot examples provided by the prompt. This capability is present in the bigger GPT-3 (Brown et al., 2020) but not in the smaller GPT-2, despite both models having similar architectures. This observation suggests that simply scaling up models produces more human-like language processing. Indeed, research in comparative neuroscience has suggested that uniquely human cognitive abilities emerged from scaling up the primate brain (Herculano-Houzel, 2012). While building and training LLMs with billions to trillions of parameters is an impressive engineering achievement, such artificial neural networks are tiny compared to cortical neural networks. In the human brain, each cubic millimeter of cortex contains a remarkable number of about 150 million synapses, and the language network can cover a few centimeters of the cortex (Cantlon & Piantadosi, 2024).

Our study focuses on one crucial question: What is the relationship between the size of an LLM and how well it can predict linguistic information encoded in the brain? In this study, we define "model size" as the number of all trainable parameters in the model. In addition to size, we also consider the model's expressivity: its capacity to predict the statistics of natural language. Perplexity measures expressivity by evaluating the average level of surprise or uncertainty the model attributes to a sequence of words. Larger models possess a greater capacity for expressing linguistic structure, which tends to yield lower (better) perplexity scores (Radford et al., 2019). In this paper, we hypothesized that larger models that capture linguistic structure more accurately (lower perplexity) would better capture neural activity.

To test this hypothesis, we used electrocorticography (ECoG) to measure neural activity in ten epilepsy patient participants while they listened to a 30-minute audio podcast. Invasive ECoG recordings more directly measure neural activity than non-invasive neuroimaging modalities like fMRI, with much higher temporal resolution. We extracted contextual embeddings at each hidden layer from multiple families of transformer-based LLMs, including GPT-2, GPT-Neo, OPT, and Llama 2 (Black et al., 2022; Radford et al., 2019; Touvron et al., 2023; S. Zhang et al., 2022), and fit electrode-wise encoding models to predict neural activity for each word in the podcast stimulus. We found that larger language models, with greater expressivity and lower perplexity, better predicted neural activity (Antonello et al., 2023). This result was consistent across all model families. Critically, we then focus on a particular family of models (GPT-Neo), which span a

broad range of sizes and are trained on the same text corpora. This allowed us to assess the effect of scaling on the match between LLMs and the human brain while keeping the size of the training set constant.

Results

To investigate scaling effects between model size and the alignment of the model internal representations (embeddings) with brain activity, we utilized four families of transformer-based language models: GPT-2, GPT-Neo, OPT, and Llama 2 (Black et al., 2022; Radford et al., 2019; Touvron et al., 2023; S. Zhang et al., 2022). These models span 82 million to 70 billion parameters and 6 to 80 layers (Table 1). Different families of models vary in architectural details and are trained on different text corpora. To control for these confounding variables, we also focused on the GPT-Neo family (Gao et al., 2020) with a comprehensive range of models that vary only in size (but not training data), spanning 125 million to 20 billion parameters. For simplicity, we renamed the four models as "SMALL" (gpt-neo-125M), "MEDIUM" (gpt-neo-1.3B), "LARGE" (gpt-neo-2.7B), and "XL" (gpt-neox-20b).

We collected ECoG data from ten epilepsy patients while they listened to a 30-minute audio podcast (*So a Monkey and a Horse Walk into a Bar*, 2017). We extracted high-frequency broadband power (70–200 Hz) in 200 ms bins at lags ranging from -2000 ms to +2000 ms relative to the onset of each word in the podcast stimulus. We ran a preliminary encoding analysis using non-contextual language embeddings (GloVe; Pennington et al., 2014) to select a subset of 160 language-sensitive electrodes across the cortical language network of eight patients; all subsequent analyses were performed within this set of electrodes (Goldstein et al., 2022). Using a podcast transcription, we next extracted contextual embeddings from each hidden layer across the four families of autoregressive large language models. We used the maximum context length of each word for each language model. We constructed linear, electrode-wise encoding models using contextual embeddings from every layer of each language model to predict neural activity for each word in the stimulus. We estimated and evaluated the encoding models using a 10-fold cross-validation procedure: ridge regression was used to estimate a weight matrix for predicting word-by-word neural signals in 9 out of 10 contiguous training segments of the podcast; for each electrode, we then calculated the Pearson correlation between predicted and actual word-by-word neural signals for the left-out test segment of the podcast. We repeated this analysis for 161 lags from -2,000 ms to 2,000 ms in 25 ms increments relative to word onset (Fig. 1).

Prior to encoding analysis, we measured the "expressiveness" of different language models—that is, their capacity to predict the structure of natural language. Perplexity quantifies expressivity as the average level of surprise or uncertainty the model assigns to a sequence of words. A lower perplexity value indicates a better alignment with linguistic statistics and a higher accuracy during next-word prediction. For each model, we computed perplexity values for the podcast transcript. Consistent with prior research (Hosseini et al., 2022; Kaplan et al., 2020), we found that perplexity decreases as model size increases (Fig. 2A). In simpler terms, we confirmed that larger models better predict the structure of natural language.

Larger language models better predict brain activity

We compared encoding model performance across language models at different sizes. For each electrode, we obtained the maximum encoding performance correlation across all lags and layers, then averaged these correlations across electrodes to derive the overall maximum correlation for each model (Fig. 2B). Using ECoG neural signals with superior temporal resolution, we replicated the previous fMRI work reporting a logarithmic relationship between model size and encoding performance (Antonello et al., 2023), indicating that larger models better predict neural activity. We also observed a plateau in the maximal encoding performance, occurring around 7 billion parameters (Fig. 2B), with a decline in performance for the OPT-66B model (Fig. S1). The size of the contextual embedding varies across models depending on the model's size and architecture. This can range from 762 in the smallest distill GPT2 model to 8192 in the largest LLAMA-2 70 billion parameter model. To control for the different embedding dimensionality

Model Family	Context Length	Model Name	Model Size	Number of Layers	Hidden Embedding Size
GPT2	1024	distilgpt2	82 M	6	768
		gpt2	124 M	12	768
		gpt2-medium	355 M	24	1024
		gpt2-large	774 M	32	1280
		gpt2-xl	1.5 B	48	1600
GPT-Neo	2048	EleutherAI/gpt-neo-125M	125 M	12	768
		EleutherAI/gpt-neo-1.3B	1.3 B	24	2048
		EleutherAI/gpt-neo-2.7B	2.7 B	32	2560
		EleutherAI/gpt-neox-20b	20 B	44	6144
OPT	2048	facebook/opt-125m	125 M	12	768
		facebook/opt-350m	350 M	24	1024
		facebook/opt-1.3b	1.3 B	24	2048
		facebook/opt-2.7b	2.7 B	32	2560
		facebook/opt-6.7b	6.7 B	32	4096
		facebook/opt-13b	13 B	40	5120
		facebook/opt-30b	30 B	48	7168
		facebook/opt-66b	66 B	64	9216
Llama-2	4096	meta-llama/Llama-2-7b-hf	7 B	32	4096
		meta-llama/Llama-2-13b-hf	13 B	40	5120
		meta-llama/Llama-2-70b-hf	70 B	80	8192

Table 1. Summary of four families of open large language models: GPT-2, GPT-Neo, OPT, and Llama-2.

Context length is the maximum context length for the model, ranging from 1024 to 4096 tokens. The model name is the model's name as it appears in the transformers package from Hugging Face (Wolf et al., 2019). Model size is the total number of parameters; M represents million, and B represents billion. The number of layers is the depth of the model, and the hidden embedding size is the internal width.

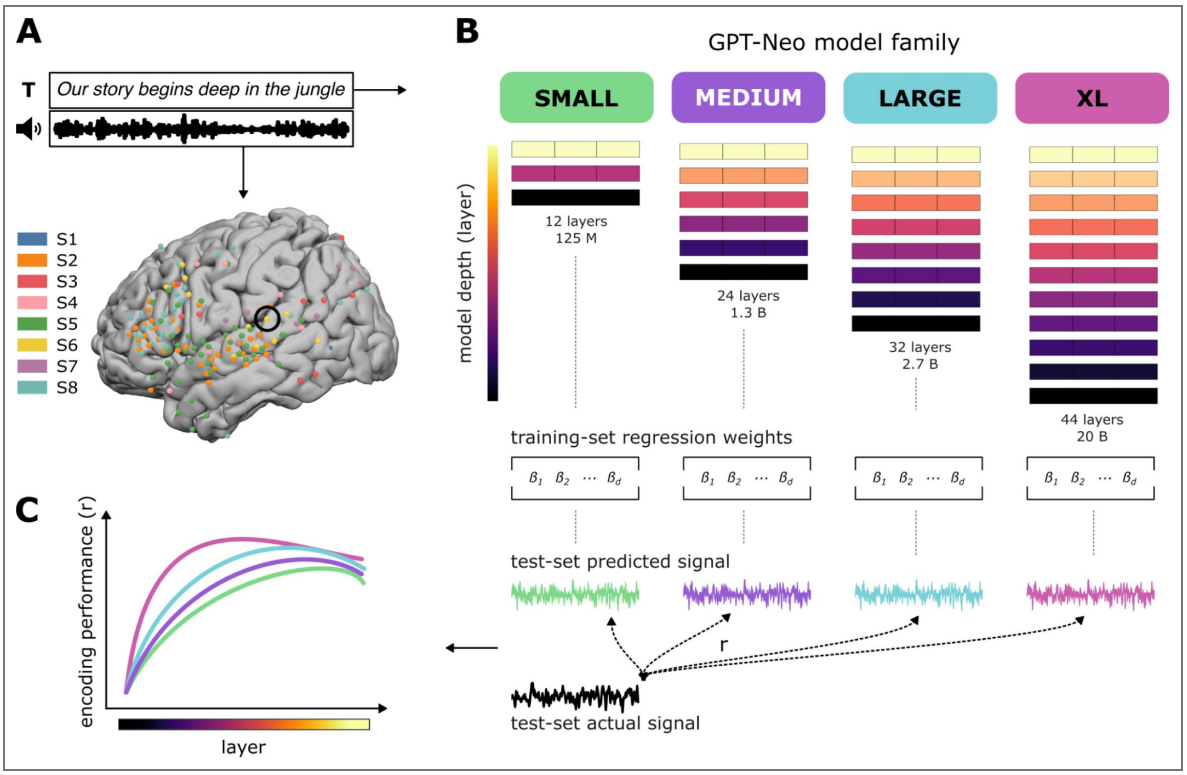


Figure 1. Naturalistic language comprehension model comparison framework.

A. Participants listened to a 30-minute story while undergoing ECoG recording. A word-level aligned transcript was obtained and served as input to four language models of varying size from the same GPT-Neo family. **B.** For every layer of each model, a separate linear regression encoding model was fitted on a training portion of the story to obtain regression weights that can predict each electrode separately. Then, the encoding models were tested on a held-out portion of the story and evaluated by measuring the Pearson correlation of their predicted signal with the actual signal. **C.** Encoding model performance (correlations) was measured as the average over electrodes and compared between the different language models.

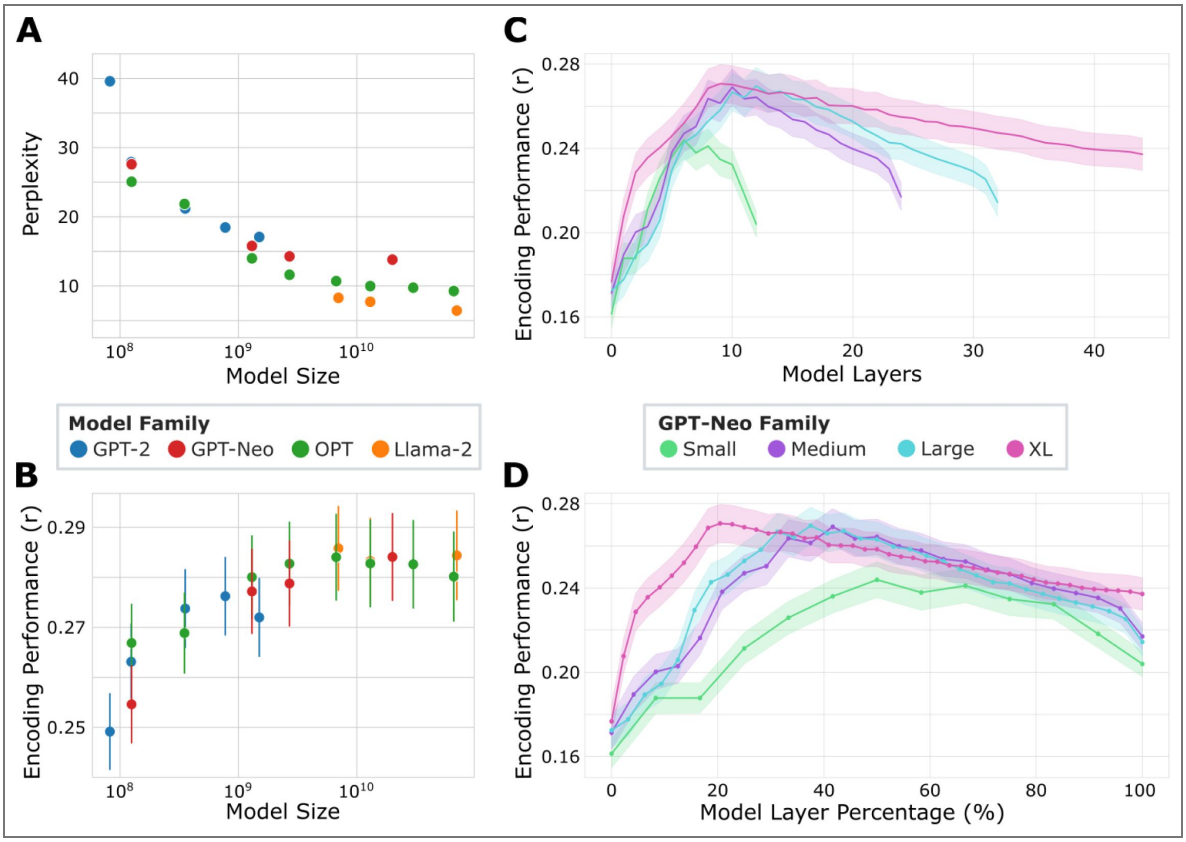


Figure 2. **Model performance improves with increasing model size.**

A. The relationship between model size (measured as the number of parameters, shown on a log scale) and perplexity: as the model size increases, perplexity decreases. Each data point corresponds to a model. **B.** The relationship between model size (shown on a log scale) and brain encoding performance: correlations for each model are calculated by averaging the maximum correlations across all lags and layers across electrodes. As the model size increases, the encoding performance increases. Each data point corresponds to a model. The error bars represent standard error. **C.** For the GPT-Neo model family, the relationship between encoding performance and layer number. Encoding performance is best for intermediate layers. The shaded colors represent standard error. **D**. Same as C, but the layer number was transformed to a layer percentage for better model comparison.

across models, we standardized all embeddings to the same size using principal component analysis (PCA) and trained linear encoding models using ordinary least-squares (OLS) regression, replicating the logarithmic relationship but with significantly lower encoding performance overall (Fig. S2). The PC features are used by the OLS models only.

To dissociate model size and control for other confounding variables, we next focused on the GPT-Neo models and assessed layer-by-layer and lag-by-lag encoding performance. For each layer of each model, we identified the maximum encoding performance correlation across all lags and averaged this maximum correlation across electrodes (Fig. 2C). Additionally, we converted the absolute layer number into a percentage of the total number of layers to compare across models (Fig. 2D). We found that correlations for all four models typically peak at intermediate layers, forming an inverted U-shaped curve, corroborating with previous fMRI findings (Caucheteux et al., 2021; Schrimpf et al., 2021; Toneva & Wehbe, 2019). Furthermore, we replicated the phenomenon observed by (Antonello et al., 2023), wherein smaller models (e.g. SMALL) achieve maximum encoding performance approximately three-quarters into the model, while larger models (e.g. XL) peak in relatively earlier layers before gradually declining. Leveraging the high temporal resolution of ECoG, we compared the encoding performance of models across various lags relative to word onset. We identified the optimal layer for each electrode and model and then averaged the encoding performance across electrodes. We found that XL significantly outperformed SMALL in encoding models for most lags from 2000 ms before word onset to 575 ms after word onset (Fig. S3). To establish a general baseline for encoding performance, we built encoding models using embeddings from the SMALL model with randomly initialized weights. The trained SMALL model exhibits significantly higher encoding performance across all layers compared to the untrained SMALL model (Fig. S4). We also assessed the encoding performance of contextual embeddings from LLMs against classic speech features and static GloVe embeddings (Table S1). The SMALL and XL embeddings achieved markedly higher encoding correlations than the speech features and GloVe embeddings (Fig. S5). We also built encoding models using subsets of the data and found that encoding performance increases as the volume of training data increases (Fig. S6).

Encoding model performance across electrodes and brain regions

Next, we examined the differences in the encoding model across electrodes and brain regions. For each of the 160 electrodes, we identified the maximum encoding performance correlation across all lags and layers (Fig. 3A). Consistent with prior studies (Goldstein et al., 2022, 2025), our encoding model for SMALL achieved the highest correlations in superior temporal gyrus (STG) and inferior frontal gyrus (IFG). We then compared the encoding performances between SMALL and the other three models, plotting the percent change in encoding performance relative to SMALL for each electrode (Fig. 3B). Across all three comparisons, we observed significantly higher encoding performance for the larger models in approximately one-third of the 160 electrodes (two-sided pairwise t-test across cross-validation folds for each electrode, $p < 0.05$, Bonferroni corrected).

We then compared the maximum correlations between SMALL and XL models across five regions of interest (ROIs) across the cortical language network (Fig. S7): middle superior temporal gyrus (mSTG, n = 28 electrodes), anterior superior temporal gyrus (aSTG, n = 13 electrodes), Brodmann area 44 (BA44, n = 19 electrodes), Brodmann area 45 (BA45, n = 26 electrodes), and temporal pole (TP, n = 6 electrodes). Encoding performance for the XL model significantly surpassed that of the SMALL model in mSTG, aSTG, BA44, and BA45 (Fig. 3C, Table S2). Additionally, we calculated the percent change in encoding performance relative to SMALL for each brain region by averaging across electrodes and plotting against model size (Fig. 3D). As model size increases, the fit to the brain nominally increases across all observed regions. However, the increase plateaued after the Medium model for regions BA45 and TP.

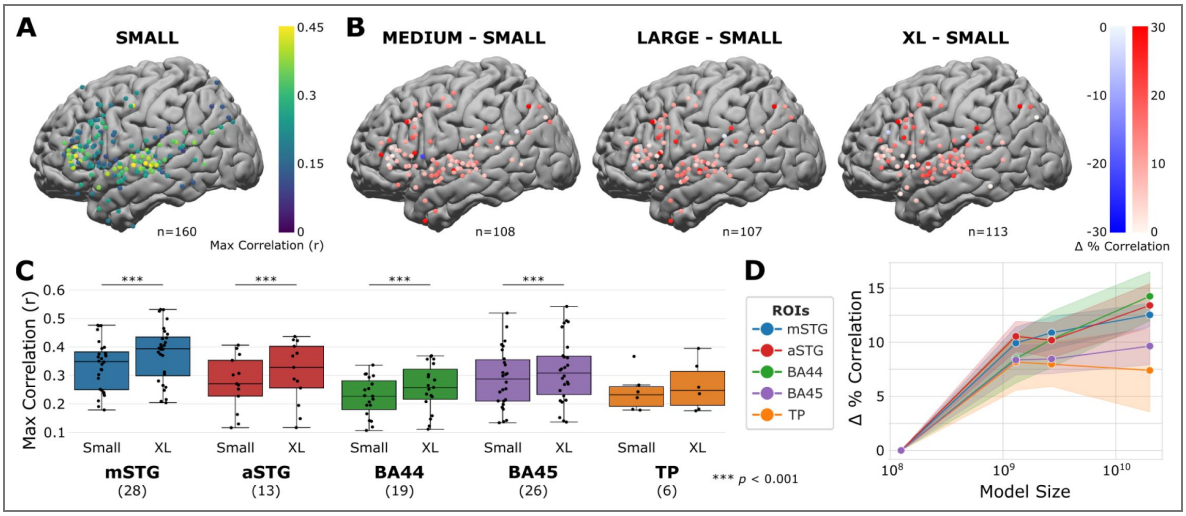


Figure 3. Model performance improves with increasing model size across electrodes and ROIs.

A. Maximum correlation per electrode for SMALL. The encoding model achieves the highest correlations in STG and IFG. **B.** For MEDIUM, LARGE, and XL, the percentage difference in correlation relative to SMALL for all electrodes with significant encoding differences. The encoding performance is significantly higher for the bigger models for almost all electrodes across the brain (pairwise t-test across cross-validation folds). **C.** Maximum encoding correlations for SMALL and XL for each ROI (mSTG, aSTG, BA44, BA45, and TP area). The encoding performance is significantly higher for XL for all ROIs except TP. Each data point corresponds to an electrode in the corresponding ROI. **D.** Percent difference in correlation relative to SMALL for all ROIs. As model size increases, the percent change in encoding performance also increases for mSTG, aSTG, and BA44. After the medium model, the percent change in encoding performance plateaus for BA45 and TP. The shaded colors represent standard error.

The best layer for encoding performance varies with model size

In the previous analyses, we observed that encoding performance peaks at intermediate to later layers for some models and relatively earlier layers for others (Fig. 1C, 1D). To examine this phenomenon more closely, we selected the best layer for each electrode based on its maximum encoding performance across lags. To account for the variation in depth across models, we computed the best layer as the percentage of each model's overall depth. We found that as models increase in size, peak encoding performance tends to occur in relatively earlier layers, being closer to the input in larger models (Fig. 4A). This was consistent across multiple model families, where we found a logarithmic relationship between model size and best encoding layers (Fig. 4B).

We further observed variations of the best encoding layers across the brain within the same model. We found that the language processing hierarchy was better reflected in the best encoding layer preference for smaller than for larger models (Fig. 4C). Specifically, in the SMALL model, peak encoding was observed in earlier layers for STG electrodes and in later layers for IFG electrodes (Fig.4C, Table S3). A similar trend is evident in MEDIUM and partially in LARGE models, but not in the XL model, where the majority of electrodes exhibited peak encoding in the first 25% of all layers (Fig. 4C). However, despite the XL model showing less variance in the best layer distributions across cortex, we found the same hierarchy present for the first half of the model (layers 0–22, Fig. 4D). In this analysis, we observed that the best relative layer nominally increases from mSTG electrodes (M = 21.916, SD = 10.556) to aSTG electrodes (M = 29.720, SD = 17.979) to BA45 (M = 30.157, SD = 16.039) and TP electrodes (M = 31.061, SD = 16.305), and finally to BA44 electrodes (M = 36.962, SD = 13.140, Fig. 4E).

The best lag for encoding performance does not vary with model size

Leveraging the high temporal resolution of ECoG, we investigated whether peak lag for encoding performance relative to word onset is affected by model size. For each ROI, we identified the optimal layer for each electrode in the ROI and then averaged the encoding performance (Fig. 5A). Within the SMALL model, we observed a trend where putatively lower-level regions of the language processing hierarchy peak earlier relative to word onset: mSTG encoding performance peaks around 25 ms before word onset, followed by aSTG encoding peak 225 ms after onset, and subsequently TP, BA44, and BA45 peak at approximately 350 ms. Within the XL model, we observed a similar trend, with mSTG encoding peaking first, followed by aSTG encoding peak, and finally, TP, BA44, and BA45 encodings. We also identified the lags when the encoding performance peaks for each electrode and visualized them on the brain map (Fig. 5B). Notably, the optimal lags for each electrode do not exhibit significant variation when transitioning from SMALL to XL.

Discussion

In this study, we investigated how the quality of model-based predictions of neural activity scales with LLM model size (i.e., the number of parameters across layers). Prior studies have shown that encoding models constructed from the internal embeddings of LLMs provide remarkably good predictions of neural activity during natural language comprehension (Caucheteux & King, 2022; Goldstein et al., 2022; Schrimpf et al., 2021). Corroborating prior work using fMRI (Antonello et al., 2023), across a range of models with 82 million to 70 billion parameters, we found that larger models are better aligned with neural activity. This result was consistent across several autoregressive LLM families varying in architectural details and training corpora and within a single model family trained on the same corpora and varying only in size. We suspect that the improved alignment with brain activity in larger models is driven by their increased expressivity and sensitivity to nuanced linguistic structure present in large-scale naturalistic datasets (Antonello et al., 2023). Our findings indicate that this trend does not trivially result from arbitrarily increasing model complexity: (a) models of varying size were estimated using regularized regression and evaluated using out-of-sample prediction to minimize the risks of

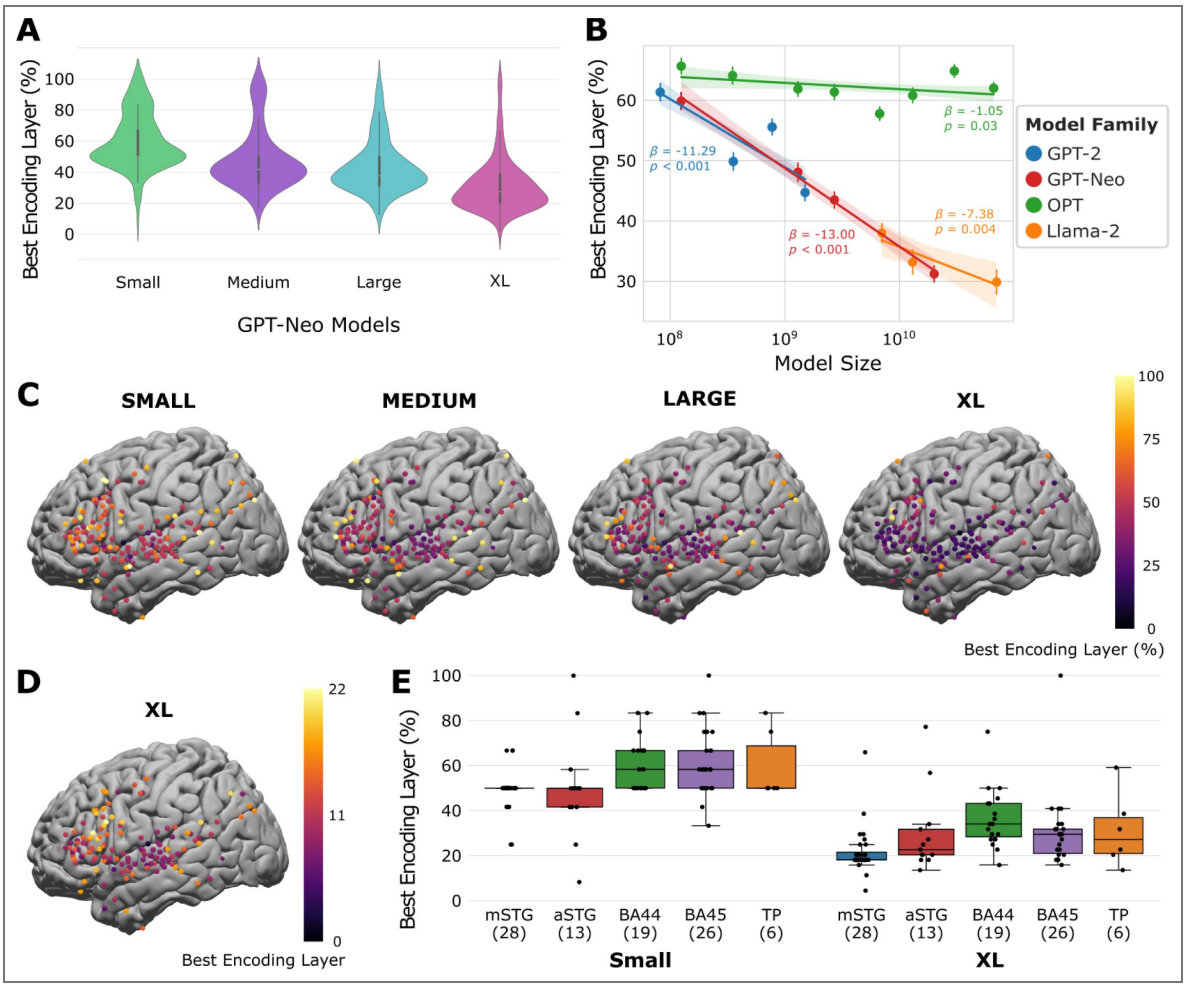


Figure 4. Relative layer preference varies with model size.

A. Relative layer (in percentage of total number of layers) with peak encoding performance for all four GPT-Neo models: the larger the model size, the earlier relative layer where the encoding performance peaks. **B.** The relationship between model size (shown on a log scale) and best encoding layer (in percentage) for all four model families: as the model size increases, the best encoding layer (in percentage) decreases, although the rate of decrease is different between model families. We estimate a linear regression model per model family of the form: best percent layer ~ log(model size). The slopes (β) indicate the decrease in the relative best-performing layer at increasing log model size; p-values are obtained from a Wald test against the null hypothesis that the slope is 0. Each data point corresponds to a model. **C.** Best relative encoding layer (in percentage) for all four GPT-Neo models. **D**. Best encoding layer for XL with electrodes that peak in the first half of the model (Layer 0 to 22). **E.** Best encoding layer (in percentage) for SMALL and XL for each ROI (mSTG, aSTG, BA44, BA45, and TP). Each data point corresponds to an electrode in the corresponding ROI.

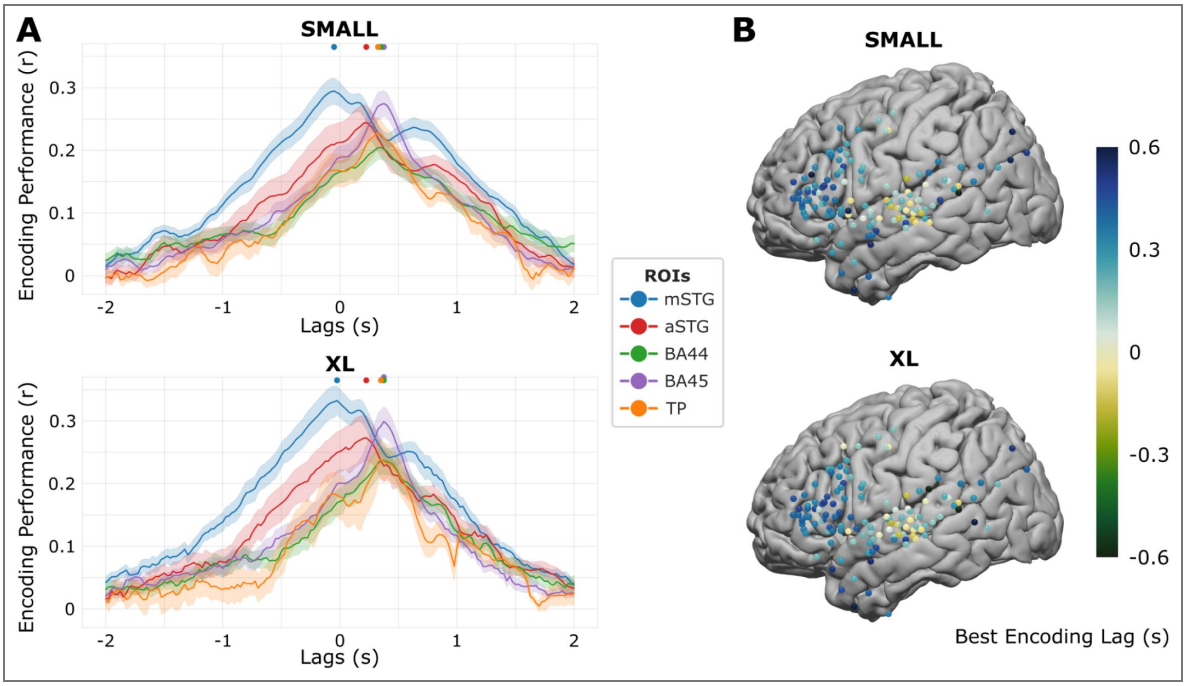


Figure 5. Encoding performance across lags does not vary with model size.

A. Average ROI encoding performance for SMALL and XL models. mSTG encoding peaks first before word onset, then aSTG peaks after word onset, followed by BA44, BA45, and TP encoding peaks at around 400 ms after onset. The dots represent the peak lag for each ROI. **B.** Lag with best encoding performance correlation for each electrode, using SMALL and XL model embeddings. Only electrodes with the best lags that fall within 600 ms before and after word onset are plotted.

overfitting, and (b) we obtained qualitatively similar results with explicitly matched dimensionality using PCA. Combined with the observation that larger models yield lower perplexity, our findings suggest that larger models' capacity for learning the structure of natural language also yields better predictions of brain activity. By leveraging their increased size and representational power, these models have the potential to provide valuable insights into the mechanisms underlying language comprehension.

We focused on a particular family of models (GPT-Neo) trained on the same corpora and varying only in size to investigate how model size impacts layerwise encoding performance across lags and ROIs. We found that model-brain alignment improves consistently with increasing model size across the cortical language network. However, the increase plateaued after the MEDIUM model for regions BA45 and TP, possibly due to already high encoding correlations for the SMALL model and a small number of electrodes in the area, respectively.

A more detailed investigation of layerwise encoding performance revealed a logarithmic relationship where peak encoding performance tends to occur in relatively earlier layers as both model size and expressivity increase (Mischler et al., 2024). This is an unexpected extension of prior work on both language (Caucheteux & King, 2022; Kumar et al., 2024; Schrimpf et al., 2021; Toneva & Wehbe, 2019) and vision (Jiahui et al., 2023), where peak encoding performance was found at late-intermediate layers. Moreover, we observed variations in best relative layers across different brain regions, corresponding to a language processing hierarchy. This is particularly evident in smaller models and early layers of larger models. The inverted U-shaped trend of encoding performance commonly found in previous research is likely due to a "two-phase abstraction process" within LLMs (Cheng & Antonello, 2024; Csordás et al., 2025). In the early and intermediate layers of the model, a composition phase occurs, where low-level input features become increasingly abstract and contextualized. The intermediate layers of the model show the highest correlation with brain activity, presumably because they capture complex semantic and contextual information in a way that generalizes well across a variety of tasks (including prediction of human neural activity) (Antonello & Huth, 2024). Subsequently, a prediction phase happens in the later layers of the model, where the representations become more specialized for the LLM's specific training objective (e.g., next-word prediction). This specialization can effectively constrict the more generalized feature representations, making these layers less optimal for predicting brain activity. Our results indicate that the initial composition phase does not take up more layers as models scale up in size. Larger models develop the necessary rich, abstract representations in the same number of layers as smaller models. Thus, as LLMs increase in size, the later layers of the model may contain representations that are increasingly divergent from the more general linguistic processing captured in brain activity. It is also possible that the later layers of larger models are overall underutilized and may not significantly contribute to benchmark performances during inference (Csordás et al., 2025; Fan et al., 2024; Gromov et al., 2024). Leveraging the high temporal resolution of ECoG, we found that putatively lower-level regions of the language processing hierarchy peak earlier than higher-level regions. However, we did not observe variations in the optimal lags for encoding performance across different model sizes. Since we exclusively employ textual LLMs, which lack inherent temporal information due to their discrete token-based nature, future studies utilizing multimodal LLMs integrating continuous audio or video streams, like Whisper or WavLM, may better unravel the relationship between model size and temporal dynamic representations in LLMs (Goldstein et al., 2025; Millet et al., 2023; Vaidya et al., 2022).

Our podcast stimulus comprised ~5,000 words over a roughly 30-minute episode. Although this is a rich language stimulus, naturalistic stimuli of this kind have relatively low power for modeling infrequent linguistic structures (Hamilton & Huth, 2020). While perplexity for the podcast stimulus continued to decrease for larger models, we observed a plateau in predicting brain activity for the largest LLMs. The largest models learn to capture relatively nuanced or rare linguistic structures, but these may occur too infrequently in our stimulus to capture much variance in brain activity. Using a subsampling approach, we found that encoding performance also scales with the volume (and diversity) of language stimuli. Encoding performance may

continue to increase for the largest models with more extensive stimuli (Antonello et al., 2023), motivating future work to pursue dense sampling with numerous, diverse naturalistic stimuli (Goldstein et al., 2025; LeBel et al., 2023).

The advent of deep learning has marked a tectonic shift in how we model brain activity in more naturalistic contexts, such as real-world language comprehension (Hasson et al., 2020; Richards et al., 2019). Traditionally, neuroscience has sought to extract a limited set of interpretable rules to explain brain function. However, deep learning introduces a new class of highly parameterized models that can challenge and enhance our understanding. The vast number of parameters in these models allows them to achieve human-like performance on complex tasks like language comprehension and production. It is important to note that LLMs have fewer parameters than the number of synapses in any human cortical functional network. Furthermore, the complexity of what these models learn enables them to process natural language in real-life contexts as effectively as the human brain does. Thus, the explanatory power of these models is in achieving such expressivity based on relatively simple computations in pursuit of a relatively simple objective function (e.g., next-word prediction). As in the human brain, while scaling alone may yield emergent cognitive abilities (Cantlon & Piantadosi, 2024; Herculano-Houzel, 2012), specialized architectural features likely also play a critical role (Friederici & Becker, 2025). As we continue to develop larger, more sophisticated models, the scientific community is tasked with advancing a framework for understanding these models to better understand the intricacies of the neural code that supports natural language processing in the human brain.

Materials and Methods

Participants

Ten patients (6 female, 20-48 years old) with treatment-resistant epilepsy undergoing intracranial monitoring with subdural grid and strip electrodes for clinical purposes participated in the study. Two patients consented to have an FDA-approved hybrid clinical research grid implanted, which includes standard clinical electrodes and additional electrodes between clinical contacts. The hybrid grid provides a broader spatial coverage while maintaining the same clinical acquisition or grid placement. All participants provided informed consent following the protocols approved by the Institutional Review Board of the New York University Grossman School of Medicine. The patients were explicitly informed that their participation in the study was unrelated to their clinical care and that they had the right to withdraw from the study at any time without affecting their medical treatment. One patient was removed from further analyses due to excessive epileptic activity and low SNR across all experimental data collected during the day.

Stimuli

Participants listened to a 30-minute auditory story stimulus, “So a Monkey and a Horse Walk Into a Bar: Act One, Monkey in the Middle,” (*So a Monkey and a Horse Walk into a Bar*, 2017) from the This American Life Podcast (Chivvis, 2017). The audio narrative is 30 minutes long and consists of approximately 5000 words. Participants were not explicitly aware that we would examine word prediction in our subsequent analyses. The onset of each word was marked using the Penn Phonetics Lab Forced Aligner (Yuan & Liberman, 2008) and manually validated and adjusted as needed. The stimulus and alignment processes are described in prior work (Goldstein et al., 2022). In this study, we use the term “structure” to refer to a variety of linguistic patterns (e.g., morphology, syntax, semantics, context) that LLMs encode.

Data acquisition and preprocessing

Across all patients, 1106 electrodes were placed on the left and 233 on the right hemispheres (signal sampled at or downsampled to 512 Hz). Brain activity was recorded from a total of 1339 intracranially implanted subdural platinum-iridium electrodes embedded in silastic sheets (2.3mm diameter contacts, Ad-Tech Medical Instrument; for the hybrid grids, 64 standard contacts had a diameter of 2 mm and an additional 64 contacts were 1mm diameter, PMT corporation,

Chananssen, MN). We also preprocessed the neural data to get the power in the high-gamma-band activity (70-200 HZ). The full description of ECoG recording procedure is provided in prior work (Goldstein et al., 2022).

Electrode-wise preprocessing consisted of four main stages: First, large spikes exceeding four quartiles above and below the median were removed, and replacement samples were imputed using cubic interpolation. Second, the data were re-referenced using common average referencing. Third, 6-cycle wavelet decomposition was used to compute the high-frequency broadband (HFBB) power in the 70–200 Hz band, excluding 60, 120, and 180 Hz line noise. In addition, the HFBB time series of each electrode was log-transformed and z-scored. Fourth, the signal was smoothed using a Hamming window with a kernel size of 50 ms. The filter was applied in both the forward and reverse directions to maintain the temporal structure. Additional preprocessing details can be found in prior work (Goldstein et al., 2022).

Electrode selection

We used a nonparametric statistical procedure with correction for multiple comparisons(Nichols & Holmes, 2002) to identify significant electrodes. We randomized each electrode's signal phase at each iteration by sampling from a uniform distribution. This disconnected the relationship between the words and the brain signal while preserving the autocorrelation in the signal. We then performed the encoding procedure for each electrode (for all lags). We used GloVe embeddings for electrode selection to avoid biasing our main results toward a particular LLM. We repeated the encoding process 5000 times. After each iteration, the encoding model's maximal value across all lags was retained for each electrode. We then took the maximum value for each permutation across electrodes. This resulted in a distribution of 5000 values, which was used to determine the significance for all electrodes. For each electrode, a *p*-value was computed as the percentile of the non-permuted encoding model's maximum value across all lags from the null distribution of 5000 maximum values. Performing a significance test using this randomization procedure evaluates the null hypothesis that there is no systematic relationship between the brain signal and the corresponding word embedding. This procedure yielded a *p*-value per electrode, corrected for the number of models tested across all lags within an electrode. To further correct for multiple comparisons across all electrodes, we used a false-discovery rate (FDR). Electrodes with *q*-values less than .01 are considered significant. This procedure identified 160 electrodes from eight patients in the left hemisphere's early auditory, motor cortex, and language areas.

Perplexity

We computed the perplexity values for each LLM using our story stimulus, employing a stride length half the maximum token length of each model (stride 512 for GPT-2 models, stride 1024 for GPT-Neo models, stride 1024 for OPT models, and stride 2048 for Llama-2 models). These stride values yield the lowest perplexity value for each model. We also replicated our results on fixed stride length across model families (stride 512, 1024, 2048, 4096).

Contextual embeddings

We extracted contextual embeddings from all layers of four families of autoregressive large language models. The GPT-2 family, particularly *gpt2-xl*, has been extensively used in previous encoding studies (Goldstein et al., 2022; Schrimpf et al., 2021). Here we include distilGPT-2 as part of the GPT-2 family. The GPT-Neo family, released by EleutherAI (Black et al., 2022), features three models plus GPT-Neox-20b, all trained on the Pile dataset (Gao et al., 2020). The OPT and Llama-2 families are released by MetaAI (Touvron et al., 2023; S. Zhang et al., 2022). For Llama-2, we use the pre-trained versions before any reinforcement learning from human feedback. All models we used are implemented in the HuggingFace environment (Wolf et al., 2019). We define "model size" as the combined width of a model's hidden layers and its number of layers, determining the total parameters. We first converted the words from the raw transcript (including punctuation and capitalization) to tokens comprising whole words or sub-words (e.g., (1) there's → (1) there (2) 's). All models within the same model family adhere to the same

tokenizer convention, except for GPT-Neox-20B, which utilizes a different tokenizer (Black et al., 2022). For each token, we utilized a context window with the maximum context length of each language model containing prior tokens from the podcast (i.e., the token and its history) and extracted the embedding for the final token in the sequence (i.e., the token itself). To facilitate a fair comparison of the encoding effect across different models, we aligned all tokens in the story across all models. We averaged the token embeddings if a word is split into several tokens, resulting in one embedding per word for each model.

Transformer-based language model consists of blocks containing a self-attention sub-block and a subsequent feedforward sub-block. The output of a block is obtained through a residual connection applied to the sum of the block's input and the output of the feedforward sub-block. The self-attention output is added to this sum later in the layer normalization step. This output is commonly referred to as the "hidden state" of language models. This hidden state is considered the contextual embedding for the preceding block. For convenience, we refer to the blocks as "layers"; that is, the hidden state at the output of block 3 is referred to as the contextual embedding for layer 3. 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 those language models automatically stores these hidden states when a forward pass of the model is conducted. Different models have different numbers of layers and embeddings of different dimensionality. For instance, gpt-neo-125M (SMALL) has 12 layers, and the embeddings at each layer are 768-dimensional vectors. Since we generate an embedding for each word at every layer, this results in 12 768-dimensional embeddings per word.

Static embeddings

We constructed static embeddings with classic speech features and GloVe to compare with the encoding performance of LLMs. First, we extracted features capturing lower-level acoustic qualities of speech. Using the stimulus transcript as input, we created one-hot vectors for phonetic and articulatory features. Phoneme classes (39 total classes) were obtained from the Carnegie Mellon Pronouncing Dictionary (*The CMU Pronouncing Dictionary*, n.d.). We further classified the phonemes based on their place of articulation (9 classes), manner of articulation (9 classes), and voiced or voiceless status (3 classes), according to the general American English consonants of the International Phonetic Alphabet. Given that each word consists of multiple phonemes, we averaged the one-hot vectors for all phonetic and articulatory features for each word. Second, we extracted linguistic features using spaCy (Honnibal et al., 2020), including part of speech (17 classes), tag (50 classes), function or content word (3 classes), dependency (45 classes), whether the word is an alpha character (binary), and whether the word is a stop word (binary). We also extracted prefix (30 classes) and suffix (44 classes) information using the Cambridge Dictionary. We constructed one-hot vectors for each multi-class feature and one-dimensional vectors for each binary feature. Third, for each word, we obtained word frequency from the Google Web Trillion Word Corpus (Brants & Franz, 2006) and from our own dataset. Fourth, we generated static word embeddings of dimension 50 using GloVe (Pennington et al., 2014).

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 (i.e., for each lag, the model learns to predict the average single +-100 ms around the lag). We estimated and evaluated the encoding models using a 10-fold cross-validation procedure: ridge regression was used to estimate a weight matrix for predicting word-by-word neural signals in 9 out of 10 contiguous training segments of the podcast; for each electrode, we then calculated the Pearson correlation between predicted and actual neural signals for the left-out test segment of the podcast. For each ridge regression model (for each fold, lag, and electrode), the alpha parameter is determined by cross-validation using the "RidgeCV" method from the "himalaya" package (Dupré la Tour et al., 2022). This procedure was performed for all layers of

contextual embeddings from each LLM. To control for the different embedding dimensionality across models, we standardized all embeddings to the same size using principal component analysis (PCA) and trained linear encoding models using ordinary least-squares (OLS) regression. The PC features are used by the OLS models only.

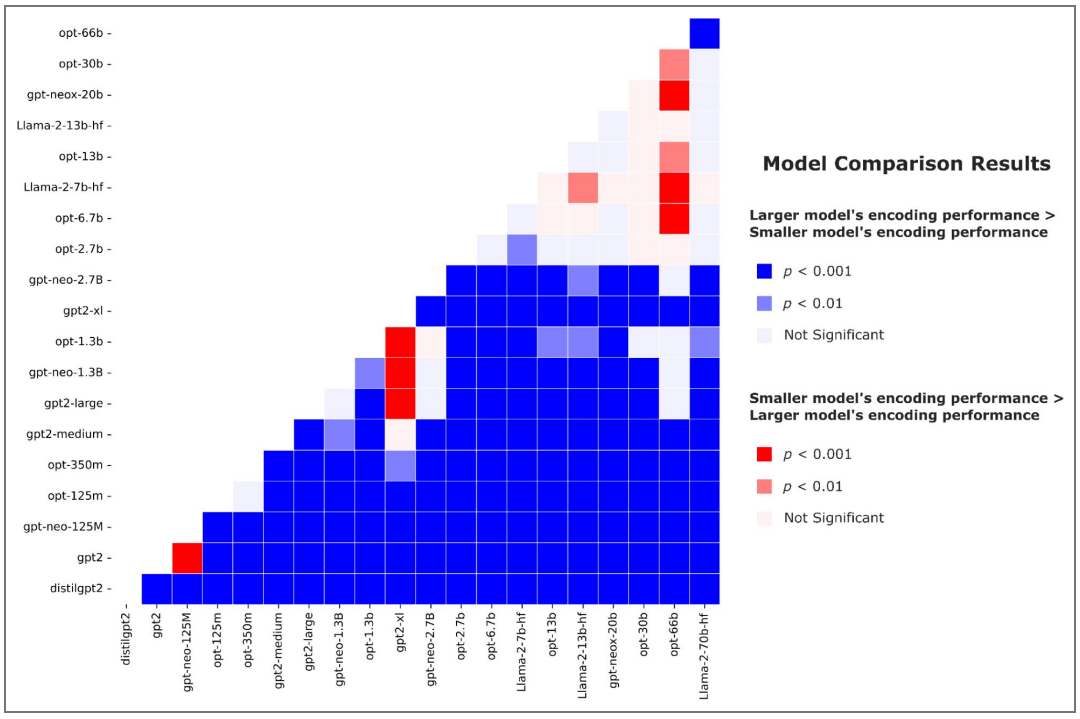
Dimensionality reduction

To control for the different hidden embedding sizes across models, we standardized all embeddings to the same size using principal component analysis (PCA) and trained linear regression encoding models using ordinary least-squares regression, replicating all results (Fig. S2). 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. We compute PCA separately on the training and testing set to avoid data leakage.

Data availability

We have recently made the data publicly available (Zada et al., 2025). We have also provided tutorials for preprocessing the data and training encoding models: *https://hassonlab.github.io/podcast-ecog-tutorials*. For this specific project, the analysis code is available at *https://github.com/hassonlab/247-pickling/tree/scaling-paper-0* and *https://github.com/hassonlab/247-encoding/tree/scaling-paper-1*.

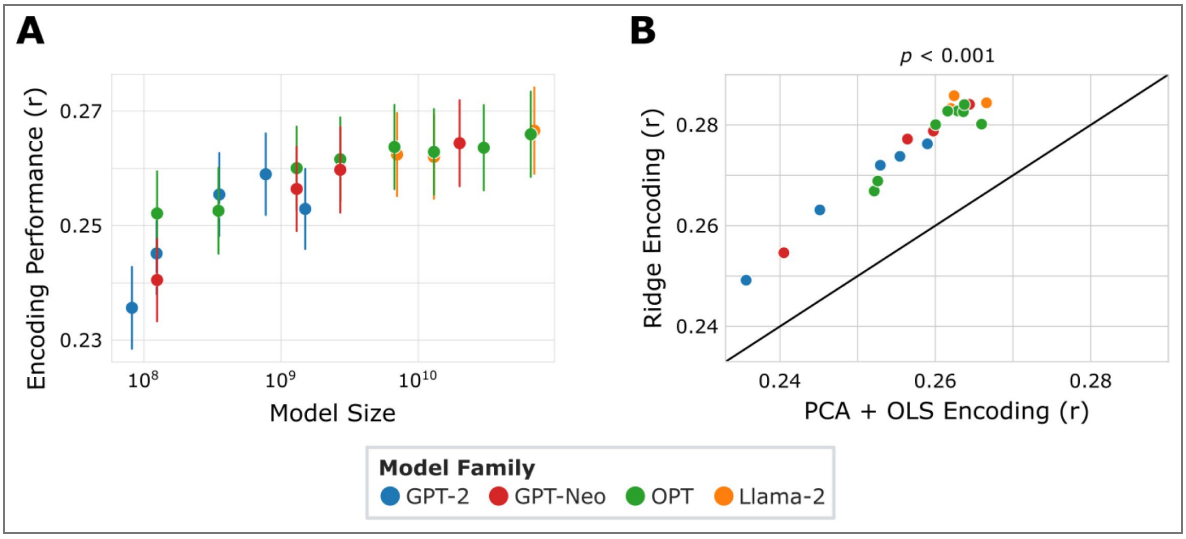
Supplementary figures and tables



Supplementary Figure 1. Heatmap comparing encoding performance between models. Encoding performance for a model is represented by the maximum correlation across lags and layers per electrode. Paired t-tests are performed across electrodes (df = 159). The result is blue if t < 0, meaning the larger model (the model with a higher number of parameters, represented on the x-axis) outperforms the smaller model (the model with a smaller number of parameters, represented on the y-axis). The result is red if t > 0, meaning the smaller model outperforms the larger model. The shades of the colors represent significance (p < 0.001 or p < 0.01 or not significant, two-sided, FDR corrected). There is a positive relationship between model size and encoding performance when models are smaller than 3 billion parameters. When models are larger than 7 billion parameters, we observed a plateau in the maximal encoding performance.

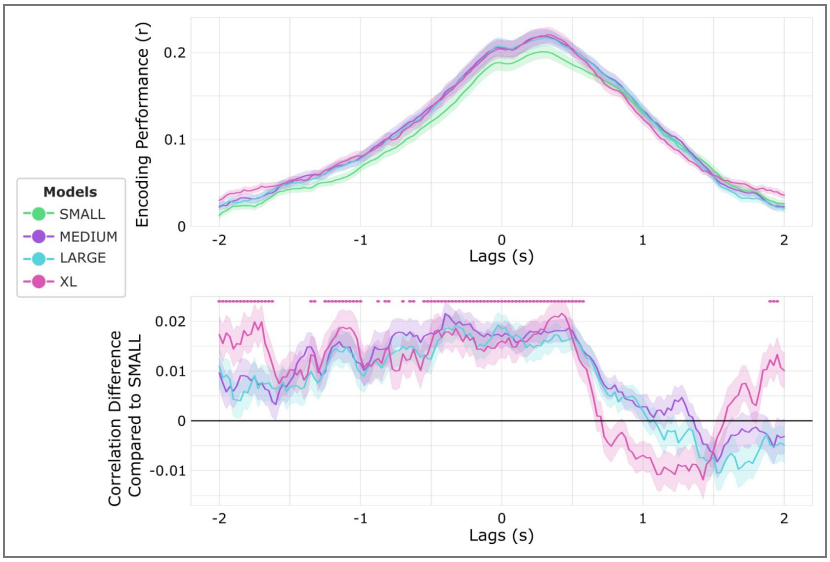
Supplementary Figure 2. Model performance improves with increasing model size.

To control for the different embedding dimensionality across models, we standardized all embeddings to the same size using principal component analysis (PCA) and trained linear encoding models using ordinary least-squares (OLS) regression (cf. Fig. 2). **A.** Replication of Fig. 2B, the relationship between model size (shown on a log scale) and brain encoding performance: encoding performance increases as model size increases. Each data point corresponds to a model. **B.** Ridge regression encoding outperforms PCA + OLS regression encoding for all 20 transformer-based language models (paired two-sided *t*-test across electrodes, df = 159, $p < 0.001$, Bonferroni corrected). Each data point corresponds to a model.



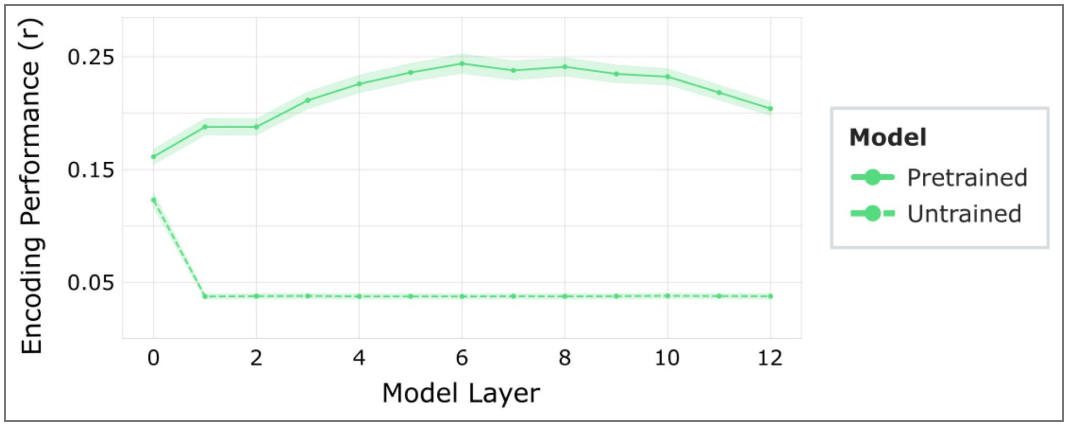
Supplementary Figure 3. Lag-wise encoding for the GPT-Neo Family.

Top. Lag-wise encoding for all four models of the GPT-Neo family, averaged across electrodes. The dots represent lags where XL significantly outperformed Small (paired two-sided *t*-test across electrodes, df = 159, $p < 0.001$, Bonferroni corrected). XL significantly outperformed Small in encoding models for most lags from 2000 ms before word onset to 575 ms after word onset. **Bottom.** Lag-wise encoding difference for the three bigger models compared to SMALL, averaged across electrodes.



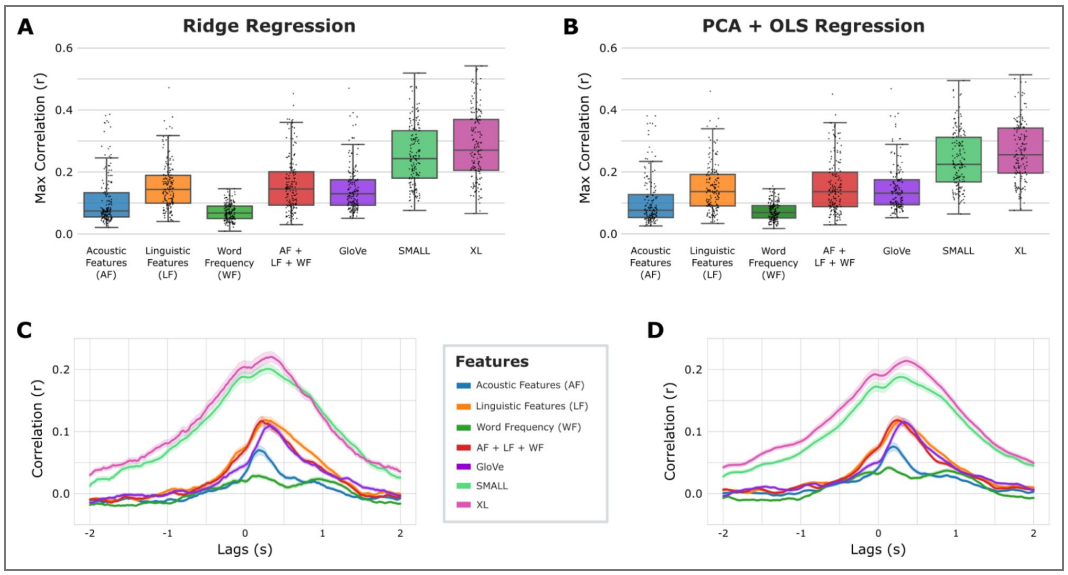
Supplementary Figure 4. The relationship between encoding performance and layer number for pretrained and untrained SMALL.

For pretrained SMALL, encoding performance is best for intermediate layers. For untrained SMALL with randomly initialized weights, encoding performance is best for the 0th layer. Encoding performance is significantly higher for pretrained SMALL than for untrained SMALL for every layer (paired two-sided t-test across electrodes, df = 159, $p < 0.001$, Bonferroni corrected). The shaded colors represent standard error.



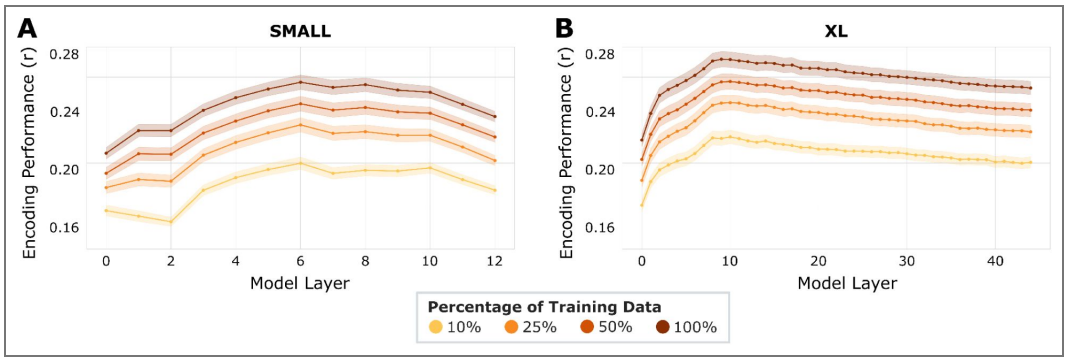
Supplementary Figure 5. Contextual embeddings from LLMs outperform classical speech features and GloVe embeddings.

A. Maximum encoding performance (across lags) for acoustic features (60 dimensions), linguistic features (191 dimensions), word frequency (2 dimensions), acoustic features + linguistic features + word frequency (253 dimensions), GloVe (50 dimensions), SMALL (768 dimensions), and XL (6144 dimensions) using ridge regression. SMALL and XL showed significantly better performance than other embeddings (paired two-sided t-test across electrodes, df = 159, $p < 0.001$). **B**. Replication of A using PCA and OLS regression. To control for the different dimensions of the different feature spaces, we standardized all feature sets to 50 dimensions (2 dimensions for word frequency) using PCA and trained OLS regression encoding models. SMALL and XL showed significantly better performance than other embeddings (paired two-sided t-test across electrodes, df = 159, $p < 0.001$). **C**. Encoding performance across lags for all features. Shaded colors represent standard error across electrodes. **D.** Replication of C using PCA and OLS regression.



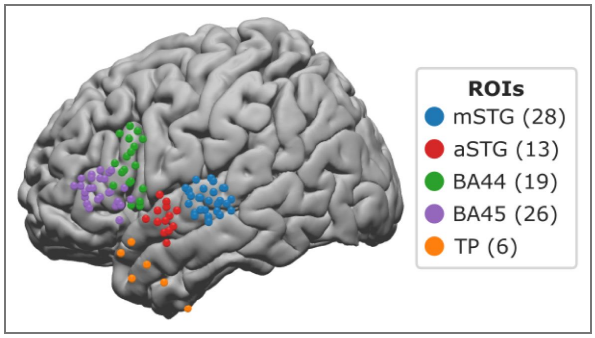
Supplementary Figure 6. Model performance improves with increasing dataset size.

A. For SMALL, the relationship between the percentage of training data and brain encoding performance across layers. Encoding performance increases as the training dataset size increases. The shaded colors represent standard error across electrodes. B. Same as A, but for model XL.



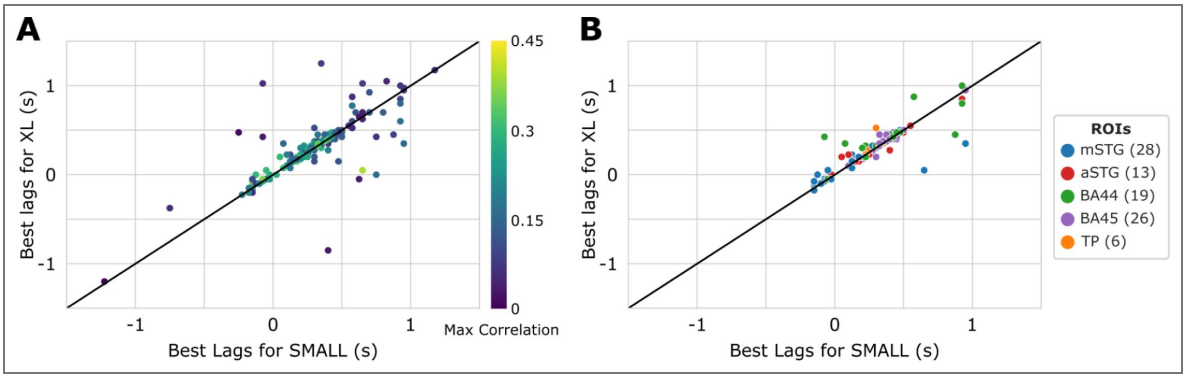
Supplementary Figure 7. Brain map of electrodes in five regions of interest (ROIs) across the cortical language network:

middle superior temporal gyrus (mSTG, n = 28 electrodes), anterior superior temporal gyrus (aSTG, n = 13 electrodes), Brodmann area 44 (BA44, n = 19 electrodes), Brodmann area 45 (BA45, n = 26 electrodes), and temporal pole (TP, n = 6 electrodes).



Supplementary Figure 8. The optimal lags for each electrode do not exhibit significant variation when transitioning between SMALL and XL models.

A. Scatter plot of best-performing lag for SMALL and XL models, colored by max correlation. Each data point corresponds to an electrode. **B.** Scatter plot of best-performing lag for SMALL and XL models, colored by ROIs. Each data point corresponds to an electrode. Only the electrodes in Fig. S7 are included.



Supplementary Table 1. **Classic speech features.**

Lower-level acoustic features include phoneme (39 classes), place of articulation (9 classes), manner of articulation (9 classes), and voiced or voiceless (3 classes). Linguistic features include part of speech (17 classes), tag (50 classes), function or content (3 classes), prefix (30 classes), suffix (44 classes), dependency (45 classes), whether the word is an alpha character (binary), and whether the word is a stop word (binary). Word frequency includes the frequency of the word in the Google Web Trillion Word Corpus and in our dataset.

Feature Family	Feature	Dimension
Acoustic Features	Phonemes	39
	Place of Articulation	9
	Manner of Articulation	9
	Voiced or voiceless	3
Linguistic Features	Part of Speech (POS)	17
	Tag (detailed POS)	50
	Function or Content	3
	Prefix	30
	Suffix	44
	Dependency	45
	Is alpha word	1
	Is stop word	1
Word Frequency	Google Web Trillion Word Corpus	1
	In our dataset	1

Supplementary Table 2. **Summary statistics and paired *t*-test results for maximum correlations between SMALL and XL models across five regions of interest.**

Encoding performance for the XL model significantly surpassed that of the SMALL model in whole brain, mSTG, aSTG, BA44, and BA45.

ROI	electrode num	SMALL corr mean	SMALL corr std	XL corr mean	XL corr std	corr diff t-value	corr diff p-value
all	160	0.255	0.097	0.284	0.109	16.791	$\mathbf{4.770e^{-37}}$
mSTG	28	0.335	0.087	0.377	0.098	11.584	$\mathbf{5.538e^{-12}}$
aSTG	13	0.274	0.096	0.311	0.108	6.111	$\mathbf{5.247e^{-5}}$
BA44	19	0.224	0.068	0.257	0.082	6.340	$\mathbf{5.663e^{-6}}$
BA45	26	0.289	0.105	0.316	0.113	6.107	$\mathbf{2.205e^{-6}}$
TP	6	0.243	0.086	0.263	0.086	1.991	0.103

ROIs	mSTG $M = 49.107$ $SD = 9.168$	aSTG $M = 50.000$ $SD = 22.82$	BA44 $M = 60.965$ $SD = 11.471$	BA45 $M = 60.897$ $SD = 14.485$	TP $M = 59.722$ $SD = 15.290$
aSTG	$t(39) = -0.180$ $p = 0.429$				
BA44	$t(45) = -3.930$ $p = \mathbf{1.450e^{-5}}$	$t(30) = -1.797$ $p = \mathbf{0.041}$			
BA45	$t(52) = -3.601$ $p = \mathbf{3.538e^{-5}}$	$t(37) = -1.820$ $p = \mathbf{0.038}$	$t(43) = 0.017$ $p = 0.507$		
TP	$t(32) = -2.276$ $p = \mathbf{0.0148}$	$t(17) = -0.943$ $p = 0.179$	$t(23) = 0.214$ $p = 0.584$	$t(30) = 0.177$ $p = 0.570$	

Supplementary Table 3. Summary statistics and paired *t*-test results for best-performing layers (in percentage) for the SMALL model across five regions of interest.

The best-performing layer (in percentage) occurred earlier for electrodes in mSTG and aSTG and later for electrodes in BA44, BA45, and TP.

Acknowledgements

This work was supported by the National Institutes of Health under award numbers DP1HD091948 (to A.G., Z.H., H.W., Z.Z., B.A., L.N., A.F., and U.H.), R01NS109367 (to A.F.), and R01DC022534 (to S.A.N.), Finding a Cure for Epilepsy and Seizures (FACES), and Schmidt Futures Foundation DataX Fund. Z.H. devised the project, performed experimental design and data analysis, and wrote the article; H.W. devised the project, performed experimental design and data analysis, and wrote the article; Z.Z. devised the project, performed experimental design and data analysis, and critically revised the article; H.G. performed data analysis; B.A. performed data analysis; L.N. performed data analysis; W.D. devised the project; S.D. devised the project; P.D. devised the project; D.F. devised the project; O.D. devised the project; A.F. devised the project; U.H. devised the project, performed experimental design, and critically revised the article; S.A.N. devised the project, performed experimental design, wrote and critically revised the article; A.G. devised the project, performed experimental design, and critically revised the article.

Additional information

Funding

Funder	Grant reference number	Author
HHS | National Institutes of Health (NIH)	DP1HD091948	Zhuoqiao Hong Haocheng Wang Uri Hasson Ariel Y Goldstein
HHS | National Institutes of Health (NIH)	R01DC022534	Samuel Nastase
HHS | National Institutes of Health (NIH)	R01NS109367	Adeen Flinker

Author ORCID iDs

Haocheng Wang: https://orcid.org/0009-0000-8500-6054
Adeen Flinker: https://orcid.org/0000-0003-1247-1283
Samuel A Nastase: https://orcid.org/0000-0001-7013-5275

References

1. **Antonello R,** Huth A (2024) Predictive coding or just feature discovery? An alternative account of why language models fit brain data. *Neurobiology of Language (Cambridge, Mass.)* **5**:64-79 https://doi.org/10.1162/nol_a_00087 | PubMed
2. **Antonello R,** Huth A, Vaidya A (2023) Scaling laws for language encoding models in fMRI. *Advances in Neural Information Processing Systems* **36**:21895-21907 https://doi.org/10.48550/arxiv.2305.11863 | PubMed
3. **Black S,** Biderman S, Hallahan E, Anthony Q, Gao L, Golding L, He H, Leahy C, McDonell K, Phang J, *et al.* (2022) GPT-NeoX-20B: An Open-Source Autoregressive Language Model. In: *Proceedings of BigScience Episode #5 -- Workshop on Challenges & Perspectives in Creating Large Language Models* Proceedings of BigScience Episode #5 -- Workshop on Challenges & Perspectives in Creating Large Language Models. https://doi.org/10.18653/v1/2022.bigscience-1.9
4. **Bommasani R,** Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, *et al.* (2021) On the Opportunities and Risks of Foundation Models. *arXiv* https://doi.org/10.48550/arxiv.2108.07258
5. **Brants T,** Franz A (2006) *Web 1T 5-gram Version 1* [Dataset]. *Linguistic Data Consortium* https://doi.org/10.35111/CQPA-A498

6. **Brown TB,** Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, *et al.* (2020) Language Models are Few-Shot Learners. *arXiv* https://doi.org/10.48550/arXiv.2005.14165
7. **Cantlon JF,** Piantadosi ST (2024) Uniquely human intelligence arose from expanded information capacity. *Nature Reviews Psychology* 1-19 https://doi.org/10.1038/s44159-024-00283-3
8. **Caucheteux C,** Gramfort A, King J.-R (2021) GPT-2's activations predict the degree of semantic comprehension in the human brain. *bioRxiv* https://doi.org/10.1101/2021.04.20.440622
9. **Caucheteux C,** King J.-R (2022) Brains and algorithms partially converge in natural language processing. *Communications Biology* **5**:134 https://doi.org/10.1038/s42003-022-03036-1 | PubMed
10. **Cheng E,** Antonello RJ (2024) Evidence from fMRI supports a two-phase abstraction process in language models. *arXiv* https://doi.org/10.48550/ARXIV.2409.05771
11. **Csordás R,** Manning CD, Potts C (2025) Do language models use their depth efficiently?. *arXiv* https://doi.org/10.48550/ARXIV.2505.13898
12. **Dupré la Tour T,** Eickenberg M, Nunez-Elizalde AO, Gallant JL (2022) Feature-space selection with banded ridge regression. *bioRxiv* https://doi.org/10.1101/2022.05.05.490831
13. **Fan S,** Jiang X, Li X, Meng X, Han P, Shang S, Sun A, Wang Y, Wang Z (2024) Not all Layers of LLMs are Necessary during Inference. In: Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence. https://doi.org/10.24963/ijcai.2024/566
14. **Friederici AD,** Becker Y (2025) The core language network separated from other networks during primate evolution. *Nature Reviews. Neuroscience* **26**:131-132 https://doi.org/10.1038/s41583-024-00897-9 | PubMed
15. **Gao L,** Biderman S, Black S, Golding L, Hoppe T, Foster C, Phang J, He H, Thite A, Nabeshima N, *et al.* (2020) The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv* https://doi.org/10.48550/arXiv.2101.00027
16. **Goldstein A,** Grinstein-Dabush A, Schain M, Wang H, Hong Z, Aubrey B, Schain M, Nastase SA, Zada Z, Ham E, *et al.* (2024) Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature Communications* **15**:2768 https://doi.org/10.1038/s41467-024-46631-y | PubMed
17. **Goldstein A,** Wang H, Niekerken L, Schain M, Zada Z, Aubrey B, Sheffer T, Nastase SA, Gazula H, Singh A, *et al.* (2025) A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature Human Behaviour* https://doi.org/10.1038/s41562-025-02105-9 | PubMed
18. **Goldstein A,** Zada Z, Buchnik E, Schain M, Price A, Aubrey B, Nastase SA, Feder A, Emanuel D, Cohen A, *et al.* (2022) Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience* **25**:369-380 https://doi.org/10.1038/s41593-022-01026-4 | PubMed
19. **Gromov A,** Tirumala K, Shapourian H, Glorioso P, Roberts DA (2024) The Unreasonable Ineffectiveness of the Deeper Layers. *arXiv* https://doi.org/10.48550/arxiv.2403.17887
20. **Hamilton LS,** Huth AG (2020) The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience* **35**:573-582 https://doi.org/10.1080/23273798.2018.1499946 | PubMed
21. **Hasson U,** Nastase SA, Goldstein A (2020) Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron* **105**:416-434 https://doi.org/10.1016/j.neuron.2019.12.002 | PubMed
22. **Herculano-Houzel S** (2012) The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences of the United States of America* **109**:10661-10668 https://doi.org/10.1073/pnas.1201895109 | PubMed
23. **Honnibal M,** Montani I, Van Landeghem S, Boyd A (2020) spaCy: Industrial-strength Natural Language Processing in Python.

24. **Hosseini EA,** Schrimpf M, Zhang Y, Bowman S, Zaslavsky N, Fedorenko E (2022) Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *bioRxiv* 2022.10.04.510681 https://doi.org/10.1101/2022.10.04.510681

25. **Jiahui G,** Feilong M, Visconti di Oleggio Castello M, Nastase SA, Haxby JV, Gobbini MI (2023) Modeling naturalistic face processing in humans with deep convolutional neural networks. *Proceedings of the National Academy of Sciences of the United States of America* **120**:e2304085120 https://doi.org/10.1073/pnas.2304085120 | PubMed

26. **Kaplan J,** McCandlish S, Henighan T, Brown TB, Chess B, Child R, Gray S, Radford A, Wu J, Amodei D (2020) Scaling Laws for Neural Language Models. *arXiv* https://doi.org/10.48550/arxiv.2001.08361

27. **Kumar S,** Sumers TR, Yamakoshi T, Goldstein A, Hasson U, Norman KA, Griffiths TL, Hawkins RD, Nastase SA (2024) Shared functional specialization in transformer-based language models and the human brain. *Nature Communications* **15**:5523 https://doi.org/10.1038/s41467-024-49173-5 | PubMed

28. **LeBel A,** Wagner L, Jain S, Adhikari-Desai A, Gupta B, Morgenthal A, Tang J, Xu L, Huth AG (2023) A natural language fMRI dataset for voxelwise encoding models. *Scientific Data* **10**:555 https://doi.org/10.1038/s41597-023-02437-z | PubMed

29. **Linzen T,** Baroni M (2021) Syntactic Structure from Deep Learning. *Annual Review of Linguistics* **7**:195-212 https://doi.org/10.1146/annurev-linguistics-032020-051035

30. **Liu J,** Shen D, Zhang Y, Dolan B, Carin L, Chen W (2022) What makes good in-context examples for GPT-3?. In: Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures.. https://doi.org/10.18653/v1/2022.deelio-1.10

31. **Manning CD,** 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**:30046-30054 https://doi.org/10.1073/pnas.1907367117 | PubMed

32. **Millet J,** Caucheteux C, Orhan P, Boubenec Y, Gramfort A, Dunbar E, Pallier C, King J.-R (2023) Toward a realistic model of speech processing in the brain with self-supervised learning. In: NeurIPS. https://neurips.cc/virtual/2022/poster/54632

33. **Mischler G,** Li YA, Bickel S, Mehta AD, Mesgarani N (2024) Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence* **6**:1467-1477 https://doi.org/10.1038/s42256-024-00925-4

34. **Nichols TE,** Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping* **15**:1-25 https://doi.org/10.1002/hbm.1058 | PubMed

35. **Pavlick E** (2022) Semantic structure in deep learning. *Annual Review of Linguistics* **8**:447-471 https://doi.org/10.1146/annurev-linguistics-031120-122924

36. **Pennington J,** Socher R, Manning C (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.3115/v1/d14-1162

37. **Radford A,** Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language Models are Unsupervised Multitask Learners. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

38. **Richards BA,** Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, Clopath C, Costa RP, de Berker A, Ganguli S, *et al.* (2019) A deep learning framework for neuroscience. *Nature Neuroscience* **22**:1761-1770 https://doi.org/10.1038/s41593-019-0520-2 | PubMed

39. **Schrimpf M,** Blank IA, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, Tenenbaum JB, 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** https://doi.org/10.1073/pnas.2105646118 | PubMed

40. **Chivvis D** So a monkey and a horse walk into a bar. This American Life. Accessed 2017, November, 1010 https://www.thisamericanlife.org/631/so-a-monkey-and-a-horse-walk-into-a-bar

41. **Sutton R** (2019) The bitter lesson. *Incomplete Ideas*
42. **Carnegie Mellon University** (n.d.) The CMU Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict
43. **Toneva M,** Wehbe L (2019) Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In: Advances in Neural Information Processing Systems. **32** https://proceedings.neurips.cc/paper/2019/hash/749a8e6c231831ef7756db230b4359c8-Abstract.html
44. **Touvron H,** Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, *et al.* (2023) Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* https://doi.org/10.48550/arxiv.2307.09288
45. **Vaidya AR,** Jain S, Huth AG (2022) Self-supervised models of audio effectively explain human cortical responses to speech. In: Icml. **2022** https://doi.org/10.48550/ARXIV.2205.14252
46. **Wolf T,** Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, *et al.* (2019) HuggingFace's transformers: State-of-the-art natural language processing. *arXiv* https://doi.org/10.48550/ARXIV.1910.03771
47. **Xie SM,** Raghunathan A, Liang P, Ma T (2021) An explanation of in-context learning as implicit Bayesian inference. *arXiv* https://doi.org/10.48550/arXiv.2111.02080
48. **Yuan J,** Liberman M (2008) Speaker identification on the SCOTUS corpus. *The Journal of the Acoustical Society of America* **123**:3878-3878 https://doi.org/10.1121/1.2935783
49. **Zada Z,** Nastase SA, Aubrey B, Jalon I, Michelmann S, Wang H, Hasenfratz L, Doyle W, Friedman D, Dugan P, *et al.* (2025) The "Podcast" ECoG dataset for modeling neural activity during natural language comprehension. *Scientific Data* **12**:1135 https://doi.org/10.1038/s41597-025-05462-2 | PubMed
50. **Zhang C,** Bengio S, Hardt M, Recht B, Vinyals O (2021) Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM* **64**:107-115 https://doi.org/10.1145/3446776
51. **Zhang S,** Roller S, Goyal N, Artetxe M, Chen M, Chen S, Dewan C, Diab M, Li X, Lin XV, *et al.* (2022) OPT: Open Pre-trained Transformer Language Models. *arXiv* https://doi.org/10.48550/arxiv.2205.01068

Peer reviews

Reviewer #1 (Public review):

Summary:

The authors perform an analysis of the relationship between the size of an LMM and the predictive performance of an ECoG encoding model made using the representations from that LMM. They find a logarithmic relationship between model size and prediction performance, consistent with previous findings in fMRI. They additionally observe that as the model size increases, the location of the "peak" encoding performance typically moves further back into the model in terms of percent layer depth, an interesting result worthy of further analysis into these representations.

Strengths:

The evidence is quite convincing, consistent across model families and complementary to other work in this field. This sort of analysis for ECoG is needed and supports the decade-long enduring trend of the "virtuous cycle" between neuroscience and AI research, where more powerful AI models have consistently yielded more effective predictions of responses in the brain. The lag analysis showing that optimal lags do not change with model size is a nice result using the higher temporal resolution of ECoG compared to other methods like fMRI.

Comments on revised version.

After the latest revision, I am pleased to remove my previous remarks about weaknesses of the paper, as I believe the additional data scaling analysis, discussion of layerwise trends, and other additional commentary makes the paper a compelling addition to the literature.

https://doi.org/10.7554/eLife.101204.2.sa3

Reviewer #2 (Public review):

Summary:

This paper investigates whether large language models (LLMs) of increasing size more accurately align with brain activity during naturalistic language comprehension. The authors extracted word embeddings from LLMs for each word in a 30-minute story and regressed them against electrocorticography (ECoG) activity time-locked to each word as participants listened to the story. The findings reveal that larger LLMs more effectively predict ECoG activity, reflecting the scaling laws observed in other natural language processing tasks.

Strengths:

(1) The study compared model activity with ECoG recordings, which offer much better temporal resolution than other neuroimaging methods, allowing for the examination of model encoding performance across various lags relative to word onset.

(2) The range of LLMs tested is comprehensive, spanning from 82 million to 70 billion parameters. This serves as a valuable reference for researchers selecting LLMs for brain encoding and decoding studies.

(3) The regression methods used are well-established in prior research, and the results demonstrate a convincing scaling law for the brain encoding ability of LLMs. The consistency of these results after PCA dimensionality reduction further supports the claim.

Comments on revised version.

I thank the authors very much for their efforts in addressing my comments. One remaining concern is the extent of the paper's conceptual advance. Several recent studies have made broadly similar claims regarding the increasing alignment between large language models and human language processing, although using fMRI data (Antonello et al., 2023; Gao et al., 2025). I would therefore encourage the authors to more clearly articulate what additional insights are gained from using ECoG. Clarifying this point would help better establish the novelty and contribution of the present study.

Antonello, R. J., Vaidya, A. R., & Huth, A. G. (2023). Scaling laws for language encoding models in fMRI. Advances in Neural Information Processing Systems, 36, 21895-21907.

Gao, C., Ma, Z., Chen, J., Li, P., Huang, S., & Li, J. (2025). Increasing alignment of large language models with language processing in the human brain. Nature Computational Science, 5(11), 1080-1090.

https://doi.org/10.7554/eLife.101204.2.sa2

Reviewer #3 (Public review):

This manuscript studies the connection between neural activity collected through electrocorticography and hidden vector representations from autoregressive language models, with the specific aim of studying the influence of language model size on this connection. Neural activity was measured from subjects that listened to a segment from a

podcast, and the representations from language models were calculated using the written transcription as the input text. The ability of vector representations to predict neural activity was evaluated using 10-fold cross-validation with ridge regression models.

The main results are that (as well summarized in section headings):

(1) Larger models predict neural activity better.

(2) The ability of language model representations to predict neural activity differs across electrodes and brain regions.

(3) The layer that best predicts neural activity differs according to model size, with the "SMALL" model showing a correspondence between layer number and the language processing hierarchy.

(4) There seems to be a similar relationship between the time lag and the ability of language model representations to predict neural activity across models.

Strengths:

(1) The experimental and modeling protocols generally seem solid, which yielded results that answer the authors' primary research question.

(2) Electrocorticography data is especially hard to collect, so these results make a nice addition to recent functional magnetic resonance imaging studies.

Weaknesses:

(1) The interpretation of some results seems unjustified, although this may just be a presentational issue.

a) Figure 2B: The authors interpret the results as "a plateau in the maximal encoding performance," when some readers might interpret this rather as a decline after 13 billion parameters. Can this be further supported by a significance test like that shown in Figure 4B?

b) Figure S1A: It looks like the drop in PCA max correlation is larger for larger models, which may suggest to some readers that the same trend observed for ridge max correlation may not hold, contra the authors' claim that all results replicate. Why not include a similar figure as Figure 2B as part of Figure S1?

(2) Discussion of what might be driving the main result about the influence of model size appears to be missing (cf. the authors aim to provide an explanation of what seems to drive the influence of the layer location in Paragraph 3 of the Discussion section). What explanations have been proposed in the previous functional magnetic resonance imaging studies? Do those explanations also hold in the context of this study?

(3) The GloVe-based selection of language-sensitive electrodes (at least to me) isn't explained/motivated clearly enough (I think a more detailed explanation should be included in the Materials and Methods section). If the electrodes are selected based on GloVe embeddings, then isn't the main experiment just showing that representations from larger language models track more closely with GloVe embeddings? What justifies this methodology?

(4) (Minor weakness) The main experiments are largely replications of previous functional magnetic resonance imaging studies, with the exception of the one lag-based analysis. Is there anything else that the electrocorticography data can reveal that functional magnetic resonance imaging data can't?

Comments on revised version.

I reread the manuscript, my previous review, and the authors' response to it. I thank the authors for clarifying any misunderstanding from my end (e.g. the different LLM tokenizers) and feel that the authors addressed my concerns very carefully.

https://doi.org/10.7554/eLife.101204.2.sa1

Author response:

The following is the authors' response to the original reviews

> ***Public Reviews:***
>
> ***Reviewer #1 (Public review):***
>
> *Summary:*
>
> *The authors perform an analysis of the relationship between the size of an LMM and the predictive performance of an ECoG encoding model made using the representations from that LMM. They find a logarithmic relationship between model size and prediction performance, consistent with previous findings in fMRI. They additionally observe that as the model size increases, the location of the "peak" encoding performance typically moves further back into the model in terms of percent layer depth, an interesting result worthy of further analysis into these representations.*
>
> *Strengths:*
>
> *The evidence is quite convincing, consistent across model families, and complementary to other work in this field. This sort of analysis for ECoG is needed and supports the decade-long enduring trend of the "virtuous cycle" between neuroscience and AI research, where more powerful AI models have consistently yielded more effective predictions of responses in the brain. The lag analysis showing that optimal lags do not change with model size is a nice result using the higher temporal resolution of ECoG compared to other methods like fMRI.*

We thank the reviewer for their thoughtful assessment! We agree that the "virtuous cycle" between neuroscience and AI research has been, and will continue to be, a driving force in advancing our understanding of brain function through more powerful predictive models. We are especially pleased that the reviewer appreciated the lag analysis, as we view this as a valuable complement to the existing fMRI work.

> *Weaknesses:*
>
> *I would have liked to have seen the data scaling trends explored a bit too, as this is somewhat analogous to the main scaling results. While better performance with more data might be unsurprising, showing good data scaling would be a strong and useful justification for additional data collection in the field, especially given the extremely limited amount of existing language ECoG data. I realize that the data here is somewhat limited (only 30 minutes per subject), but authors could still in principle train models on subsets of this data.*

We thank the reviewer for their valuable suggestion. For the revised manuscript, we performed a new analysis where we trained encoding models using subsets of the data (randomly sampling contiguous chunks of 50%, 25%, and 10% of all words in each of the training folds) and tested these models on all words in the test fold. As expected, we found that encoding performance increases as the training dataset size increases, suggesting that model performance scales with data quantity even within the constraints of our relatively small dataset. This result reinforces the importance of collecting dense ECoG data. We have

added the following text to our Results section: “We also built encoding models using subsets of the data and found that encoding performance increases as the volume of training data increases (Fig. S6)” and included the results as a supplementary figure 6 in the revised manuscript.

> *Separately, it would be nice to have better justification of some of these trends, in particular the peak layerwise encoding performance trend and the overall upside-down U-trend of encoding performance across layers more generally. There is clearly something very fundamental going on here, about the nature of abstraction patterns in LLMs and in the brain, and this result points to that. I don't see the lack of justification here as a critical issue, but the paper would certainly be better with some theoretical explanation for why this might be the case.*

We thank the reviewer for this insightful comment. The general inverted U-shaped trend of encoding performance across layers has been a frequently observed phenomenon in studies comparing LLM representations to brain activity (Goldstein, Ham, et al., 2025; Schrimpf et al., 2021). A potential explanation is the existence of a “two-phase abstraction process” within LLMs (Cheng & Antonello, 2024; Csordás et al., 2025). In the initial layers, models begin by processing relatively low-level input features. As layers get deeper, representations become increasingly abstract and richly contextualized in semantic features relevant for understanding language. These intermediate layers often show the highest correlation with brain activity in language areas, presumably because they capture complex semantic and contextual information in a way that generalizes well across a variety of tasks (including prediction of human neural activity) (Antonello & Huth, 2024). Subsequently, a prediction phase happens in the later layers, where the representations become more specialized for the LLM's specific training objective (e.g., next-word prediction). This specialization can effectively constrict the more generalized feature representations, making these layers less optimal for predicting brain activity. These observations suggest that it is primarily the abstractive, contextual features developed in the intermediate layers of LLMs that drive their alignment with brain activity. As models become more potent at prediction, their most predictive layers (for the LLM’s natural language task) and their most generalizable layers (for brain activity) can diverge.

A key finding in our study is that the initial processing phase does not scale and take up more layers as models scale up in size and layers. Larger models develop the necessary rich, abstract representations in the same number of layers as smaller models. Consequently, the prediction phase may begin relatively earlier in these larger models, and the later layers could develop highly specialized representations that are increasingly divergent from the more general linguistic processing captured in brain activity. For example, these layers may specialize in capturing very specific patterns of language (thus lowering their perplexity) that do not actually occur often or at all in our naturalistic dataset. It is also possible that the later layers of larger models are overall underutilized and do not contribute as much to linguistic processing and next-word prediction (Csordás et al., 2025).

We have added the following text to our Discussion section:

“The inverted U-shaped trend of encoding performance commonly found in previous research is likely due to a "two-phase abstraction process" within LLMs (Cheng & Antonello, 2024; Csordás et al., 2025). In the early and intermediate layers of the model, a composition phase occurs, where low-level input features become increasingly abstract and contextualized. The intermediate layers of the model show the highest correlation with brain activity, presumably because they capture complex semantic and contextual information in a way that generalizes well across a variety of tasks (including prediction of human neural activity) (Antonello & Huth, 2024). Subsequently, a prediction phase happens in the later layers of the model, where the representations become more specialized for the LLM's specific training objective (e.g., next-word prediction). This specialization can effectively

constrict the more generalized feature representations, making these layers less optimal for predicting brain activity. Our results indicate that the initial composition phase does not take up more layers as models scale up in size. Larger models develop the necessary rich, abstract representations in the same number of layers as smaller models. Thus, as LLMs increase in size, the later layers of the model may contain representations that are increasingly divergent from the more general linguistic processing captured in brain activity. It is also possible that the later layers of larger models are overall underutilized and may not significantly contribute to benchmark performances during inference (Csordás et al., 2025; Fan et al., 2024; Gromov et al., 2024)."

> *Lastly, I would have wanted to see a similar analysis here done for audio encoding models using Whisper or WavLM as this is the modality where you might see real differences between ECoG and other slower scanning approaches. Again, I do not see this omission as a fundamental issue, but it does seem like the sort of analysis for which the higher temporal resolution of ECoG might grant some deeper insight.*

We appreciate this suggestion. In a separate project, we focused on multimodal audio-to-speech-to-language large language models (LLMs), building encoding models using Whisper embeddings (from both the encoder and decoder stacks) to predict electrocorticographic (ECoG) signals during naturalistic conversations (Goldstein, Wang, et al., 2025). The higher temporal resolution of ECoG enables us to trace the temporal flow of information from the superior temporal gyrus (STG) and somatomotor areas (SM) to the inferior frontal gyrus (IFG) during speech comprehension. Conversely, during speech production, encoding in IFG peaked significantly earlier than in the STG and SM. We agree that scaling encoding models using multimodal approaches and our ECoG conversation datasets could yield valuable insights, and we look forward to exploring this in future work. However, we feel that the added complexity of multimodal encoding models falls beyond the scope of this paper.

We have modified the following text to our Discussion section:

"Since we exclusively employ textual LLMs, which lack inherent temporal information due to their discrete token-based nature, future studies utilizing multimodal LLMs integrating continuous audio or video streams, like Whisper or WavLM may better unravel the relationship between model size and temporal dynamic representations in LLMs (Goldstein, Wang, et al., 2025; Millet et al., 2023; Vaidya et al., 2022)."

> ***Reviewer #2 (Public review):***
>
> *Summary:*
>
> *This paper investigates whether large language models (LLMs) of increasing size more accurately align with brain activity during naturalistic language comprehension. The authors extracted word embeddings from LLMs for each word in a 30-minute story and regressed them against electrocorticography (ECoG) activity time-locked to each word as participants listened to the story. The findings reveal that larger LLMs more effectively predict ECoG activity, reflecting the scaling laws observed in other natural language processing tasks.*
>
> *Strengths:*
>
> *(1) The study compared model activity with ECoG recordings, which offer much better temporal resolution than other neuroimaging methods, allowing for the examination of model encoding performance across various lags relative to word onset.*
>
> *(2) The range of LLMs tested is comprehensive, spanning from 82 million to 70 billion parameters. This serves as a valuable reference for researchers selecting LLMs for brain encoding and decoding studies.*

> *(3) The regression methods used are well-established in prior research, and the results demonstrate a convincing scaling law for the brain encoding ability of LLMs. The consistency of these results after PCA dimensionality reduction further supports the claim.*

We thank the reviewer for their thoughtful and positive feedback.

> *Weaknesses:*
>
> *(1) Some claims of the paper are less convincing. The authors suggested that "scaling could be a property that the human brain, similar to LLMs, can utilize to enhance performance", however, many other animals have brains with more neurons than the human brain, making it unlikely that simple scaling alone leads to better language performance.*

We thank the reviewer for this insightful comment. We agree that simply having more neurons does not automatically confer more complex or human-like cognitive or linguistic capabilities. This suggestion deserves a more nuanced treatment than we had included in the original manuscript.

Research in comparative neuroscience has argued that human cognitive abilities emerge from scaling up the primate brain (Herculano-Houzel, 2012). However, the critical aspect is not merely the number of neurons, but how these neurons contribute to computational power within a specific evolutionary and cultural context. The uniqueness of human cognition has been argued to result from a global adaptation for increased information processing capacity (Cantlon & Piantadosi, 2024). Moreover, the language network in humans is likely grounded in the evolution of particular structural networks in the primate brain (Friederici & Becker, 2025). This suggests that the way brain regions are connected and the expansion of certain pathways are critical, not just the overall scale. Furthermore, the specialized structure must be tuned by its learning environment and training data. For example, both humans and LLMs learn from language data generated by other humans, which reflects world knowledge that has accumulated over many generations.

We have modified the following text in the Introduction:

“Research in comparative neuroscience has suggested that uniquely human cognitive abilities emerge from scaling up the primate brain (Herculano-Houzel, 2012).”

We also added a caveat to the Discussion on this point:

“As in the human brain, while scaling alone may yield emergent cognitive abilities (Cantlon & Piantadosi, 2024; Herculano-Houzel, 2012), specialized architectural features likely also play a critical role (Friederici & Becker, 2025).”

> *Additionally, the authors claim that their results show 'larger models better predict the structure of natural language.' However, it remains unclear to what extent the embeddings of LLMs capture the "structure" of language better than the lexical semantics of language.*

We appreciate the reviewer's point about how well LLM embeddings capture the "structure" of language versus just lexical semantics. It's true that distinguishing these aspects is complex. From our perspective, a model's ability to predict/produce natural language entails that the model captures various levels of linguistic structure, including morphology, syntax, semantics, and contextual dependencies. We use "structure" inclusively in this sense. A model cannot achieve high predictive accuracy without representing, to some extent, all of these structural elements (Linzen & Baroni, 2021; Manning et al., 2020; Pavlick, 2022). There is a very active field of research into understanding exactly *how* these models represent these

different structures of language (Ameisen et al., 2025; Chemla et al., 2024; Elhage et al., 2021, 2022; Hewitt & Manning, 2019). Our results confirm the core trend that larger models tend to better reproduce the various structures of language (i.e., yield lower perplexity; Fig. 2A).

In previous work, we have shown that LLM embeddings better predict neural activity during natural language processing than lexical embeddings (e.g., GloVe) that do not contain other elements of linguistic structure (Goldstein et al., 2022; Kumar et al., 2024; Zada et al., 2024). In response to the following comment, we also compare LLMs to simpler models capturing specific speech and language features (see next comment). To clarify our intended use of the word "structure", we've added a brief explanation in the Methods section:

"In this study, we use the term "structure" to refer to a variety of linguistic patterns (e.g., morphology, syntax, semantics, context) that LLMs encode in order to better predict natural language."

> *(2) The study lacks control LLMs with randomly initialized weights and control regressors, such as word frequency and phonetic features of speech, making it unclear what the baseline is for the model-brain correlation.*

We've added several supplementary analyses to the revised manuscript to address these concerns. To establish a baseline, we extracted embeddings from each layer of the SMALL model with randomly initialized weights and constructed encoding models. The encoding performance is significantly higher for pretrained SMALL than for untrained SMALL for every layer (Fig. S4). For the untrained model, the performance is the highest for the 0th layer and decreases in subsequent layers. This is because at the 0th layer, every instance of the same word receives an identical, albeit random, embedding (See Supplementary Figure 4).

We also compared the encoding performance of LLMs with more classical speech/language features and static GloVe embeddings (Goldstein, Wang, et al., 2025; Kumar et al., 2024). First, we extracted features capturing lower-level speech features. Using the stimulus transcript as input, we created one-hot vectors for phonetic and articulatory features. Phoneme classes (39 total classes) were obtained from the Carnegie Mellon Pronouncing Dictionary (The CMU Pronouncing Dictionary, n.d.). We further classified the phonemes based on their place of articulation (9 classes), manner of articulation (9 classes), and voiced or voiceless status (3 classes), according to the general American English consonants of the International Phonetic Alphabet. Given that each word consists of multiple phonemes, we averaged the one-hot vectors for all phonetic and articulatory features for each word.

Second, we extracted linguistic features using spaCy (Honnibal et al., 2020), including part of speech (17 classes), tag (50 classes), function or content word (3 classes), dependency (45 classes), whether the word is an alpha character (binary), and whether the word is a stop word (binary). We also extracted prefix (30 classes) and suffix (44 classes) information using the Cambridge Dictionary. We constructed one-hot vectors for each multi-class feature and one-dimensional vectors for each binary feature.

Third, for each word, we obtained word frequency from the Google Web Trillion Word Corpus (Brants & Franz, 2006) and from our own dataset.

Fourth, we generated static word embeddings of dimension 50 using GloVe (Pennington et al., 2014).

We then built encoding models in the same way as the contextual embeddings for each of the three categories of speech features, all speech features concatenated, and the GloVe embeddings. To control for the different dimensions of the embeddings, we also standardized all embeddings to the same size (50 dimensions) using principal component analysis (PCA) and trained linear encoding models using ordinary least-squares (OLS) regression. For both

ridge and OLS encoding, our contextual embeddings from LLMs showed significantly better performance than the classic speech features and GloVe embeddings.

We have added the following text to our manuscript and updated our Figures S4, S5, Table S1, and the methods section:

"To establish a general baseline for encoding performance, we built encoding models using embeddings from the SMALL model with randomly initialized weights. The trained SMALL model exhibits significantly higher encoding performance across all layers compared to the untrained SMALL model (Fig. S4). We also assessed the encoding performance of contextual embeddings from LLMs against classic speech features and static GloVe embeddings (Table S1). The SMALL and XL embeddings achieved markedly higher encoding correlations than the speech features and GloVe embeddings (Fig. S5)."

> *(3) The finding that peak encoding performance tends to occur in relatively earlier layers in larger models is somewhat surprising and requires further explanation. Since more layers mean more parameters, if the later layers diverge from language processing in the brain, it raises the question of what aspects of the larger models make them more brain-like.*

We thank the reviewer for this insightful comment; this point was also highlighted by Reviewer 1. We agree that this result is somewhat surprising, and we aim to provide a more detailed explanation in the revised manuscript. The general inverted U-shaped trend of encoding performance across layers has been a frequently observed phenomenon in studies comparing LLM representations to brain activity (Goldstein, Ham, et al., 2025; Schrimpf et al., 2021). A potential explanation is the existence of a "two-phase abstraction process" within LLMs (Cheng & Antonello, 2024; Csordás et al., 2025). In the initial layers, models begin by processing relatively low-level input features. As layers get deeper, representations become increasingly abstract and richly contextualized in semantic features relevant for understanding language. These intermediate layers often show the highest correlation with brain activity in language areas, presumably because they capture complex semantic and contextual information in a way that generalizes well across a variety of tasks (including prediction of human neural activity) (Antonello & Huth, 2024). Subsequently, a prediction phase happens in the later layers, where the representations become more specialized for the LLM's specific training objective (e.g., next-word prediction). This specialization can effectively constrict the more generalized feature representations, making these layers less optimal for predicting brain activity. These observations suggest that it is primarily the abstractive, contextual features developed in the intermediate layers of LLMs that drive their alignment with brain activity. As models become more potent at prediction, their most predictive layers (for the LLM's natural language task) and their most generalizable layers (for brain activity) can diverge.

A key finding in our study is that the initial processing phase does not scale and take up more layers as models scale up in size and layers. Larger models develop the necessary rich, abstract representations in the same number of layers as smaller models.

Consequently, the prediction phase may begin relatively earlier in these larger models, and the later layers could develop highly specialized representations that are increasingly divergent from the more general linguistic processing captured in brain activity. For example, these layers may specialize in capturing very specific patterns of language (thus lowering their perplexity) that do not actually occur often or at all in our naturalistic dataset. It is also possible that the later layers of larger models are overall underutilized and do not contribute as much to linguistic processing and next-word prediction (Csordás et al., 2025).

We have added the following text to our Discussion section:

“The inverted U-shaped trend of encoding performance commonly found in previous research is likely due to a "two-phase abstraction process" within LLMs (Cheng & Antonello, 2024; Csordás et al., 2025). In the early and intermediate layers of the model, a composition phase occurs, where low-level input features become increasingly abstract and contextualized. The intermediate layers of the model show the highest correlation with brain activity, presumably because they capture complex semantic and contextual information in a way that generalizes well across a variety of tasks (including prediction of human neural activity) (Antonello & Huth, 2024). Subsequently, a prediction phase happens in the later layers of the model, where the representations become more specialized for the LLM's specific training objective (e.g., next-word prediction). This specialization can effectively constrict the more generalized feature representations, making these layers less optimal for predicting brain activity. Our results indicate that the initial composition phase does not take up more layers as models scale up in size. Larger models develop the necessary rich, abstract representations in the same number of layers as smaller models. Thus, as LLMs increase in size, the later layers of the model may contain representations that are increasingly divergent from the more general linguistic processing captured in brain activity. It is also possible that the later layers of larger models are overall underutilized and may not significantly contribute to benchmark performances during inference (Csordás et al., 2025; Fan et al., 2024; Gromov et al., 2024).”

> ***Reviewer #3 (Public review):***
>
> *This manuscript studies the connection between neural activity collected through electrocorticography and hidden vector representations from autoregressive language models, with the specific aim of studying the influence of language model size on this connection. Neural activity was measured from subjects who listened to a segment from a podcast, and the representations from language models were calculated using the written transcription as the input text. The ability of vector representations to predict neural activity was evaluated using 10-fold cross-validation with ridge regression models.*
>
> *The main results are that (as well summarized in section headings):*
>
> *(1) Larger models predict neural activity better.*
>
> *(2) The ability of language model representations to predict neural activity differs across electrodes and brain regions.*
>
> *(3) The layer that best predicts neural activity differs according to model size, with the "SMALL" model showing a correspondence between layer number and the language processing hierarchy.*
>
> *(4) There seems to be a similar relationship between the time lag and the ability of language model representations to predict neural activity across models.*
>
> *Strengths:*
>
> *(1) The experimental and modeling protocols generally seem solid, which yielded results that answer the authors' primary research question.*
>
> *(2) Electrocorticography data is especially hard to collect, so these results make a nice addition to recent functional magnetic resonance imaging studies.*

We thank the reviewer for their thoughtful and positive feedback.

> *Weaknesses:*

> *(1) The interpretation of some results seems unjustified, although this may just be a presentational issue.*
>
> *(a) Figure 2B: The authors interpret the results as "a plateau in the maximal encoding performance," when some readers might interpret this rather as a decline after 13 billion parameters. Can this be further supported by a significance test like that shown in Figure 4B?*

We agree that this could be a subjective interpretation, so we conducted an additional analysis. We performed paired two-sided *t*-tests between best layer encoding performances averaged across electrodes (df = 159 electrodes), comparing all models with larger models. We found that after 13 billion parameters, only the encoding performance for OPT-66B, the largest model in the OPT family, is significantly worse than the encoding performance of some other smaller models, supporting the claim that the maximal encoding performance declines after 13 billion parameters. However, we did not find conclusive statistical evidence of a decline in encoding performance for other model families.

We have added the statistical results as Supplementary Figure 1.

We have also modified the following text in the manuscript:

“We also observed a plateau in the maximal encoding performance, occurring around 7 billion parameters (Fig. 2B), with a decline in performance for the OPT-66B model (Fig. S1).”

> *(b) Figure S1A: It looks like the drop in PCA max correlation is larger for larger models, which may suggest to some readers that the same trend observed for ridge max correlation may not hold, contra the authors' claim that all results replicate. Why not include a similar figure as Figure 2B as part of Figure S1?*

PCA is an unsupervised dimensionality reduction technique and may discard model features with small eigenvalues that nonetheless contribute to encoding performance. Ridge regression, a supervised method, can capitalize on these features. We suspect that this is why there appears to be a larger drop in model performance for larger models with PCA than with ridge regression. We replicated the logarithmic relationship between model size and encoding performance using PCA and ordinary least-squares (OLS) regression encoding models. We have updated Supplementary Figure 2.

> *(2) Discussion of what might be driving the main result about the influence of model size appears to be missing (cf. the authors aim to provide an explanation of what seems to drive the influence of the layer location in Paragraph 3 of the Discussion section). What explanations have been proposed in the previous functional magnetic resonance imaging studies? Do those explanations also hold in the context of this study?*

We suspect that the increased expressivity of larger models - that is, their improved sensitivity to nuanced structure in natural language - yields improved alignment to brain activity (given large enough samples of brain activity) (Antonello et al., 2023). This effect persists even when dimensionality is tightly controlled in our PCA-based analysis, indicating that the improved alignment with the brain is not a modeling artifact of dimensionality alone, but results from the structural representations learned by these larger models.

We have added the following text to our Discussion section:

“We suspect that the improved alignment with brain activity in larger models is driven by their increased expressivity and sensitivity to nuanced linguistic structure present in large-scale naturalistic datasets (Antonello et al., 2023).”

> *(3) The GloVe-based selection of language-sensitive electrodes (at least to me) isn't explained/motivated clearly enough (I think a more detailed explanation should be included in the Materials and Methods section). If the electrodes are selected based on GloVe embeddings, then isn't the main experiment just showing that representations from larger language models track more closely with GloVe embeddings? What justifies this methodology?*

We selected electrodes based on previously established methods (Goldstein et al., 2022). Our use of GloVe embeddings for electrode selection does not imply that larger language model representations are simply more closely aligned with GloVe embeddings. On the contrary, contextual embeddings from LLMs, which incorporate the word's previous context, consistently outperform static embeddings like GloVe or word2vec (Fig. S3). Selecting electrodes using LLM embeddings would likely result in a slightly different, potentially larger set of electrodes (Goldstein et al., 2022), but would be more circular (Kriegeskorte et al., 2009). The GloVe-based electrode selection represents a more conservative approach by identifying words encoding linguistic content without biasing the selection directly toward any LLMs.

We have added the following text to our Method section:

"We used GloVe embeddings for electrode selection to avoid biasing our main results toward a particular LLM."

> *(4) (Minor weakness) The main experiments are largely replications of previous functional magnetic resonance imaging studies, with the exception of the one lag-based analysis. Is there anything else that the electrocorticography data can reveal that functional magnetic resonance imaging data can't?*

We thank the reviewer for this thoughtful question. While we agree that a key contribution of our work corroborates previous fMRI findings, we would argue that using ECoG is not merely a replication but a crucial validation and extension of that work. It is important to validate these effects across distinct measurement modalities. In our work, we further observed a novel trend where the peak encoding performance tends to occur in relatively earlier layers for larger models. This is supported by recent studies suggesting that later layers of large LLMs may not significantly contribute to benchmark performance (Csordás et al., 2025). While scaling has been an effective method to improve LLM performance, including in encoding models, future research should explore the potential underutilization of the later layers as models scale.

Furthermore, ECoG data offers temporal resolution on the order of milliseconds, far superior to fMRI's. Although we did not observe a relationship between model size and temporal lags in this study, future work should investigate the temporal dynamics of encoding that are accessible with ECoG (Goldstein, Ham, et al., 2025; Goldstein, Wang, et al., 2025).

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations for the authors):***
>
> *Thank you to the authors for the fun and personally useful read.*
>
> *I see in Supplementary Figure 1 the authors show a comparison of the performance between OLS vs. Ridge regression. Is the OLS model the only one that is working over PC features, or are both models using PC features? The current text is a bit unclear. My current understanding is that the comparison is between (OLS + PCA) and (Ridge with no PCA), but I am not sure.*

The OLS model is the only one that works over PC features, following previous methods (Goldstein et al., 2022).

We have added the following text to our Results and Methods section for clarity:

"To control for the different embedding dimensionality across models, we standardized all embeddings to the same size using principal component analysis (PCA) and trained linear encoding models using ordinary least-squares (OLS) regression, replicating the logarithmic relationship but with significantly lower encoding performance overall (Fig. S2). The PC features are used by the OLS models only."

> *Clarification in the text would be appropriate. If this is the correct understanding, the authors should note in the main text that the ridge approach is more effective than the PCA approach, which is still the dominant approach to building linear encoding models in the field for some unjustifiable reason.*

We thank the reviewer for pointing out the confusion. We have updated Supplementary Figure 2.

> *How were the alpha values for ridge regression determined? Do you use the same ridge parameter for all electrodes or fit a different parameter for each electrode? This is not mentioned anywhere.*

The alpha values are determined by cross-validation using the "RidgeCV" method from the "himalaya" package (Dupré la Tour et al., 2022). Specifically, we perform a grid search over cross-validation folds in the training data to find the best-performing alpha. The alpha parameter is specific to each ridge regression model, meaning each fold, lag, and electrode has a different alpha parameter.

We have added the following text to our manuscript:

"For each ridge regression model (for each fold, lag, and electrode), the alpha parameter is determined by cross-validation using the "RidgeCV" method from the "himalaya" package (Dupré la Tour et al., 2022)."

> *It's not entirely clear to me how the authors handle tokens that do not terminate in words (such as the "there" + "'s" example in the text). My current reading of the text is that authors essentially ignore these half-word embeddings, doing one forward pass per word, rather than per token, but the current text is somewhat ambiguous.*

If a word is tokenized into several tokens, like "there" and "'s", we average the token embeddings to get a word embedding.

We have added the following text to our Method section:

"To facilitate a fair comparison of the encoding effect across different models, we aligned all tokens in the story across all models. We averaged the token embeddings if a word is split into multiple tokens, resulting in one embedding per word for each model."

> *The authors describe the scaling relationship they find as a "log-linear" relationship. I believe this is a misnomer derived from the original paper describing this relationship in fMRI as log-linear (Antonello et al.) The correct term is simply "logarithmic", and for what it's worth, the authors of the original fMRI work have made this correction as well.*

Thank you! We have made this correction.

> *Is the data publicly available? If not, there should be some basic justification as to why (consent reasons, etc.).*

We have recently made the data publicly available (Zada et al., 2025). We have also provided tutorials for preprocessing the data and training encoding models: *https://hassonlab.github.io/podcast-ecog-tutorials*. For this specific project, the analysis code is available at *https://github.com/hassonlab/247-pickling/tree/scaling-paper-0* and *https://github.com/hassonlab/247-encoding/tree/scaling-paper-1*.

> *The authors assert that ECoG has "superior spatiotemporal resolution". While this is unquestionably true for temporal resolution, the story is a bit more complicated for spatial resolution, where ECoG has far less cortical coverage than fMRI. Perhaps this sentence should be revised.*

Thank you for pointing out the typo! We have changed it to “superior temporal resolution”.

> *Minor Points:*
>
> *The bolded title of Figure 3 probably shouldn't be bolded, as this is just actually the title of Figure 3A.*

Fixed.

> *Figure 4d is has a typo: "Best Encoidng Layer".*

Fixed.

> ***Reviewer #2 (Recommendations for the authors):***
>
> *The authors could consider adding control regressors such as word rate, word frequency, phonetic features, and syntactic features like node counts, as well as control LLMs of comparable size to serve as baselines. The authors could also include correlation analyses of the embeddings from different layers of the same LLM to further illustrate how distinct the layers are within the models.*

We have added untrained LLM embeddings as a baseline and included a comparison of encoding models between LLM contextual embeddings and classical speech features. We have also performed some preliminary correlation analyses of embeddings. In some models, we found evidence of the “two-phase abstraction process” (Cheng & Antonello, 2024). However, the result is inconclusive across different LLM families. Since each LLM layer accesses and modifies the residual stream (Elhage et al., 2021), the embeddings across layers are inherently correlated. Future work could instead explore the isolated transformations within each layer to illustrate the distinct information across layers (Kumar et al., 2024).

> *The analysis codes and data should be made available.*

We have recently made the data publicly available (Zada et al., 2025). We have also provided tutorials for preprocessing the data and training encoding models: *https://hassonlab.github.io/podcast-ecog-tutorials*. For this specific project, the analysis code is available at *https://github.com/hassonlab/247-pickling/tree/scaling-paper-0* and *https://github.com/hassonlab/247-encoding/tree/scaling-paper-1*.

> ***Reviewer #3 (Recommendations for the authors):***
>
> *Most of my concrete recommendations are in the public review. Below are some additional minor ones:*

> *(1) Introduction: "Remarkably, these models learn from much the same shared space as humans: from real-world language generated by humans."*
>
> *I think this is an extremely strong claim due to e.g. the different nature of child-directed speech vs. written text corpora, the lack of multimodality and grounding in language models, etc. I might suggest re-wording this sentence or removing it entirely.*

We thank the reviewer for their suggestion! We have removed the sentence from the manuscript.

> *(2) Introduction: "EleutherAI, n.d." reference for GPT-Neo*
>
> *GPT-NeoX-20B has an associated paper, which the authors might cite instead: https://aclanthology.org/2022.bigscience-1.9*

Thank you! We have added the reference for GPT-NeoX-20B (Black et al., 2022).

> *(3) Figure 4D: Encoidng -> Encoding*

Fixed.

> *(4) Materials and Methods, Contextual embeddings: "except for GPT-Neox-20b, which assigns additional tokens to whitespace characters."*
>
> *What do the authors mean by "additional tokens to whitespace characters?" The tokenizer for GPT-NeoX-20B works in much the same way as that of GPT-Neo, just with a different vocabulary set.*
>
> *>>> t1 = AutoTokenizer.from_pretrained("EleutherAI/gpt-neo-125M")*
>
> *>>> t2 = AutoTokenizer.from_pretrained("EleutherAI/gpt-neox-20b")*
>
> *>>> t1.convert_ids_to_tokens(t1("The quick brown fox jumps over the lazy dog.").input_ids) ['The', 'Ġquick', 'Ġbrown', 'Ġfox', 'Ġjumps', 'Ġover', 'Ġthe', 'Ġlazy', 'Ġdog', '.']*
>
> *>>> t2.convert_ids_to_tokens(t2("The quick brown fox jumps over the lazy dog.").input_ids)*
>
> *['The', 'Ġquick', 'Ġbrown', 'Ġfox', 'Ġjumps', 'Ġover', 'Ġthe', 'Ġlazy', 'Ġdog', '.']*
>
> *If the authors are referring to Ġ as the "additional token to whitespace characters," then these are in all other tokenizers as well (not only that for GPT-Neo, but also those for GPT-2 and OPT).*

We agree that “additional tokens to whitespace characters” is an oversimplification. The GPT-Neo model family, which includes the 125M, 1.3B, and 2.7B models, utilizes the same Byte Pair Encoding (BPE) tokenizer as GPT-2. This common tokenizer has a vocabulary size of 50,257 tokens, providing compatibility and seamless integration across the models.

The GPT-NeoX-20B model introduces a modified tokenizer to address limitations observed in the GPT-2 tokenizer (Black et al., 2022). As detailed in Section 3.2, this new tokenizer incorporates a few key improvements:

(1) New BPE tokenizer: A more general-purpose BPE tokenizer was trained using the Pile dataset.

(2) Space Delimitation: Unlike the GPT-2 tokenizer, which treats tokenization at the start of a string as a non-space-delimited token, the GPT-NeoX-20B tokenizer applies consistent space

delimitation regardless. This change resolves inconsistencies related to the presence of prefix spaces in the tokenization input.

(3) Whitespace Handling: The tokenizer includes tokens for repeated space characters (up to 24 consecutive spaces), enhancing efficiency in tokenizing text with substantial whitespace, such as program source code or LaTeX documents.

These modifications result in the GPT-NeoX-20B tokenizer representing the Pile validation set with approximately 10% fewer tokens than the GPT-2 tokenizer. This efficiency gain is particularly beneficial for processing texts with extensive whitespace.

In our analysis, we extracted embeddings by setting àdd_prefix_space = True ̀to all tokenizers, so space delimitation does not result in tokenizer differences. We highlight here examples of the other two tokenizer differences using the Huggingface AutoTokenizer:̀

>>> t1 = AutoTokenizer.from_pretrained("EleutherAI/gpt-neo-125M")

>>> t2 = AutoTokenizer.from_pretrained("EleutherAI/gpt-neox-20b")

>>> t1.convert_ids_to_tokens(t1("The Downing Street.").input_ids) ['The', 'ĠDowning', 'ĠStreet']

>>> t2.convert_ids_to_tokens(t2("The Downing Street.").input_ids)

['The', 'ĠDown', 'ing', 'ĠStreet']

>>> t1.convert_ids_to_tokens(t1("Hello !").input_ids)

['Hello', 'Ġ', 'Ġ', 'Ġ', 'Ġ', 'Ġ', 'Ġ', 'Ġ!']

>>> t2.convert_ids_to_tokens(t2("Hello !").input_ids)

['Hello', ' ', '!']

More examples showing the differences between the GPT-2 tokenizer and the GPT-NeoX-20B tokenizer can be found in Appendix F: Tokenizer Analysis (Black et al., 2022).

We have added the following text to our manuscript for simplicity:

“All models within the same model family adhere to the same tokenizer convention, except for GPT-Neox-20B, which utilizes a different tokenizer (Black et al., 2022).”

References

Ameisen, E., Lindsey, J., Pearce, A., Gurnee, W., Turner, N. L., Chen, B., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., … Batson, J. (2025). Circuit Tracing: Revealing Computational Graphs in Language Models. Transformer Circuits Thread. https://transformer-circuits.pub/2025/attribution-graphs/methods.html

Antonello, R., & Huth, A. (2024). Predictive coding or just feature discovery? An alternative account of why language models fit brain data. Neurobiology of Language (Cambridge, Mass.), 5(1), 64–79.

Antonello, R., Vaidya, A., & Huth, A. G. (2023). Scaling laws for language encoding models in fMRI. NeurIPS 2023. *https://doi.org/10.48550/ARXIV.2305.11863*

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U. S., Purohit, S., Reynolds, L., Tow, J., Wang, B., & Weinbach, S. (2022). GPT-NeoX-20B: An Open-Source Autoregressive Language Model. Proceedings of BigScience Episode #5 -- Workshop on Challenges & Perspectives in Creating

Large Language Models. Proceedings of BigScience Episode #5 -- Workshop on Challenges & Perspectives in Creating Large Language Models, virtual+Dublin. *https://doi.org/10.18653/v1/2022.bigscience-1.9*

Brants, T., & Franz, A. (2006). Web 1T 5-gram Version 1 [Dataset]. Linguistic Data Consortium. *https://doi.org/10.35111/CQPA-A498*

Cantlon, J. F., & Piantadosi, S. T. (2024). Uniquely human intelligence arose from expanded information capacity. Nature Reviews Psychology, 3(4), 275–293.

Chemla, E., D'Ascoli, S., Diego-Simón, P., King, J.-R., & Lakretz, Y. (2024). A Polar coordinate system represents syntax in large language models. Advances in Neural Information Processing Systems 37, 105375–105396.

Cheng, E., & Antonello, R. J. (2024). Evidence from fMRI supports a two-phase abstraction process in language models. In arXiv [cs.CL]. arXiv. *http://arxiv.org/abs/2409.05771*

Csordás, R., Manning, C. D., & Potts, C. (2025). Do language models use their depth efficiently? In arXiv [cs.LG]. *https://doi.org/10.48550/ARXIV.2505.13898*

Dupré la Tour, T., Eickenberg, M., Nunez-Elizalde, A. O., & Gallant, J. L. (2022). Feature-space selection with banded ridge regression. In bioRxiv. *https://doi.org/10.1101/2022.05.05.490831*

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). Toy Models of Superposition. Transformer Circuits Thread.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., … Olah, C. (2021). A Mathematical Framework for Transformer Circuits. Transformer Circuits Thread.

Fan, S., Jiang, X., Li, X., Meng, X., Han, P., Shang, S., Sun, A., Wang, Y., & Wang, Z. (2024). Not all Layers of LLMs are Necessary during Inference. In arXiv [cs.CL]. arXiv. *http://arxiv.org/abs/2403.02181*

Friederici, A. D., & Becker, Y. (2025). The core language network separated from other networks during primate evolution. Nature Reviews. Neuroscience, 26(2), 131–132.

Goldstein, A., Ham, E., Schain, M., Nastase, S. A., Aubrey, B., Zada, Z., Grinstein-Dabush, A., Gazula, H., Feder, A., Doyle, W., Devore, S., Dugan, P., Friedman, D., Brenner, M., Hassidim, A., Matias, Y., Devinsky, O., Siegelman, N., Flinker, A., … Hasson, U. (2025). Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. Nature Communications, 16(1), 10529.

Goldstein, A., Wang, H., Niekerken, L., Schain, M., Zada, Z., Aubrey, B., Sheffer, T., Nastase, S. A., Gazula, H., Singh, A., Rao, A., Choe, G., Kim, C., Doyle, W., Friedman, D., Devore, S., Dugan, P., Hassidim, A., Brenner, M., … Hasson, U. (2025). A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. Nature Human Behaviour. *https://doi.org/10.1038/s41562-025-02105-9*

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.

Gromov, A., Tirumala, K., Shapourian, H., Glorioso, P., & Roberts, D. A. (2024). The Unreasonable Ineffectiveness of the Deeper Layers. In arXiv [cs.CL]. arXiv. *http://arxiv.org/abs/2403.17887*

Herculano-Houzel, S. (2012). The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. Proceedings of the National Academy of Sciences of the United States of America, 109 Suppl 1(supplement_1), 10661–10668.

Hewitt, J., & Manning, C. D. (2019). A Structural Probe for Finding Syntax in Word Representations. In J. Burstein, C. Doran, & T. Solorio (Eds.), Proceedings of the 2019 Conference of the North (pp. 4129–4138). Association for Computational Linguistics.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. Nature Neuroscience, 12(5), 535–540.

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (2024). Shared functional specialization in transformer-based language models and the human brain. Nature Communications, 15(1), 5523.

Linzen, T., & Baroni, M. (2021). Syntactic Structure from Deep Learning. Annual Review of Linguistics, 7(1), 195–212.

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.

Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., & King, J.-R. (2023). Toward a realistic model of speech processing in the brain with self-supervised learning. NeurIPS 2022. *https://doi.org/10.48550/ARXIV.2206.01685*

Pavlick, E. (2022). Semantic structure in deep learning. Annual Review of Linguistics, 8(1), 447–471.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar. *https://doi.org/10.3115/v1/d14-1162*

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.

The CMU Pronouncing Dictionary. (n.d.). Retrieved May 27, 2025, from *http://www.speech.cs.cmu.edu/cgi-bin/cmudict*

Vaidya, A. R., Jain, S., & Huth, A. G. (2022). Self-supervised models of audio effectively explain human cortical responses to speech. ICML 2022. *https://doi.org/10.48550/ARXIV.2205.14252*

Zada, Z., Goldstein, A., Michelmann, S., Simony, E., Price, A., Hasenfratz, L., Barham, E., Zadbood, A., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Devore, S., Flinker, A., Devinsky, O., Nastase, S. A., & Hasson, U. (2024). A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. Neuron, S0896627324004604.
Zada, Z., Nastase, S. A., Aubrey, B., Jalon, I., Michelmann, S., Wang, H., Hasenfratz, L., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Devore, S., Flinker, A., Devinsky, O., Goldstein, A., &

Hasson, U. (2025). The “Podcast” ECoG dataset for modeling neural activity during natural language comprehension. Scientific Data, 12(1), 1135.

https://doi.org/10.7554/eLife.101204.2.sa0