nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-022-01026-4

Check for updates

OPEN

Shared computational principles for language processing in humans and deep language models

Ariel Goldstein[1,2]✉, Zaid Zada[1,8], Eliav Buchnik[2,8], Mariano Schain[2,8], Amy Price[1,8], Bobbi Aubrey[1,3,8], Samuel A. Nastase[1,8], Amir Feder[2,8], Dotan Emanuel[2,8], Alon Cohen[2,8], Aren Jansen[2,8], Harshvardhan Gazula[1], Gina Choe[1,3], Aditi Rao[1,3], Catherine Kim[1,3], Colton Casto[1], Lora Fanda[3], Werner Doyle[3], Daniel Friedman[3], Patricia Dugan[3], Lucia Melloni[4], Roi Reichart[5], Sasha Devore[3], Adeen Flinker[3], Liat Hasenfratz[1], Omer Levy[6], Avinatan Hassidim[2], Michael Brenner[2,7], Yossi Matias[2], Kenneth A. Norman[1], Orrin Devinsky[3] and Uri Hasson[1,2]

Departing from traditional linguistic models, advances in deep learning have resulted in a new type of predictive (autoregressive) deep language models (DLMs). Using a self-supervised next-word prediction task, these models generate appropriate linguistic responses in a given context. In the current study, nine participants listened to a 30-min podcast while their brain responses were recorded using electrocorticography (ECoG). We provide empirical evidence that the human brain and autoregressive DLMs share three fundamental computational principles as they process the same natural narrative: (1) both are engaged in continuous next-word prediction before word onset; (2) both match their pre-onset predictions to the incoming word to calculate post-onset surprise; (3) both rely on contextual embeddings to represent words in natural contexts. Together, our findings suggest that autoregressive DLMs provide a new and biologically feasible computational framework for studying the neural basis of language.

The outstanding success of autoregressive (predictive) DLMs is striking from theoretical and practical perspectives because they have emerged from a very different scientific paradigm than traditional psycholinguist models[1]. In traditional psycholinguistic approaches, human language is explained with interpretable models that combine symbolic elements (for example, nouns, verbs, adjectives and adverbs) with rule-based operations[2,3]. In contrast, autoregressive DLMs learn language from real-world textual examples 'in the wild', with minimal or no explicit prior knowledge about language structure. Autoregressive DLMs do not parse words into parts of speech or apply explicit syntactic transformations. Rather, they learn to encode a sequence of words into a numerical vector, termed a contextual embedding, from which the model decodes the next word. After learning, the next-word prediction principle allows the generation of well-formed, novel, context-aware texts[1,4,5].

Autoregressive DLMs have proven to be extremely effective in capturing the structure of language[6–9]. It is unclear, however, if the core computational principles of autoregressive DLMs are related to the way the human brain processes language. Past research has leveraged language models and machine learning to extract semantic representation in the brain[10–18]. But such studies did not view autoregressive DLMs as feasible cognitive models for how the human brain codes language. In contrast, recent theoretical papers argue that there are fundamental connections between DLMs and how the brain processes language[1,19,20].

In agreement with this theoretical perspective, we provide empirical evidence that the human brain processes incoming speech similarly to an autoregressive DLM (Fig. 1). In particular, the human brain and autoregressive DLMs share three computational principles: (1) both are engaged in continuous context-dependent next-word prediction before word onset; (2) both match pre-onset predictions to the incoming word to induce post-onset surprise (that is, prediction-error signals); (3) both represent words using contextual embeddings. The main contribution of this study is the provision of direct evidence for the continuous engagement of the brain in next-word prediction before word onset (computational principle 1). In agreement with recent publications[14,16,21–24], we also provide neural evidence in support of computational principles 2 and 3.

Principle 1: next-word prediction before word onset. The constant engagement in next-word prediction before word onset is the bedrock objective of autoregressive DLMs[6–9,25]. Similarly, the claim that the brain is constantly engaged in predicting the incoming input is fundamental to numerous predictive coding theories[26–30]. However, even after decades of research, behavioral and neural evidence for the brain's propensity to predict upcoming words before they are perceived during natural language processing has remained indirect. On the behavioral level, the ability to predict upcoming words has typically been tested with highly controlled sentence stimuli (that is, the cloze procedure[31–33]). Thus, we still do not know how accurate

[1]Department of Psychology and the Neuroscience Institute, Princeton University, Princeton, NJ, USA. [2]Google Research, Mountain View, CA, USA. [3]New York University Grossman School of Medicine, New York, NY, USA. [4]Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany. [5]Faculty of Industrial Engineering and Management, Technion, Israel Institute of Technology, Haifa, Israel. [6]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel. [7]School of Engineering and Applied Science, Harvard University, Cambridge, MA, USA. [8]These authors contributed equally: Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen.
✉e-mail: ariel.y.goldstein@gmail.com

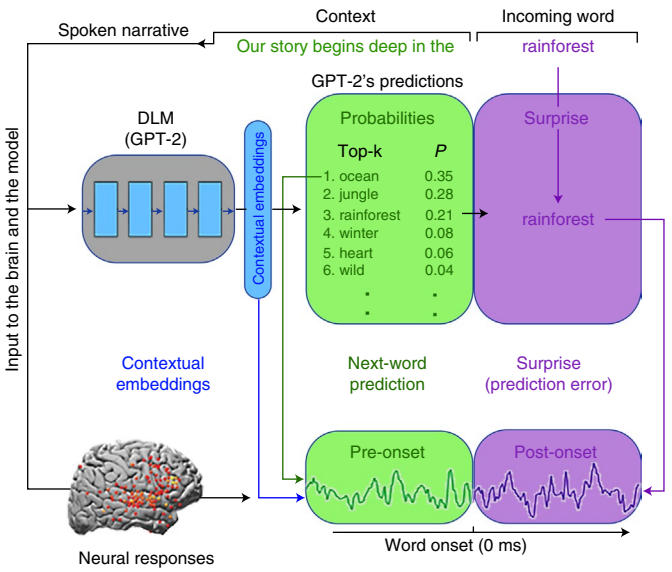


Fig. 1 | Shared computational principles between the brain and autoregressive deep language models in processing natural language. For each sequence of words in the text, GPT-2 generates a contextual embedding (blue), which is used to assign probabilities to the identity of the next word (green box). Next, GPT-2 uses its pre-onset predictions to calculate its surprise level (that is, error signal) when the next word is revealed (purple box). The minimization of the surprise is the mechanism by which GPT-2 is trained to generate well-formed outputs. Each colored box and arrow represents the relationship between a given computational principle of the autoregressive DLM and the neural responses. The green boxes represent that both the brain and autoregressive DLMs are engaged in context-dependent prediction of the upcoming word before word onset. The green arrow indicates that we take the actual predictions from GPT-2 (for example, the top-one prediction 'ocean' for the example sentence 'our story begins deep in the…') to model the neural responses before word onset (Fig. 4). The purple boxes represent that both the brain and autoregressive DLMs are engaged in the assessment of their predictions after word onset. The purple arrow indicates that we take the actual perceived next word ('rainforest' in our example), as well as GPT-2's surprise level for the perceived word (cross-entropy) to model the neural responses after word onset (Figs. 4b and 5b). The blue boxes represent that both the brain and autoregressive DLMs use contextual embeddings to represent words. The blue arrow indicates that we take the contextual embeddings from GPT-2 to model the neural responses (Figs. 6 and 8). We argue here that, although the internal word-processing mechanisms are not the same for the brain and DLMs, they do share three core computational principles: (1) continuous context-dependent next-word prediction before word onset; (2) reliance on the pre-onset prediction to calculate post-word-onset surprise; and finally, (3) context-specific representation of meaning.

listeners' predictions are in open-ended natural contexts. The first section of the paper provides new behavioral evidence that humans can predict forthcoming words in a natural context with remarkable accuracy, and that, given a sufficient context window, next-word predictions in humans and an autoregressive DLM (GPT-2)[8] match. On the neuronal level, we provide new evidence that the brain is spontaneously engaged in next-word prediction before word onset during the processing of natural language. These findings provide the missing evidence that the brain, like autoregressive DLMs, is constantly involved in next-word prediction before word onset as it processes natural language (Fig. 1).

Principle 2: pre-onset predictions are used to calculate post-word-onset surprise. Detecting increased neural activity 400 ms after word onset for unpredictable words, documented across many studies[34,35], has traditionally been used as indirect evidence for pre-onset predictions. Recent development of autoregressive DLMs, like GPT-2, provides a powerful new way to quantify the surprise and confidence levels for each upcoming word in natural language[14,22,23]. Specifically, autoregressive DLMs use the confidence in pre-onset next-word predictions to calculate post-word-onset surprise level (that is, prediction error; Fig. 1)[14,22,23]. Here we map the temporal coupling between confidence level (entropy) in pre-onset prediction and post-word-onset surprise signals (cross-entropy). Our findings provide compelling evidence that, similarly to DLMs, the biological neural error signals after word onset are coupled to pre-onset neural signals associated with next-word predictions.

Principle 3: contextual vectorial representation in the brain. Autoregressive DLMs encode the unique, context-specific meaning of words based on the sequence of prior words. Concurrent findings demonstrate that contextual embeddings derived from GPT-2 provide better modeling of neural responses in multiple brain areas than static (that is, non-contextual) word embeddings[16,17]. Our paper goes beyond these findings by showing that contextual embeddings encapsulate information about past contexts as well as next-word predictions. Finally, we demonstrate that contextual embeddings are better than non-contextual embeddings at predicting word identity from cortical activity (that is, decoding) before (and after) word onset. Taken together, our findings provide compelling evidence for core computational principles of pre-onset prediction, post-onset surprise, and contextual representation, shared by autoregressive DLMs and the human brain. These results support a unified modeling framework for studying the neural basis of language.

Results

Prediction before word onset. *Comparison of next-word prediction behavior in autoregressive deep language models and humans.* We developed a sliding-window behavioral protocol to directly quantify humans' ability to predict every word in a natural context (Fig. 2a,b). Fifty participants proceeded word by word through a 30-min transcribed podcast ('Monkey in the Middle', This American Life podcast[36]) and provided a prediction of each upcoming word. The procedure yielded 50 predictions for each of the story's 5,113 words (Fig. 2c and Methods). We calculated the mean prediction performance for each word in the narrative, which we refer to as a 'predictability score' (Fig. 2d). A predictability score of 100% indicates that all participants correctly guessed the next word, and a predictability score of 0% indicates that no participant predicted the upcoming word. This allowed us to address the following questions: First, how good are humans at next-word prediction? Second, how closely do human predictions align with autoregressive DLM predictions?

Word-by-word behavioral prediction during a natural story. Participants were able to predict many upcoming words in a complex and unfamiliar story. The average human predictability score was 28% (s.e. = 0.5%), in comparison to a predictability score of 6% for blindly guessing the most frequent word in the text ('the'). About 600 words had a predictability score higher than 70%. Interestingly, high predictability was not confined to the last words in a sentence and applied to words from all parts of speech (21.44% nouns, 14.64% verbs, 41.62% function words, 4.35% adjectives and adverbs, and 17.94% other). This suggests that humans are proficient in predicting upcoming words in real-life contexts when asked to do so.

Human and deep language model next-word predictions and probabilities. Next, we compared human and autoregressive DLM predictions of the words of the podcast as a function of prior context. As an instance of an autoregressive DLM, we chose to work with

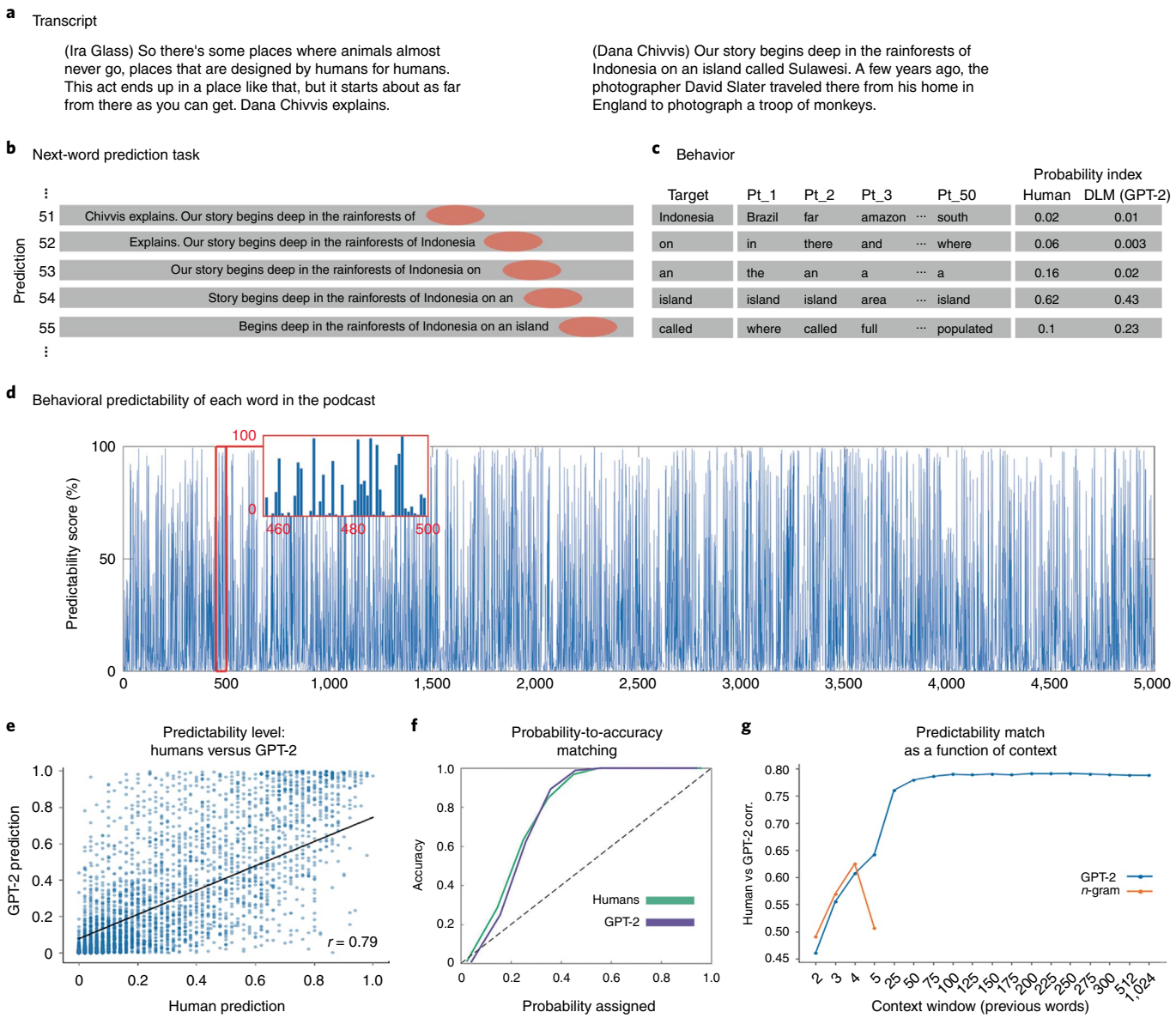


Fig. 2 | Behavioral assessment of the human ability to predict forthcoming words in a natural context. a, The stimulus was transcribed for the behavioral experiment. **b**, A ten-word sliding window was presented in each trial, and participants were asked to type their prediction of the next word. Once entered, the correct word is presented, and the window slides forward by one word. **c**, For each word, we calculated the proportion of participants that predicted the forthcoming word correctly. **d**, Human predictability scores across words. **e**, Human predictability scores versus GPT-2's predictability scores for each upcoming word in the podcast. **f**, Match between assigned probability for humans and GPT-2 and the actual accuracy for their top-one predictions. **g**, Correlation between human predictions and GPT-2 predictions (as reported in **d**) for different context window lengths ranging from 2 to 1,024 preceding tokens (blue). Correlation between human predictions and 2- to 5-gram model predictions (orange).

GPT-2 (ref. [8]). GPT-2 is a pretrained autoregressive language model with state-of-the-art performance on tasks related to reading comprehension, translation, text summarization and question answering. GPT-2 is trained by maximizing the log-probability of a token given its 1,024 past tokens (context, for a full description see ref. [7]). For each word in the transcript, we extracted the most probable next-word prediction as a function of context. For example, GPT-2 assigned a probability of 0.82 to the upcoming word 'monkeys' when it received the preceding words in the story as contextual input: '… So after two days of these near misses, he changed strategies. He put his camera on a tripod and threw down some cookies to try to entice the _____'. Human predictability scores and GPT-2 estimations of predictability were highly correlated (Fig. 2e; $r = 0.79$, $P < 0.001$). In this case, the most probable next-word prediction for both GPT-2 and humans was 'monkeys'. In 49.1% of the cases, the most probable human prediction and the most probable GPT-2 prediction matched (irrespective of accuracy). For baseline comparison, we reported the same agreement measure with human prediction for 2- to 5-gram models in Extended Data Fig. 1 (Methods). Regarding accuracy, GPT-2 and humans jointly correctly and incorrectly predicted 27.6% and 54.7% of the words, respectively. Only 9.2% of the words that humans predicted correctly were not correctly predicted by GPT-2, and only 8.4% of the words correctly predicted by GPT-2 were not correctly predicted by humans (Extended Data Fig. 2).

Finally, we compared the match between the confidence level and the accuracy level of GPT-2 and human predictions. For example, if

the model (or humans) assigned a 20% probability, would it (or they) produce only 20% correct predictions? Both humans and GPT-2 had a remarkably similar confidence-to-accuracy function (Fig. 2f). Specifically, GPT-2 and humans displayed under-confidence in their predictions and were above 95% correct when the probabilities were higher than 40%. These analyses suggest that next-word predictions of GPT-2 and humans are similar in natural contexts.

Prediction as a function of contextual window size. In natural comprehension (for example, listening to or reading a story), predictions for upcoming words are influenced by information accumulated over multiple timescales: from the most recent words to the information gathered over multiple paragraphs[37]. We tested if GPT-2's predictions would improve as a function of the context window as they do in humans. To that end, we varied GPT-2's input window size (from 2 tokens up to 1,024 tokens) and examined the impact of contextual window size on the match with human behavior. The correlation between human and GPT-2 word predictions improved as the contextual window increased (from $r=0.46$, $P<0.001$ at 2-word context to an asymptote of $r=0.79$ at 100-word context; Fig. 2g). For baseline comparison, we also plotted the correlations of 2- to 5-gram models with human predictions (Fig. 2g and Methods).

Neural activity before word onset reflects next-word predictions. The behavioral study indicates that listeners can accurately predict upcoming words in a natural open-ended context when explicitly instructed. Furthermore, it suggests human predictions and autoregressive DLM predictions are matched in natural contexts. Next, we asked whether the human brain, like an autoregressive DLM, is continuously engaged in spontaneous next-word prediction before word onset without explicit instruction. To that end, we used electrocorticography signals from nine participants with epilepsy who volunteered to participate in the study (see Fig. 3a for a map of all electrodes). All participants listened to the same spoken story used in the behavioral experiment. In contrast to the behavioral study, the participants engaged in free listening—with no explicit instructions to predict upcoming words. Comprehension was verified using a post-listening questionnaire. Across participants, we had better coverage in the left hemisphere (1,106 electrodes) than in the right hemisphere (233 electrodes). Thus, we focused on language processing in the left hemisphere, but we also present the encoding results for the right hemisphere in Extended Data Fig. 3. The raw signal was preprocessed to reflect high-frequency broadband (70–200 Hz) power activity (for full preprocessing procedures, see Methods).

Below we provide multiple lines of evidence that the brain, like autoregressive DLMs, is spontaneously engaged in next-word prediction before word onset. The first section focuses solely on the pre-onset prediction of individual words by using static (that is, non-contextual) word embeddings (GloVe[38] and word2vec[39]). In the third section, we investigate how the brain adjusts its responses to individual words as a function of context, by relying on contextual embeddings.

Localizing neural responses to natural speech. We used a linear encoding model and static semantic embeddings (GloVe) to localize electrodes containing reliable responses to single words in the narrative (Fig. 3a,b and Methods). Encoding models learn a mapping to predict brain signals from a representation of the task or stimulus[40]. The model identified 160 electrodes in early auditory areas, motor cortex and language areas in the left hemisphere (see Fig. 3c for left-hemisphere electrodes and Extended Data Fig. 3 for right-hemisphere electrodes).

Encoding neural responses before word onset. In the behavioral experiment (Fig. 2), we demonstrated people's capacity to predict upcoming words in the story. Next, we tested whether the neural signals also contain information about the identity of the predicted words before they are perceived (that is, before word onset). The word-level encoding model (based on GloVe word embeddings) yielded significant correlations with predicted neural responses to upcoming words up to 800 ms before word onset (Fig. 4a; for single electrodes encoding models see Extended Data Fig. 4). The encoding analysis was performed in each electrode with significant encoding for GloVe embeddings ($n=160$), and then averaged across electrodes (see map of electrodes in Fig. 3c). Peak encoding performance was observed 150–200 ms after word onset (lag 0), but the models performed above chance up to 800 ms before word onset. As a baseline for the noise level, we randomly shuffled the GloVe embeddings, assigning a different vector to the occurrence of each word in the podcast. The analysis yielded a flat encoding value around zero (Fig. 4a). The encoding results using GloVe embeddings were replicated using 100-dimensional static embeddings from word2vec (Extended Data Fig. 5). To control for the contextual dependencies between adjacent words in the GloVe embeddings, we demonstrated that the significant encoding before word onset holds even after removing the information of the previous GloVe embedding (Extended Data Fig. 6a). This supports the claim that the brain continuously predicts semantic information about the meaning of upcoming words before they are perceived.

To test whether GloVe-based encoding is affected by the semantic knowledge embedded in the model, we shuffled the word embeddings. Interestingly, when assigning a non-match GloVe embedding (from the story) to each word such that multiple occurrences of the same word received the same (but non-match) GloVe embedding, the encoding decreased (Extended Data Fig. 7). This indicates that the relational linguistic information encoded in GloVe embeddings is also embedded in the neural activity.

Encoding neural responses before word onset. To test if the significant encoding before word onset is driven by contextual dependencies between adjacent words in the GloVe embeddings, we also trained encoding models to predict neural responses using 50-dimensional static arbitrary embeddings, randomly sampled from a uniform [−1, 1] distribution. Arbitrary embeddings effectively removed the contextual information about the statistical relationship between words included in GloVe embeddings (Fig. 4a). Even for arbitrary embeddings, we were able to obtain significant encoding before word onset as to the identity of the upcoming word (for single-electrode encoding models, see Extended Data Fig. 4). To make sure that the analysis does not rely on local dependencies among adjacent words, we repeated the arbitrary-based encoding analysis after removing bi-grams that repeated more than once in the dataset (Extended Data Fig. 6b). The ability to encode the neural activity for the upcoming words before word onset with the arbitrary embeddings remained significant.

To further demonstrate that predicting the next word before word onset goes above and beyond the contextual information embedded in the previous word, we ran an additional control analysis. In the control analysis, we encoded the neural activity using the arbitrary word embedding assigned to the previous word (Extended Data Fig. 6c). Next, we ran an encoding model using the concatenation of the previous and current word embeddings (Extended Data Fig. 6c). We found a significant difference between these two models before word onset. This indicates that the neural responses before word onset contained information related to the next word above and beyond the contextual information embedded in the previous word. Together, these results suggest that the brain is constantly engaged in the prediction of upcoming words before they are perceived as it processes natural language.

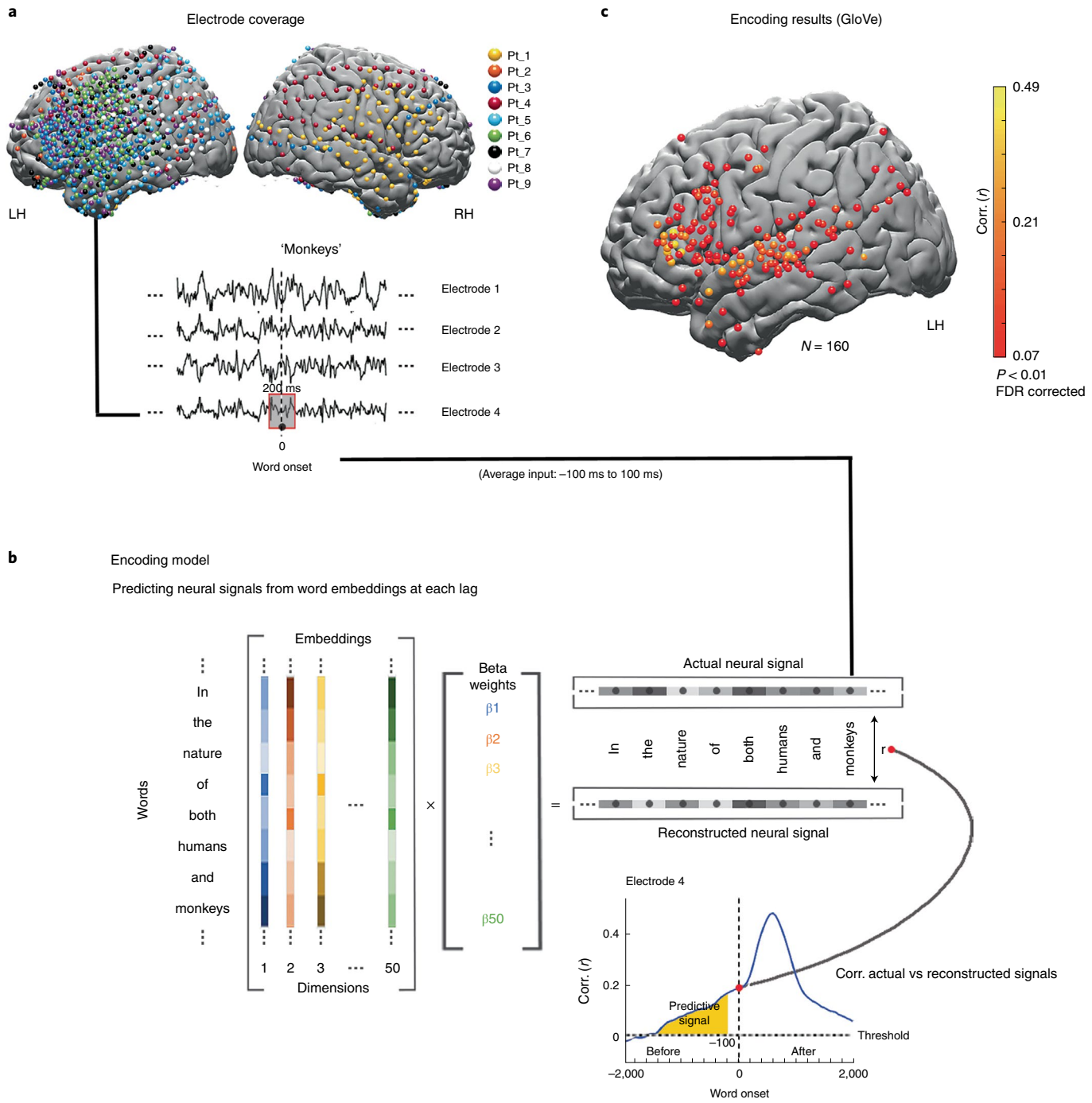


Fig. 3 | Linear encoding model used to predict the neural responses to each word in the narrative before and after word-onset. a, Brain coverage consisted of 1,339 electrodes (across nine participants). The words are aligned with the neural signal; each word's onset (moment of articulation) is designated at lag 0. Responses are averaged over a window of 200 ms and provided as input to the encoding model. **b**, A series of 50 coefficients corresponding to the features of the word embeddings is learned using linear regression to predict the neural signal across words from the assigned embeddings. The model was evaluated by computing the correlation between the reconstructed signal and the actual signal for a held-out test word. This procedure was repeated for each lag and each electrode, using a 25-ms sliding window. The dashed horizontal line indicates the statistical threshold ($q < 0.01$, FDR corrected). Lags of −100 ms or more preceding word onset contained only neural information sampled before the word was perceived (yellow). **c**, Electrodes with significant correlation at the peaked lag between predicted and actual word responses for semantic embeddings (GloVe). LH, left hemisphere; RH, right hemisphere.

Predictive neural signals for listeners' incorrect predictions. Pre-onset activity associated with next-word prediction should match the prediction content even when the prediction was incorrect. In contrast, post-onset activity should match the content of the incoming word, even if it was unpredicted. To test this hypothesis, we divided the signal into correct and incorrect predictions using GPT-2 (Methods) and computed encoding models. We also ran the same analyses using human predictions. We modeled the neural activity using: (1) the GloVe embeddings of the correctly predicted words (Fig. 4b); in this condition, the pre-onset word prediction matched the identity of the perceived incoming word; (2) the GloVe embedding for the incorrectly predicted words (Fig. 4b); and (3) the GloVe embedding

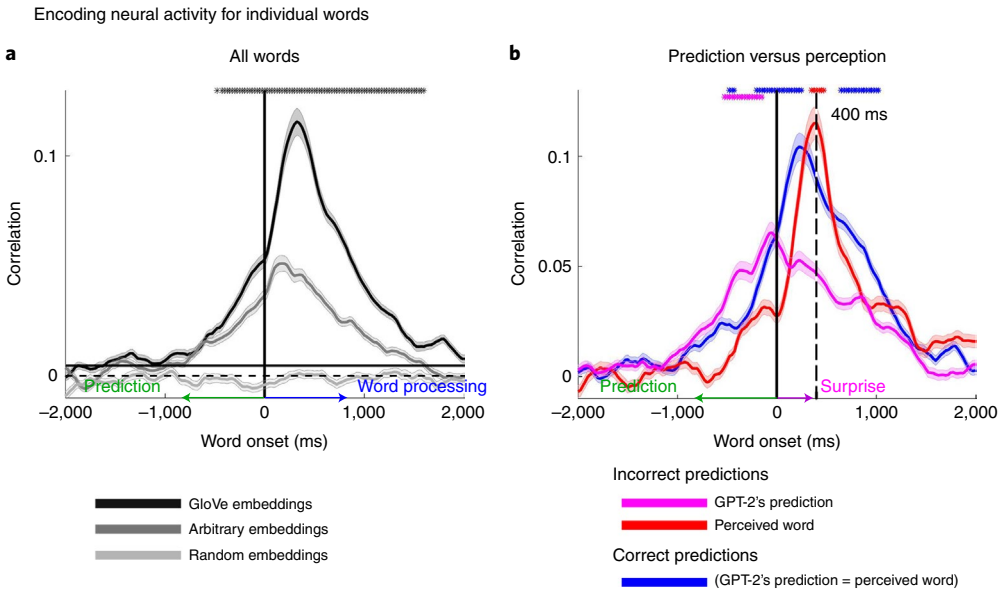


Fig. 4 | Modeling of neural signals before and after word onset for predictable, unpredictable and incorrectly predicted words. a, Estimating neural signals for all individual words from word embeddings (encodings). The encoding analysis was performed in each electrode with significant encoding for GloVe embeddings ($n=160$), and then averaged across electrodes (see map of electrodes in Fig. 3c). The shaded regions indicate the s.e. of the encoding models across electrodes. Using arbitrary embeddings, we managed to encode information as to the identity of the incoming word before and after word onset. Using word embeddings (GloVe), which contain contextual information as to the relation among words in natural language, further improves the encoding models before and after word onset. Furthermore, we observed a robust encoding to upcoming words starting −1,000 ms before word onset. The horizontal continuous black line specifies the statistical threshold. Black asterisks indicate lags for which the encoding based on GloVe embeddings significantly outperformed the encoding based on arbitrary embeddings. **b**, Estimating neural signals for correctly predicted words (blue), incorrectly predicted words (magenta) and the actual unexpected perceived word (red). Note that encoding before word onset was aligned with the content of the predicted words, whereas the encoding after word onset was aligned with the content of the perceived words. Moreover, we observed an increase in encoding performance for surprising words compared to predicted words 400 ms after word onset. Magenta asterisks represent significant differences between incorrect GPT-2 predictions (magenta line) and correct predictions (blue line). Red asterisks represent significantly higher values for incorrectly predicted words (red line) than correctly predicted words (blue line). Blue asterisks represent significantly higher values for correctly predicted words (blue line) than incorrectly predicted words (red line). The shaded regions indicate the s.e. of the encoding models across electrodes.

of the actual unpredictable words that humans perceived (Fig. 4b) because in the incorrect predictions condition the predicted word did not match the identity of the perceived word.

The neural responses before word onset contained information about human predictions regarding the identity of the next word. Crucially, the encoding was high for both correct and incorrect predictions (Fig. 4b and Extended Data Fig. 8). This demonstrated that pre-word-onset neural activity contains information about what listeners actually predicted, irrespective of what they subsequently perceived. Similar results were obtained using human predictions (Extended Data Fig. 8). In contrast, the neural responses after word onset contained information about the words that were actually perceived, irrespective of GPT-2's predictions (Fig. 4b). The analysis of the incorrect predictions unequivocally disentangles the pre-word-onset processes associated with word prediction from the post-word-onset comprehension-related processes. Furthermore, it demonstrates how autoregressive DLMs predictions can be used for modeling human predictions at the behavioral and neural levels.

In summary, these multiple pieces of evidence, which are based on encoding analyses, suggest that the brain, like autoregressive DLMs, is constantly predicting the next word before onset as it processes incoming natural speech. Next, we provide more evidence for coupling between pre-onset prediction and post-onset surprise level and error signals.

Pre-onset predictions and post-onset word surprise. Autoregressive language models provide a unified framework for modeling pre-onset next-word predictions and post-onset surprise (that is, prediction-error signals). We used pretrained GPT-2's internal estimates for each upcoming word (Fig. 1) to establish a connection between pre-onset prediction and post-onset surprise at the neural level.

Increased activity for surprise 400 ms after word onset. Autoregressive DLMs, such as GPT-2, use their pre-onset predictions to calculate the post-onset surprise level as to the identity of the incoming word. It was already shown that the activation level after onset is correlated with the surprise level[14,21–23,41]. We replicated this finding in our data. In addition, high-quality intracranial recordings allowed us to link pre-onset confidence level and the post-onset surprise level. Pre-onset confidence level was assessed using entropy (Methods), which is a measure of GPT-2's uncertainty level in its prediction before word onset. High entropy indicates that the model is uncertain about its predictions, whereas low entropy indicates that the model is confident. Post-onset surprise level was assessed using a cross-entropy measure that depends on the probability assigned to the incoming word before it is perceived (Fig. 1 and Methods). Assigning a low probability to the word before word onset will result in a post-onset high surprise when the word is perceived, and vice versa for high-probability words.

Pre-onset activity (using the same 160 electrodes used for Fig. 4a,b) increased for correct predictions, whereas, in agreement with prior research, post-onset activity increased for incorrect predictions (Fig. 5a). The activity level was averaged for all words that were correctly or incorrectly predicted. We observed increased activity for incorrect predictions 400 ms after word onset (Fig. 5a). In addition,

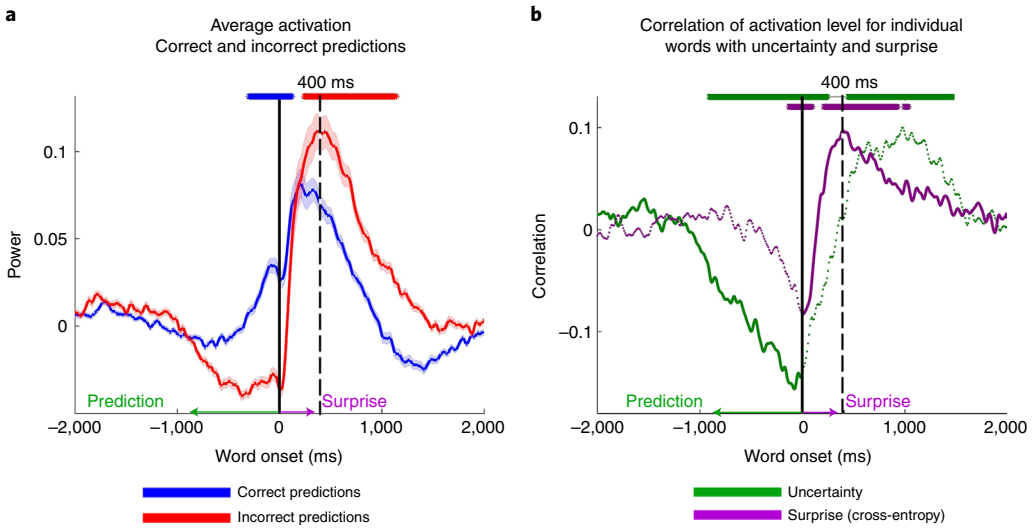


Fig. 5 | Uncertainty and surprise levels computed by GPT-2 correlate with pre-onset and post-onset neural activity respectively. a, Trigger-averaged activity (and s.e.) across 160 electrodes with significant GloVe encoding for words (Fig. 3c). Averaging was performed separately for words correctly predicted and incorrectly predicted. Note the increase in signal activity for predictable words before onset and for unpredictable words 400 ms after word onset. **b**, Partial correlations between uncertainty (entropy) and signal power controlling for cross-entropy (green line). Partial correlations between surprise (cross-entropy) and neural signals controlling for correlation with entropy (red and purple lines). Asterisks indicate correlation significance (FDR corrected, $q < 0.01$).

GPT-2's uncertainty (entropy) was negatively correlated with the activation level before word onset (Fig. 5b). In other words, before onset, the higher the confidence (low uncertainty), the higher the activation level. In contrast, after word onset, the level of surprise (cross-entropy) was correlated with activation, and peaked around 400 ms (Fig. 5b). Because uncertainty correlates with surprise, we computed partial correlations between entropy, surprise and neural signals. This analysis directly connects GPT-2's internal predictions and neural activity before word onset and GPT-2's internal surprise and the surprise (that is, prediction error) embedded in the neural responses after word onset.

In summary, based on encoding and event-related activity, we introduce multiple pieces of evidence to link pre-onset next-word prediction processes with post-onset surprise processes using GPT-2's internal estimates. This section further supports the claim that autoregressive DLMs can serve a theoretical framework for language comprehension-related processes. Next, we provide more evidence that GPT-2 tracks human neural signals and specifically that humans represent words in a context-dependent fashion, similarly to DLMs.

Contextual representation. *Contextual embeddings predict neural responses to speech.* A next-word prediction objective enables autoregressive DLMs to compress a sequence of words into a contextual embedding from which the model decodes the next word. The present results have established that the brain, similarly to autoregressive DLMs, is also engaged in spontaneous next-word prediction as it listens to natural speech. Given this shared computational principle, we investigated whether the brain, like autoregressive DLMs, compresses word sequences into contextual representation.

In natural language, each word receives its full meaning based on the preceding words[42–44]. For instance, consider how the word 'shot' can have very different meanings in different contexts, such as 'taking a shot with the camera', 'drinking a shot at the bar' or 'making the game-winning shot'. Static word embeddings, like GloVe, assign one unique vector to the word 'shot' and, as such, cannot capture the context-specific meaning of the word. In contrast, contextual embeddings assign a different embedding (vector) to every word as a function of its preceding words. Here we tested whether autoregressive DLMs that compress context into contextual embeddings provide a better cognitive model for neural activity during linguistic processing than static embeddings. To test this, we extracted the contextual embeddings from an autoregressive DLM (GPT-2) for each word in the story. To extract the contextual embedding of a word, we provided the model with the preceding sequence of all prior words (up to 1,024 tokens) in the podcast and extracted the activation of the top embedding layer (Methods).

Localizing neural responses using contextual embeddings. Replacing static embeddings (GloVe) with contextual embeddings (GPT-2) improved encoding model performance in the prediction of neural responses to words (Fig. 6a and Extended Data Fig. 3). Encoding based on contextual embeddings resulted in statistically significant correlations in 208 electrodes in the left hemisphere (and 34 in the right hemisphere), 71 of which were not significantly predicted by the static embeddings (GloVe). The additional electrodes revealed by contextual embedding were mainly located in higher-order language areas with long processing timescales along the inferior frontal gyrus, temporal pole, posterior superior temporal gyrus, parietal lobe and angular gyrus[37]. In addition, there was a noticeable improvement in the contextual embedding-based encoding model in the primary and supplementary motor cortices. The improvement was seen both at the peak of the encoding model and in the model's ability to predict neural responses to words up to 4 s before word onset (for the 160 electrodes with significant GloVe encoding; Fig. 6b and Extended Data Figs. 4 and 9). The improvement in the ability to predict neural signals to each word while relying on autoregressive DLM's contextual embeddings was robust and apparent even at the single-electrode level (Extended Data Fig. 4). These results agree with concurrent studies demonstrating that contextual embeddings model neural responses to words better than static semantic embeddings[15,16,45,46]. Next, we asked which aspects of the contextual embedding drive the improvement in modeling the neural activity.

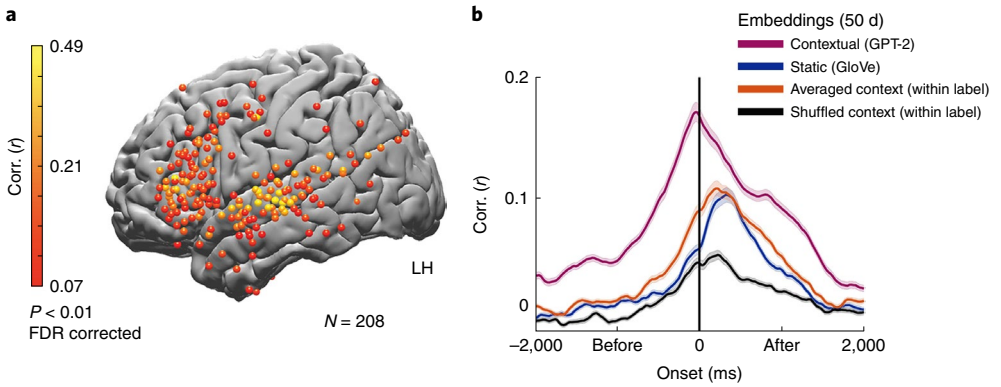


Fig. 6 | Contextual (GPT-2) embeddings improve the modeling of neural responses before word onset. a, Peak correlation between predicted and actual word responses for the contextual (GPT-2) embeddings. Using contextual embeddings significantly improved the encoding model's ability to predict the neural signals for unseen words across many electrodes. **b**, Encoding model performance for contextual embeddings (GPT-2) aggregated across 160 electrodes with significant encoding for GloVe (Fig. 3c): contextual embeddings (purple), static embeddings (GloVe, blue), contextual embeddings averaged across all occurrences of a given word (orange), contextual embeddings shuffled across context-specific occurrence of a given word (black). The shaded regions indicate the s.e. of the encoding models across electrodes.

Modeling the context versus predicting the upcoming word. The improved ability to predict neural responses before word onset using contextual embedding can be attributed to two related factors that are absent in the static word embeddings (for example, GloVe): (1) the brain, like GPT-2, aggregates information about the preceding words in the story into contextual embeddings; and (2) GPT-2 embeddings contain additional predictive information, not encoded in static embeddings, about the identity of the upcoming word in the sequence. By carefully manipulating the contextual embeddings and developing an embedding-based decoder, we show how both context and next-word prediction contribute to contextual embeddings' improved ability to model the neural responses.

Representing word meaning in unique contexts. Going above and beyond the information encoded in GloVe, GPT-2's capacity for representing context captures additional information in neural responses. A simple way to represent the context of prior words is to combine (that is, concatenate) the static embeddings of the preceding sequence of words. To test this simpler representation of context, we concatenated GloVe embeddings for the ten preceding words in the text into a longer 'context' vector and compared the encoding model performance to GPT-2's contextual embeddings (after reducing both vectors to 50 dimensions using principal-component analysis). While the concatenated static embeddings were better in predicting the prior neural responses than the original GloVe vectors, which only capture the current word, they still underperformed GPT-2's encoding before word articulation (Extended Data Fig. 9). This result suggests that GPT-2's contextual embeddings are better suited to compress the contextual information embedded in the neural responses than static embeddings (even when concatenated).

A complementary way to demonstrate that contextual embeddings uncover aspects of the neural activity that static embeddings cannot capture is to remove the unique contextual information from GPT-2 embeddings. We removed contextual information from GPT-2's contextual embeddings by averaging all embeddings for each unique word (for example, all occurrences of the word 'monkey') into a single vector. This analysis was limited to words that have at least five repetitions (Methods). Thus, we collapsed the contextual embedding into a 'static' embedding in which each unique word in the story is represented by one unique vector. The resulting embeddings were still specific to the overall topic of this particular podcast (unlike GloVe). Still, they did not contain the local context for each occurrence of a given word (for example, the context in which 'monkey' was used in sentence 5 versus the context in which it was used in sentence 50 of the podcast). Indeed, removing context from the contextual embedding by averaging the vector for each unique word effectively reduced the encoding model's performance to that of the static GloVe embeddings (Fig. 6b).

Finally, we examined how the specificity of the contextual information in the contextual embeddings improved the ability to model the neural responses to each word. To that end, we scrambled the embeddings across different occurrences of the same word in the story (for example, switched the embedding of the word 'monkey' in sentence 5 with the embedding for the word 'monkey' in sentence 50). This manipulation tests whether contextual embeddings are necessary for modeling neural activity for a specific sequence of words. Scrambling the same word occurrences across contexts substantially reduced the encoding model performance (Fig. 6b), pointing to the contextual dependency represented in the neural signals. Taken together, these results suggest that contextual embeddings provide us with a new way to model the context-dependent neural representations of words in natural contexts.

Predicting words from neural signal using contextual embeddings. Finally, we applied a decoding analysis to demonstrate that, in addition to better modeling the neural responses to context, contextual embeddings also improved our ability to read information from the neural responses as to the identity of upcoming words. This demonstrates the duality of representing the context and the next-word prediction in the brain.

The encoding model finds a mapping from the embedding space to the neural responses that is used during the test phase for predicting neural responses. The decoding analysis inverts this procedure to find a mapping from neural responses, across multiple electrodes and time points, to the embedding space[47]. This decoding analysis provides complementary insights to the encoding analysis by aggregating across electrodes and quantifies how much predictive information about each word's identity is embedded in the spatio-temporal neural activity patterns before and after word onset.

The decoding analysis was performed in two steps. First, we trained a deep convolutional neural network to aggregate neural responses (Fig. 7a and Appendix I) and mapped this neural signal to the arbitrary, static (GloVe-based) and to the contextual (GPT-2-based) embedding spaces (Fig. 7b). To conservatively compare the

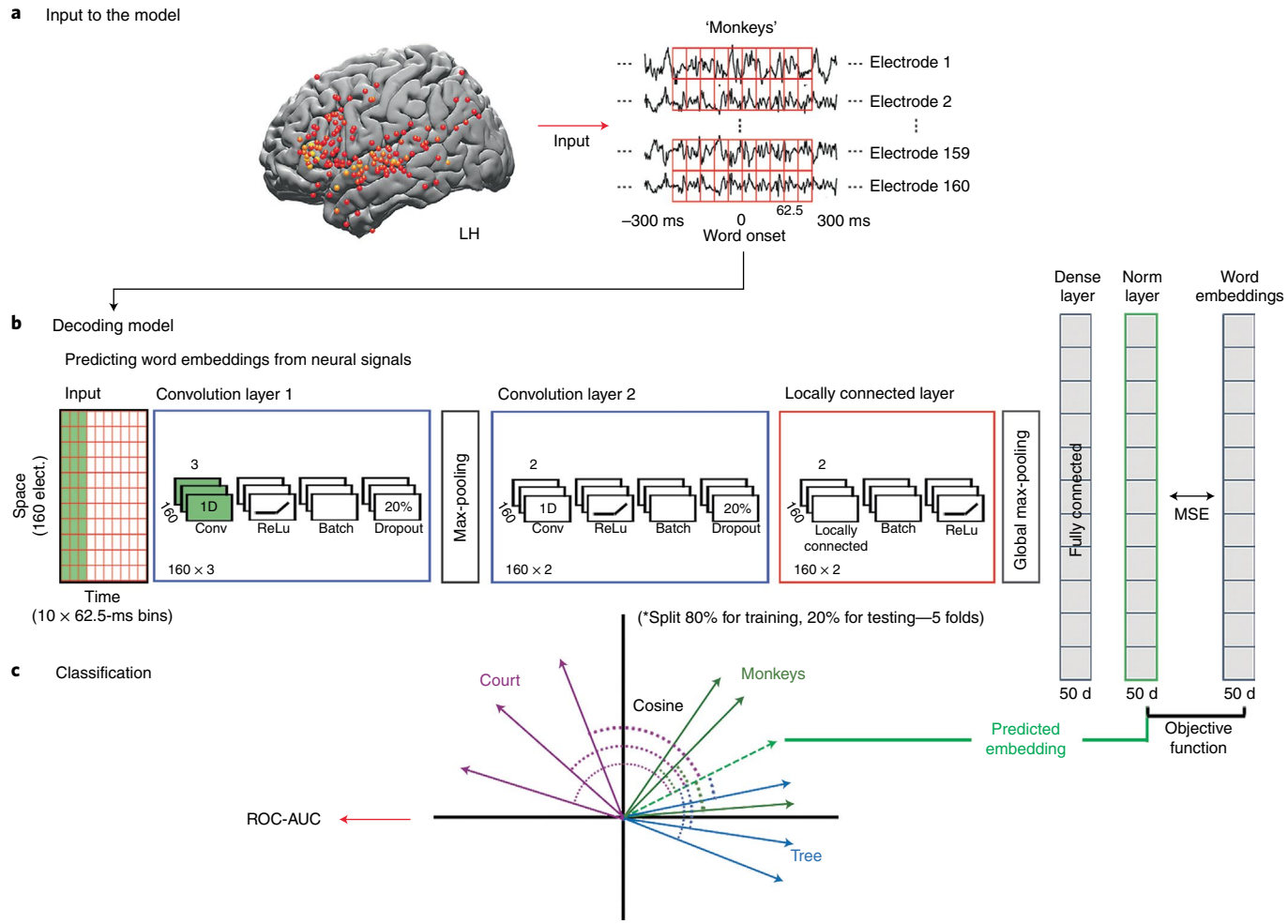


Fig. 7 | Deep nonlinear decoding model used to predict words from neural responses before and after word onset. a, Neural data from left-hemisphere electrodes with significant encoding model performance using GloVe embeddings were used as input to the decoding model. For each fold, electrode selection was performed on 80% of the data that were not used for testing the model. The stimulus was segmented into individual words and aligned to the brain signal at each lag. **b**, Schematic of the feed-forward deep neural network model that learns to project the neural signals for the words into the arbitrary embedding, static semantic embedding (GloVe) or contextual embedding (GPT-2) space (Appendix I). The input (currently represented as a 160×10 matrix) changes its dimensions for each of the five folds based on the number of significant electrodes for each fold. The model was trained to minimize the mean squared error (MSE) when mapping the neural signals into the embedding space. **c**, The decoding model was evaluated using a word classification task. The quality of word classification is based on the embedding space used to construct ROC-AUC scores. This enabled us to assess how much information about specific words is extractible from the neural activity via the linguistic embedding space.

performance of GPT-2-based embedding to GloVe-based embedding, we used as input the signal from the electrodes that were found to be significant for GloVe-based encoding. To further ensure that the decoding results were not affected by the electrode selection procedure, for each test fold, we selected the electrodes using the remaining 80% of the data (Methods). To obtain a reliable estimation of accuracy per word label we included words with at least five repetitions, which included 69% of the words in the story (for the full list of words, see Appendix II). Second, the predicted word embeddings were used for word classification based on their cosine distance from all embeddings in the dataset (Fig. 7c). Although we evaluated the decoding model using classification, the classifier predictions were constrained to rely only on the embedding space. This is a more conservative approach than an end-to-end word classification, which may capitalize on acoustic information in the neural signals that are not encoded in the language models.

Using a contextual decoder greatly improved our ability to classify word identity over decoders relying on static or arbitrary embeddings (Fig. 8). We evaluated classification performance using the receiver operating characteristic (ROC) analysis with corresponding area under the curve (AUC). A model that only learns to use word frequency statistics (for example, only guessing the most frequent word) will result in a ROC-AUC curve that falls on the diagonal line (AUC = 0.5) suggesting that the classifier does not discriminate between the words[48]. Classification using GPT-2 (average AUC of 0.74 for lag 150) outperformed GloVe and arbitrary embeddings (average AUC of 0.68 and 0.68, respectively) before and after word onset. To compare the performance of the classifiers based on GPT-2 and GloVe at each lag, we performed a paired-sample *t*-test between the AUCs of the words in the two models. Each unique word (class) in each lag had an AUC value computed from the GloVe-based model and an AUC value computed from the GPT-2-based model. The results were corrected for multiple tests by controlling the false discovery rate (FDR)[49].

A closer inspection of the GPT-2-based decoder indicated that the classifier managed to detect reliable information about the identity of words several hundred milliseconds before word onset (Fig. 8). In particular, starting at about −1,000 ms before word onset, when

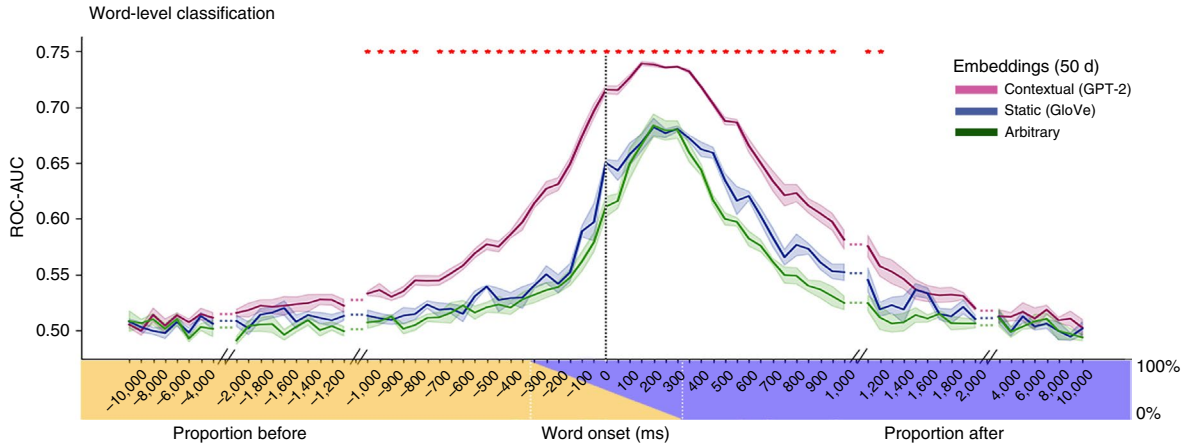


Fig. 8 | Using a decoding model for classification of words before and after word onset. Word-level classification. Classification performance for contextual (GPT-2), static (GloVe) and arbitrary (green) embeddings. The averaged values were weighted by the frequency of the words in the test set. The *x*-axis labels indicate the center of each 625-ms window used for decoding at each lag (between −10 and 10 s). The colored stripe indicates the proportion of pre-word onset (yellow) and post-word onset (blue) time points in each lag. Shaded regions denote the s.e. across five test folds. Note that contextual embeddings improved classification performance over GloVe both before and after word onset. Significance was assessed using a paired-sample *t*-test of the AUCs for each unique word, comparing the AUCs of the GloVe-based decoding and GPT-2-based decoding. The comparison was performed for each leg separately and results were FDR corrected ($q < .01$).

the neural signals were integrated across a window of 625 ms, the classifier detected predictive information about the next word's identity. The information about the next word's identity gradually increased and peaked at an average AUC of 0.74 at a lag of 150 ms after word onset, when the signal was integrated across a window from −162.5 ms to 462.5 ms. GloVe embeddings showed a similar trend with a marked reduction in classifier performance (Fig. 8). The decoding model's capacity to classify words before word onset demonstrates that the neural signal contains a considerable amount of predictive information about the meaning of the next word, up to a second before it is perceived. At longer lags (more than 2 s), all decoders' performance dropped to chance.

Discussion

DLMs provide a new modeling framework that drastically departs from classical psycholinguistic models. They are not designed to learn a concise set of interpretable syntactic rules to be implemented in novel situations, nor do they rely on part of speech concepts or other linguistic terms. Rather, they learn from surface-level linguistic behavior to predict and generate the contextually appropriate linguistic outputs. The current paper provides compelling behavioral and neural evidence for shared computational principles between the way the human brain and autoregressive DLMs process natural language.

Spontaneous prediction as a keystone of language processing. Autoregressive DLMs learn according to the simple self-supervised objective of context-based next-word prediction. The extent to which humans are spontaneously engaged in next-word predictions as they listen to continuous, minutes-long, natural speech has been underspecified. Our behavioral results revealed a robust capacity for next-word prediction in real-world stimuli, which matches a modern autoregressive DLM (Fig. 2). Neurally, our findings demonstrate that the brain constantly and actively represents forthcoming words in context during listening to natural speech. The predictive neural signals are robust, and can be detected hundreds of milliseconds before word onset. Notably, the next-word prediction processes are associated with listeners' contextual expectations and can be dissociated from the processing of the actually perceived words after word onset (Fig. 4b and Extended Data Fig. 8).

Pre-onset prediction is coupled with post-onset surprise. Autoregressive DLMs (like GPT-2) provide a unified computational framework that connects pre-onset word prediction with post-onset surprise (error signals). Our results show that we can rely on GPT-2's internal pre-onset confidence (entropy) and post-onset surprise (cross-entropy) to model the brain's internal neural activity as it processes language. The correlations between GPT-2's internal surprise level and the neural signals corroborate the link between the two systems[50].

Context-specific meaning as a keystone of language processing. As each word attains its full meaning in the context of preceding words over multiple timescales, language is fundamentally contextual[51]. Even a single change to one word or one sentence at the beginning of a story can alter the neural responses to all subsequent sentences[43,52]. We demonstrated that the contextual word embeddings learned by DLMs provide a new way to compress linguistic context into a numeric vector space, which outperforms the use of static semantic embeddings (Figs. 6b and 8 and Extended Data Figs. 4 and 9). While static embeddings and contextual embeddings are different, our neural results also hint at how they relate to each other. Our results indicate that both static and contextual embeddings can predict neural responses to single words in many language areas[16] along the superior temporal cortex, parietal lobe and inferior frontal gyrus. Switching from static to contextual embeddings boosted our ability to model neural responses during natural speech processing across many of these brain areas. Finally, averaging contextual embeddings associated with a given word removed the contextual information and effectively changed GPT-2's contextual embedding back into static word embeddings (Fig. 6b). Taken together, these results suggest that the brain is coding for the semantic relationship among words contained in static embeddings while also being tuned to the unique contextual relationship between the specific word and the preceding words in the sequence[53].

Using an autoregressive language model as a cognitive model. We describe three shared computational principles that reveal a strong link between the way the brain and DLMs process natural language. These shared computational principles, however, do not imply that the human brain and DLMs implement these computations in a

similar way. For example, many state-of-the-art DLMs rely on transformers, a type of neural network architecture developed to solve sequence transduction. Transformers are designed to parallelize a task that is largely computed serially, word by word, in the human brain. Therefore, while transformer models are an impressive engineering achievement, they are not biologically feasible. Many other ways, however, are possible to transduce a sequence into a contextual embedding vector. To the extent that the brain relies on a next-word prediction objective to learn how to use language in context, it likely uses a different implementation[54].

Psycholinguistic models versus deep language models. DLMs were engineered to solve a fundamentally different problem than psycholinguistic language models. Psycholinguistic language models aim to uncover a set of generative (learned or innate) rules to be used in infinite, novel situations[55]. Finding a set of linguistic rules, however, was challenging. There are numerous exceptions for every rule, conditioned by discourse context, meaning, dialect, genre, and many other factors[51,56–58]. In contrast, DLMs aim to provide the appropriate linguistic output given the prior statistics of language use in similar contexts[20,59]. In other words, psycholinguistic theories aim to describe observed language in terms of a succinct set of explanatory constructs. DLMs, in contrast, are performance oriented and are focused on learning how to generate formed linguistic outputs as a function of context while de-emphasizing interpretability[60]. The reliance on performance creates an interesting connection between DLMs and usage (context)-based constructionist approaches to language[58,61]. Furthermore, DLMs avoid the circularity built into many psycholinguistic language models that rely on linguistic terms to explain how language is encoded in neural substrates[19,62]. Remarkably, the internal contextual embedding space in DLMs can capture many aspects of the latent structure of human language, including syntactic trees, voice, co-references, morphology and long-range semantic and pragmatic dependencies[1,63,64]. This discussion demonstrates the power (over the more traditional approaches) of applying brute-force memorization and interpolation for learning how to generate the appropriate linguistic outputs in light of prior contexts[20].

Observational work in developmental psychology suggests that children are exposed to tens of thousands of words in contextualized speech each day, creating a large data volume available for learning[65,66]. The capacity of DLMs to learn language relies on gradually exposing the model to millions of real-life examples. Our finding of spontaneous predictive neural signals as participants listen to natural speech suggests that active prediction may underlie humans' lifelong language learning. Future studies, however, will have to assess whether these cognitively plausible, prediction-based feedback signals are at the basis of human language learning and whether the brain is using such predictive signals to guide language acquisition. Furthermore, as opposed to autoregressive DLMs, it is likely that the brain relies on additional simple objectives at different timescales to facilitate learning[20,67].

Conclusion

This paper provides evidence for three shared core computational principles between DLMs and the human brain. While DLMs may provide a building block for our high-level cognitive faculties, they undeniably lack certain central hallmarks of human cognition. Linguists were primarily interested in how we construct well-formed sentences, exemplified by the famous grammatically correct but meaningless sentence 'colorless green ideas sleep furiously', composed by Noam Chomsky[2]. Similarly, DLMs are generative in the narrow linguistic sense of being able to generate new sentences that are grammatically, semantically and even pragmatically well-formed at a superficial level. However, although language may play a central organizing role in our cognition, linguistic competence is insufficient to capture thinking. Unlike humans, DLMs cannot think, understand or generate new meaningful ideas by integrating prior knowledge. They simply echo the statistics of their input[68]. Going beyond the importance of language as having a central organizing role in our cognition, DLMs indicate that linguistic competence may be insufficient to capture thinking. A core question for future studies in cognitive neuroscience and machine learning is how the brain can leverage predictive, contextualized linguistic representations, like those learned by DLMs, as a substrate for generating and articulating new thoughts.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-022-01026-4.

Received: 13 January 2021; Accepted: 27 January 2022;
Published online: 7 March 2022

References

1. Linzen, T. & Baroni, M. Syntactic structure from deep learning. *Annu. Rev. Linguist.* **7**, 195–212 (2021).
2. Chomsky, N. Syntactic structures. https://doi.org/10.1515/9783112316009 (1957).
3. Jacobs, R. A. & Rosenbaum, P. S. *English Transformational Grammar* (Blaisdell, 1968).
4. Brown, T. B. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
5. Cho, W. S. et al. Towards coherent and cohesive long-form text generation. in *Proceedings of the First Workshop on Narrative Understanding* https://doi.org/10.18653/v1/w19-2401 (2019).
6. Yang, Z. et al. XLNet: generalized autoregressive pretraining for language understanding. in *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) 5753–5763 (Curran Associates, 2019).
7. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. *OpenAI Blog* (2018).
8. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI Blog* (2019).
9. Rosset, C. Turing-nlg: A 17-billion-parameter language model by microsoft. *Microsoft Blog* (2019).
10. Pereira, F. et al. Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* **9**, 963 (2018).
11. Makin, J. G., Moses, D. A. & Chang, E. F. Machine translation of cortical activity to text with an encoder–decoder framework. *Nat. Neurosci.* **23**, 575–582 (2020).
12. Schwartz, D. et al. Inducing brain-relevant bias in natural language processing models. in *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) 14123–14133 (Curran Associates, 2019).
13. Gauthier, J. & Levy, R. Linking artificial and human neural representations of language. in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* https://doi.org/10.18653/v1/d19-1050 (2019).
14. Donhauser, P. W. & Baillet, S. Two distinct neural timescales for predictive speech processing. *Neuron* **105**, 385–393 (2020).
15. Jain, S. & Huth, A. G. Incorporating context into language encoding models for fMRI. in *Advances in Neural Information Processing Systems* https://doi.org/10.1101/327601 (2018).
16. Schrimpf, M. et al. Artificial neural networks accurately predict language processing in the Brain. *Cold Spring Harbor Laboratory* https://doi.org/10.1101/2020.06.26.174482 (2020).
17. Caucheteux, C., Gramfort, A. & King, J. -R. GPT-2's activations predict the degree of semantic comprehension in the human brain. https://doi.org/10.1101/2021.04.20.440622 (2021).
18. Athanasiou, N., Iosif, E. & Potamianos, A. Neural activation semantic models: computational lexical semantic models of localized neural activations. in *Proceedings of the 27th International Conference on Computational Linguistics* 2867–2878 (Association for Computational Linguistics, 2018).
19. McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J. & Schütze, H. Placing language in an integrated understanding system: next steps toward human-level performance in neural language models. *Proc. Natl Acad. Sci. USA* **117**, 25966–25974 (2020).

20. Hasson, U., Nastase, S. A. & Goldstein, A. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron* **105**, 416–434 (2020).
21. Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P. & de Lange, F. P. A hierarchy of linguistic predictions during natural language comprehension. Preprint at *bioRxiv* https://doi.org/10.1101/2020.12.03.410399 (2020).
22. Weissbart, H., Kandylaki, K. D. & Reichenbach, T. Cortical tracking of surprisal during continuous speech comprehension. *J. Cogn. Neurosci.* **32**, 155–166 (2020).
23. Frank, S. L., Otten, L. J., Galli, G. & Vigliocco, G. The ERP response to the amount of information conveyed by words in sentences. *Brain Lang.* **140**, 1–11 (2015).
24. Caucheteux, C., Gramfort, A. & King, J.-R. GPT-2's activations predict the degree of semantic comprehension in the human brain. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.20.440622 (2021).
25. Lewis, M. et al. BART: denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. Preprint at https://arxiv.org/abs/1910.13461 (2019).
26. Huang, Y. & Rao, R. P. N. Predictive coding. *Wiley Interdiscip. Rev. Cogn. Sci.* **2**, 580–593 (2011).
27. Lupyan, G. & Clark, A. Words and the world: predictive coding and the language–perception–cognition interface. *Curr. Dir. Psychol. Sci.* **24**, 279–284 (2015).
28. Barron, H. C., Auksztulewicz, R. & Friston, K. Prediction and memory: a predictive coding account. *Prog. Neurobiol.* **192**, 101821 (2020).
29. Goldstein, A., Rivlin, I., Goldstein, A., Pertzov, Y. & Hassin, R. R. Predictions from masked motion with and without obstacles. *PLoS ONE* **15**, e0239839 (2020).
30. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).
31. Taylor, W. L. 'Cloze Procedure': a new tool for measuring readability. *Journal Q.* **30**, 415–433 (1953).
32. Kliegl, R., Nuthmann, A. & Engbert, R. Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *J. Exp. Psychol. Gen.* **135**, 12–35 (2006).
33. Laurinavichyute, A. K., Sekerina, I. A., Alexeeva, S., Bagdasaryan, K. & Kliegl, R. Russian sentence corpus: benchmark measures of eye movements in reading in Russian. *Behav. Res. Methods* **51**, 1161–1178 (2019).
34. Kutas, M. & Federmeier, K. D. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* **62**, 621–647 (2011).
35. Kutas, M. & Hillyard, S. A. Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* **207**, 203–205 (1980).
36. Chivvis & Dana. 'So a Monkey and a Horse Walk Into a Bar' (2017).
37. Hasson, U., Chen, J. & Honey, C. J. Hierarchical process memory: memory as an integral component of information processing. *Trends Cogn. Sci.* **19**, 304–313 (2015).
38. Pennington, J., Socher, R. & Manning, C. GloVe: global vectors for word representation. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* 1532–1543 (Association for Computational Linguistics, 2014).
39. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. in *Advances in Neural Information Processing Systems* (eds. Burges et al.) 3111–3119 (Curran Associates, 2013).
40. van Gerven, M. A. J. A primer on encoding models in sensory neuroscience. *J. Math. Psychol.* **76**, 172–183 (2017).
41. Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P. & van den Bosch, A. Prediction during natural language comprehension. *Cereb. Cortex* **26**, 2506–2516 (2016).
42. Chen, J., Hasson, U. & Honey, C. J. Processing timescales as an organizing principle for primate cortex. *Neuron* **88**, 244–246 (2015).
43. Yeshurun, Y., Nguyen, M. & Hasson, U. Amplification of local changes along the timescale processing hierarchy. *Proc. Natl Acad. Sci. USA* **114**, 9475–9480 (2017).
44. Hasson, U., Yang, E., Vallines, I., Heeger, D. J. & Rubin, N. A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* **28**, 2539–2550 (2008).
45. Wehbe, L., Vaswani, A., Knight, K. & Mitchell, T. Aligning context-based statistical models of language with brain activity during reading. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* 233–243 (Association for Computational Linguistics, 2014).
46. Toneva, M. & Wehbe, L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). in *33rd Conference on Neural Information Processing Systems* (2019).
47. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *Neuroimage* **56**, 400–410 (2011).
48. Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol.* **5**, 1315–1316 (2010).
49. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300 (1995).
50. Schwartz, D. & Mitchell, T. Understanding language-elicited EEG data by predicting it from a fine-tuned language model. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1, 43–57 (2019).
51. Goldberg, A. E. *Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions* (Princeton University Press, 2019).
52. Yeshurun, Y. et al. Same story, different story. *Psychol. Sci.* **28**, 307–319 (2017).
53. Ethayarajh, K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. Preprint at https://arxiv.org/abs/1909.00512 (2019).
54. Richards, B. A. et al. A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).
55. Chomsky, N. Aspects of the theory of syntax. https://doi.org/10.21236/ad0616323 (1964).
56. Ten Hacken, P. Andrew Radford. *Syntactic Theory and the Structure of English: a minimalist approach* (Cambridge University Press, 1997). *Andrew Radford. Syntax: a Minimalist Introduction*. (Cambridge University Press, 1997). *Natural Language Engineering*, **7**, 87–97 (2001).
57. Boer, Bde & de Boer, B. The atoms of language: the mind's hidden rules of grammar; foundations of language: brain, meaning, grammar, evolution. *Artif. Life* **9**, 89–91 (2003).
58. Bybee, J. & McClelland, J. L. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review* **22**, 381–410 (2005).
59. Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L. & Lewis, M. Generalization through memorization: nearest neighbor language models. in *International Conference on Learning Representations* (2020).
60. Breiman, L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *SSO Schweiz. Monatsschr. Zahnheilkd.* **16**, 199–231 (2001).
61. Goldberg, A. E. *Explain me This: Creativity, Competition, and the Partial Productivity of Constructions* (Princeton University Press, 2019).
62. Hasson, U., Egidi, G., Marelli, M. & Willems, R. M. Grounding the neurobiology of language in first principles: the necessity of non-language-centric explanations for language comprehension. *Cognition* **180**, 135–157 (2018).
63. Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U. & Levy, O. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl Acad. Sci. USA* **117**, 30046–30054 (2020).
64. Mamou, J. et al. Emergence of separable manifolds in deep language representations. *ICML* (2020).
65. Hart, B. & Risley, T. R. *Meaningful Differences In The Everyday Experience of Young American Children* (Brookes Publishing, 1995).
66. Weisleder, A. & Fernald, A. Talking to children matters: early language experience strengthens processing and builds vocabulary. *Psychol. Sci.* **24**, 2143–2152 (2013).
67. Tan, H. & Bansal, M. Vokenization: improving language understanding with contextualized, visual-grounded supervision. *EMNLP* (2020).
68. Marcus, G. F. *The Algebraic Mind: Integrating Connectionism and Cognitive Science* (MIT Press, 2019).

Methods

Transcription and alignment. The stimuli for the behavioral and electrocorticography experiments were extracted from the story 'So a Monkey and a Horse Walk Into a Bar: Act One, Monkey in the Middle'. The story was manually transcribed. Sounds such as laughter, breathing, lip smacking, applause and silent periods were also marked to improve the alignment's accuracy. The audio was downsampled to 11 kHz and the Penn Phonetics Lab Forced Aligner was used to automatically align the audio to the transcript[69]. After automatic alignment was complete, the alignment was manually evaluated and improved by an independent listener.

Behavioral word-prediction experiment. To obtain a continuous measure of prediction, we developed a sliding-window behavioral paradigm where healthy adult participants made predictions for each upcoming word in the story. A total of 300 participants completed a behavioral experiment on Mechanical Turk for a fee of $10 (data about age and gender were not collected). We divided the story into six segments and recruited six nonoverlapping groups of 50 participants to predict every upcoming word within each segment of the story. The first group was exposed to the first two words in the story and then asked to predict the upcoming word. After entering their prediction, the actual next word was revealed, and participants were asked again to predict the next upcoming word in the story. Once ten words were displayed on the screen, the left-most word was removed and the next word was presented (Fig. 2b). The procedure was repeated, using a sliding window until the first group provided predictions for each word in the story's first segment. Each of the other five groups listened uninterruptedly to the prior segments of the narrative and started to predict the next word at the beginning of their assigned segments. Due to a technical error, data for 33 words were omitted, and thus the final data contained 5,078 words. Importantly, before calculating the scores we used Excel's spellchecker to locate and correct spelling mistakes.

n-gram models. We trained 2- to 5-gram models using the NLTK Python package and its built-in 'Brown' corpus. All punctuations were removed and letters lowercased. We trained separate models using no-smoothing, Laplace smoothing or Kneser–Ney smoothing. Then we used each model to extract the probability of a word given its preceding $n-1$ context in the podcast transcript. We also extracted the most likely next word prediction to compare agreement with human responses.

Electrocorticography experiment. Ten participants (five females, aged 20–48 years) listened to the same story stimulus from beginning to end. Participants were not explicitly made aware that we would be examining word prediction in our subsequent analyses. One participant was removed from further analyses due to excessive epileptic activity and low signal-to-noise ratio across all experimental data collected. All participants volunteered for this study via the New York University School of Medicine Comprehensive Epilepsy Center. All participants had elected to undergo intracranial monitoring for clinical purposes and provided oral and written informed consent before study participation, according to the New York University Langone Medical Center Institutional Review Board. Participants were informed that participation in the study was unrelated to their clinical care and that they could withdraw from the study at any point without affecting their medical treatment.

For each participant, electrode placement was determined by clinicians based on clinical criteria. One participant consented to have a US Food and Drug Administration-approved hybrid clinical-research grid implanted, which includes standard clinical electrodes as well as additional electrodes in between clinical contacts. The hybrid grid provides a higher spatial coverage without changing clinical acquisition or grid placement. Across all participants, 1,106 electrodes were placed on the left hemisphere and 233 on the right hemisphere. Brain activity was recorded from a total of 1,339 intracranially implanted subdural platinum–iridium electrodes embedded in silastic sheets (2.3-mm-diameter contacts, Ad-Tech Medical Instrument; for the hybrid grids 64 standard contacts had a diameter of 2 mm and additional 64 contacts were 1 mm in diameter, PMT). Decisions related to electrode placement and invasive monitoring duration were determined solely on clinical grounds without reference to this or any other research study. Electrodes were arranged as grid arrays (8×8 contacts, 10- or 5-mm center-to-center spacing), or linear strips.

Recordings from grid, strip and depth electrode arrays were acquired using one of two amplifier types: NicoletOne C64 clinical amplifier (Natus Neurologics), band-pass filtered from 0.16–250 Hz, and digitized at 512 Hz; NeuroWorks Quantum Amplifier recorded at 2,048 Hz, high-pass filtered at 0.01 Hz and then resampled to 512 Hz. Intracranial electroencephalography signals were referenced to a two-contact subdural strip facing toward the skull near the craniotomy site. All electrodes were visually inspected, and those with excessive noise artifacts, epileptiform activity, excessive noise or no signal were removed from subsequent analysis (164 of 1,065 electrodes removed).

Presurgical and postsurgical T1-weighted magnetic resonance imaging (MRI) scans were acquired for each participant, and the location of the electrodes relative to the cortical surface was determined from co-registered magnetic resonance imaging or computed tomography scans following the procedure described by Yang and colleagues[70]. Co-registered, skull-stripped T1 images were nonlinearly registered to an MNI152 template and electrode locations were then extracted in Montreal Neurological Institute space (projected to the surface) using the co-registered image. All electrode maps were displayed on a surface plot of the template, using the electrode localization toolbox for MATLAB available at https://github.com/HughWXY/ntools_elec/.

Preprocessing. Data analysis was performed using the FieldTrip toolbox[71], along with custom preprocessing scripts written in MATLAB 2019a (MathWorks). In total, 66 electrodes from all participants were removed due to faulty recordings. The analyses described are at the electrode level. Large spikes exceeding four quartiles above and below the median were removed and replacement samples were imputed using cubic interpolation. We then re-referenced the data to account for shared signals across all channels using either the common average referencing method[71,72] or an independent component analysis-based method[73] (based on the participant's noise profile). High-frequency broadband power frequency provided evidence for a high positive correlation between local neural firing rates and high gamma activity[74].

Broadband power was estimated using six-cycle wavelets to compute the power of the 70–200 Hz band, excluding 60, 120 and 180 Hz line noise. Power was further smoothed with a Hamming window with a kernel size of 50 ms. To preserve the temporal structure of the signal, we used zero-phase symmetric filters. The estimate of the broadband power using wavelets and symmetric filters, by construction, induces some temporal uncertainty, given that information over tens of milliseconds is combined. The amount of temporal uncertainty, however, is small relative to the differences between pre-onset and post-onset effects reported in the paper. First, as the wavelet computation was done using six cycles and the lower bound of the gamma band was 70 Hz, the wavelet computation introduces a 6/70-s uncertainty window centered at each time point. Thus, there is a leak from no more than 43 ms of future signal to data points in the preprocessed signal. Second, the smoothing procedure applied to the broadband power introduces a leak of up to 50 ms from the future. Overall, the leak from the future is at a maximum of 93 ms. As recommended by Cheveigné et al.[75], this was empirically verified by examining the preprocessing procedure on an impulse response (showing a leak of up to ~90 ms,; Extended Data Fig. 10).

The procedure is as follows:

Despike

- Remove recordings that deviate more than three times the interquartile range from the mean value of the electrode.
- Interpolate the removed values using cubic interpolation.

Detrend

- Common average referencing/remove independent component analysis components.

Broadband power

- Use six-cycle wavelets to compute the power of the 70–200 Hz band, excluding 60 and 180 Hz.
- Natural log transformation
- *z*-score transformation

Temporal smoothing

- Use a filter to smooth the data with a Hamming window with a kernel size of 50 ms. Apply the filter in forward and reverse directions to maintain the temporal structure, specifically the encoding peak onset (zero phase).

Encoding analysis. In this analysis, a linear model is implemented for each lag for each electrode relative to word onset, and is used to predict the neural signal from word embeddings (Fig. 3b). The calculated values are the correlations between the predicted signal and the held-out actual signal at each lag (separately for each electrode), indicating the linear model's performance. Before fitting the linear models for each time point, we implemented running window averaging across a 200-ms window. We assessed the linear models' performance (model for each lag) in predicting neural responses for held-out data using a tenfold cross-validation procedure. The neural data were randomly split into a training set (that is, 90% of the words) for model training and a testing set (that is, 10% of the words) for model validation. On each fold of this cross-validation procedure, we used ordinary least-squares multiple linear regression to estimate the regression weights from 90% of the words. We then applied those weights to predict the neural responses to the other 10% of the words. The predicted responses for all ten folds were concatenated so a correlation between the predicted signal and actual signal was computed over all the words of the story. This entire procedure was repeated at 161 lags from −2,000 to 2000 ms in 25-ms increments relative to word onset.

Part of the encoding analysis involves the selection of words to include in the analysis. For each analysis, we included the relevant words. Figure 4a includes all the words in the transcription that have a GloVe embedding totaling 4,843 words. Figure 4b–d comprises 2,886 accurately predicted words (796 unique words) and 1,762 inaccurately predicted words (562 unique words). Lastly, Fig. 6b comprises 3,279 words (165 unique words) that have both GloVe and GPT-2 embeddings, to allow for comparison between the two, and at least five repetitions for the average context and shuffle context conditions.

Accuracy split. To model the brain's prediction, the podcast's transcription words were split into two groups. Each word was marked whether it was one of the top-five most probable words in the distribution that GPT-2 predicted given its past context (up to 1,024 previous tokens) or not (Fig. 1). Around 62% of the words included in the top-five predicted words given their context and were classified as correctly predicted using this accuracy measure. The other words were classified as incorrectly predicted (38%). To control possible confounds stemming from the accurately predicted words that are bigger than the inaccurately predicted group, we also report results from classifying the words according to top-one probability in Extended Data Fig. 8. Using the top-one measure for human prediction, we obtained a group of 36% correctly predicted words.

Confidence and surprise measures. We associate pre-onset neural activity with confidence in prediction and post-onset neural activity with surprise (prediction error). Both could be estimated using GPT-2. Given a sequence of words, autoregressive DLMs induce a distribution of probabilities for the next possible word. We used the entropy of this distribution as a measure for the confidence in prediction:[14,22,41]

$$H(X) = \sum_{i=1}^{n} P(x_i) \times \log P(x_i)$$

Where n is the vocabulary size and $P(x_i)$ is the probability (assigned by the model) of the i-th word in the vocabulary.

To estimate the surprise, we used the cross-entropy measure. Cross-entropy is the loss function used to attenuate the autoregressive DLMs weights, given its predictions (that is, the distribution) and the actual word. The lower the probability of the actual word before its onset, the higher the surprise it induces. It is defined by:

$$\text{Cross-entropy}\,(x_{\text{actual}}) = -\log(x_{\text{actual}})$$

While entropy represents the distance of a distribution from the uniform distribution, the cross-entropy describes the distance between the distribution to the 1-hot distribution.

Significance tests. To identify significant electrodes, we used a randomization procedure. At each iteration, we randomized each electrode's signal phase uniform distribution, thus disconnecting the relationship between the words and the brain signal but preserving the autocorrelation in the signal[76]. Then, we performed the entire encoding procedure for each electrode. We repeated this process 5,000 times. After each iteration, the encoding model's maximal value across all 161 lags was retained for each electrode. We then took the maximum value for each permutation across electrodes. This resulted in a distribution of 5,000 values, which was used to determine significance for all electrodes. For each electrode, a P value (Fig. 3c and 6a and Extended Data Figs. 3 and 4) was computed as the percentile of the non-permuted maximum value of the encoding model across all lags from the null distribution of 5,000 maximum values. Performing a significance test using this randomization procedure evaluates the null hypothesis such that there is no systematic relationship between the brain signal and the corresponding word embedding. This procedure yielded a P value for each electrode. To correct for multiple electrodes, we used the FDR[49]. Electrodes with q values less than 0.01 were considered significant.

To test each lag's significance for two different encoding models for the same group of electrodes (Figs. 4a,b and Extended Data Figs. 5, 8 and 9), we used a permutation test. Each electrode has encoding values for two encoding models. We randomly swapped the assignment of the encoding values between the two models. Then we computed the average of the pairwise differences to generate a null distribution at each lag. To account for multiple tests across lags, we adjusted the resulting P values to control the FDR[49]. A threshold was chosen to control the FDR at $q = 0.01$.

To set a threshold above which average encoding values are significant (Fig. 4 and Extended Data Figs. 6 and 7), we used a bootstrapping method[77]. For each bootstrap, a sample matching the subset size was drawn with replacement from the encoding performance values for the subset of electrodes. The mean of each bootstrap sample was computed. This resulted in a bootstrap distribution of 5,000 mean performance values for each lag. The bootstrap distribution was then shifted by the observed value to perform a null hypothesis test[77]. To account for multiple tests across lags, we adjusted the resulting P values to control the FDR[49]. A threshold was chosen to control the FDR at $q = 0.01$.

To statistically assess the pre-onset prediction for neural responses to correctly predicted words (Fig. 5), we completed a permutation test (such as the one described for Fig. 4a,b); however, we were also constrained to lags at which the neural responses were significant on their own (not with respect to the neural response of the inaccurate conditional brain response). The same procedure was implemented for the significant test of post-onset surprise.

Contextual embedding extraction. We extracted contextualized word embeddings from GPT-2 for our analysis. We used the pretrained version of the model implemented in the Hugging Face environment[78]. We first converted the words from the raw transcript (including punctuation and capitalization) to tokens which were either whole words or sub-words. We used a sliding window of 1,024 tokens, moving one token at a time, to extract the embedding for the final word in the sequence. Encoding these tokens into integer labels, we then fed them into the model, and in return, we received the activations at each layer in the network (also known as a hidden state). GPT-2 has 48 layers, but we focused only on the final one, before the classification layer. Finally, the token of interest was the final word of the sequence, yet we used the second-to-last token as the hidden state for the last word because it was the same activation embedding that was used to predict that word. With embeddings for each word in the raw transcript, we aligned this list with our spoken-word transcript that did not include punctuation, thus retaining only full words.

Decoding analysis. The input neural data were averaged in ten 62.5-ms bins spanning 625 ms for each lag. Each bin consisted of 32 data points (the neural recording sampling rate was 512 Hz). The neural network decoder (Appendix I) was trained to predict a word's embedding from the neural signal at a specific lag. The data were split into five nonoverlapping temporal folds and used in a cross-validation procedure. Each fold consisted of a mean of 717.04 training words (s.d. = 1.32). Three folds were used for training the decoder (training set), one fold was used for early stopping (development set) and one fold was used to assess model generalization (test set). The neural net was optimized to minimize the MSE when predicting the embedding. The decoding performance was evaluated using a classification task assessing how well the decoder can predict the word label from the neural signal. We used the ROC-AUC.

To ensure that the decoding ability was not affected by the electrode selection procedure, we used the training and validation folds (80% of the data) to choose the electrodes for each model. We used the same significance test as the one used to locate GloVe-based significant encodings. This procedure yielded a different number of electrodes ranging from 114 to 132.

To calculate the ROC-AUC, we computed the cosine distance between each of the predicted embeddings and the embeddings of all instances of each unique word label. The distances were averaged across unique word labels, yielding one score for each word label (that is, logit). We used a Softmax transformation on these scores (logits). For each label (classifier), we used the logits and the information of whether the instance matched the label to compute a ROC-AUC for the label. We plotted the weighted ROC-AUC according to the word's frequency in the test set. To obtain reliable ROC-AUC scores, we chose words with at least five repetitions in the training set (69% of the words in the transcript).

To improve the performance of the decoder, we implemented an ensemble of models. We independently trained ten decoders with randomized weight initializations and randomized the batch order. This procedure generated ten predicted embeddings. Thus, for each predicted embedding, we repeated the distance calculation from each word label ten times. These ten values were averaged and later used for ROC-AUC.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The dataset will become available 6 months after paper publication. Pending anonymization process.

Code availability

All the scripts for analyses can be found at https://github.com/orgs/hassonlab/repositories/.

References

69. Yuan, J. & Liberman, M. Speaker identification on the SCOTUS corpus. *J. Acoust. Soc. Am.* **123**, 3878–3878 (2008).
70. Yang, A. I. et al. Localization of dense intracranial electrode arrays using magnetic resonance imaging. *NeuroImage* **63**, 157–165 (2012).
71. Oostenveld, R., Fries, P., Maris, E. & Schoffelen, J.-M. FieldTrip: open-source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* **2011**, 156869 (2011).
72. Lachaux, J. P., Rudrauf, D. & Kahane, P. Intracranial EEG and human brain mapping. *J. Physiol. Paris* **97**, 613–628 (2003).
73. Michelmann, S. et al. Data-driven re-referencing of intracranial EEG based on independent component analysis. *J. Neurosci. Methods* **307**, 125–137 (2018).
74. Jia, X., Tanabe, S. & Kohn, A. Gamma and the coordination of spiking activity in early visual cortex. *Neuron* **77**, 762–774 (2013).
75. Cheveigné, Ade, de Cheveigné, A. & Nelken, I. Filters: when, why and how (not) to use them. *Neuron* **102**, 280–293 (2019).
76. Gerber, E. M. PhaseShuffle (https://www.mathworks.com/matlabcentral/fileexchange/71738-phaseshuffle), MATLAB Central File Exchange (2021).
77. Hall, P. & Wilson, S. R. Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**, 757–762 (1991).

78. Tunstall, L., von Werra, L. & Wolf, T. *Natural Language Processing With Transformers: Building Language Applications With Hugging Face* (O'Reilly, 2022).

Acknowledgements
We thank A. Goldberg, R. Goldstein, S. Michelmann, M. Meshulam, M. Kumar, M. Slaney and A. Huth for technical and conceptual assistance that motivated and informed this paper's writing. This work was supported by the National Institutes of Health under award numbers DP1HD091948 (to A.G., Z.Z., A.P., B.A., G.C., A.R., C.K., F.L., A.F. and U.H.), R01MH112566 (to S.A.N.) and R01NS109367-01 to A.F., Finding A Cure for Epilepsy and Seizures (FACES) and Schmidt Futures Foundation DataX Fund.

Author contributions
A.G. devised the project, performed experimental design and data analysis, and wrote the paper; Z.Z., E.B. and M.S. performed data analysis; A.P. designed behavioral test and performed data analysis; B.A. performed data collection and data analysis; S.A.N. critically revised the article; A.F., D.E., A.C., A.J. and H.G. performed data analysis; G.C., A.R. and C.K. conducted data collection; C.C. performed data collection and data analysis; L.F. performed data collection; W.D. performed neurosurgery; D.F. and P.D. conducted data collection; L.M. devised the project; R.R. critically revised the article; S.D. conducted data collection and provided comments on the manuscript; A.F. devised the project, performed data collection and critically revised the article; L.H. critically revised the article; O.L. conducted experiment design; A.H., M.B. and Y.M. devised the project; K.A.N. and O.D. devised the project and performed critical revision of the article; and U.H. devised the project and wrote the paper.

Competing interests
The authors declare no competing interests.

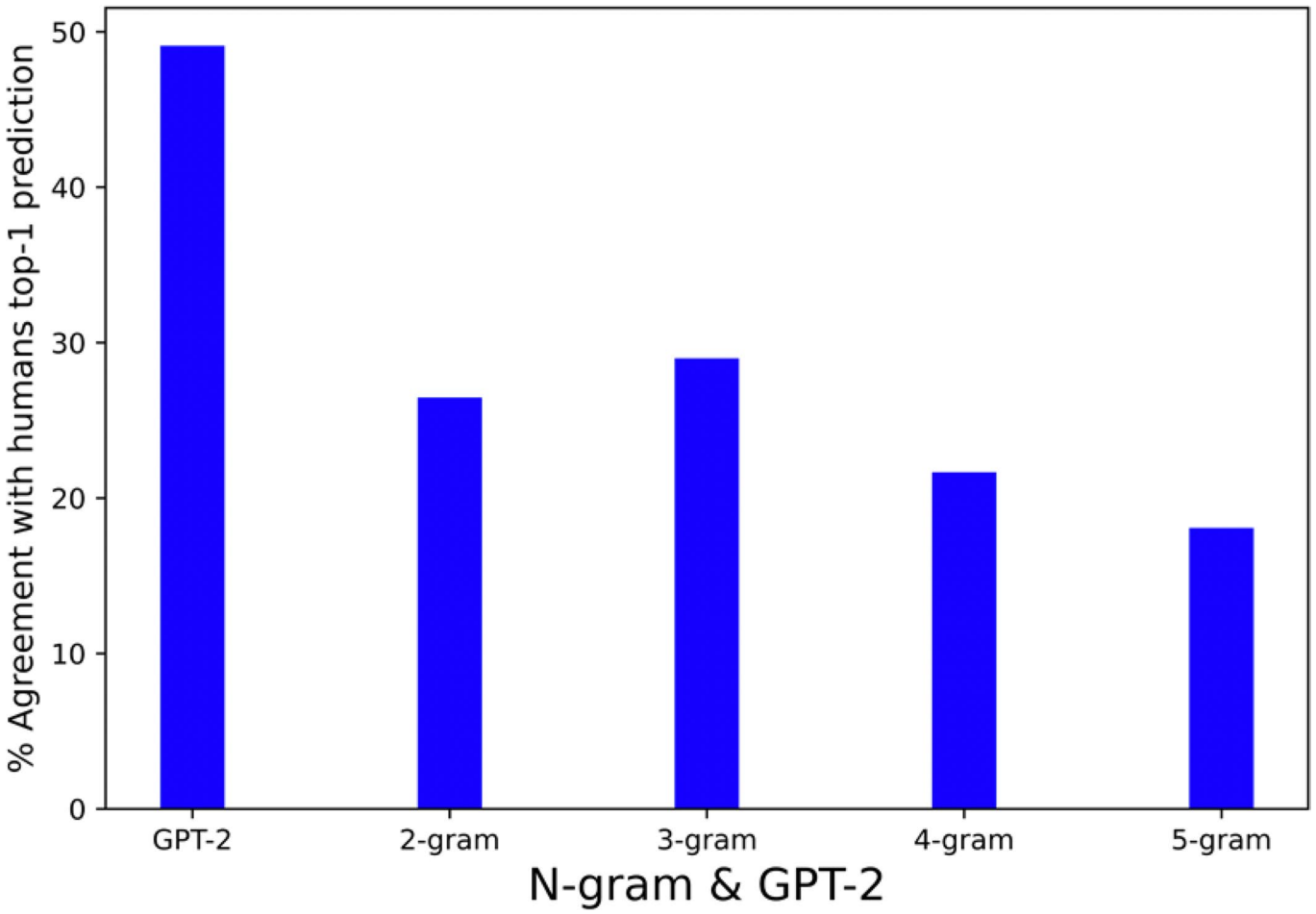
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41593-022-01026-4.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-022-01026-4.

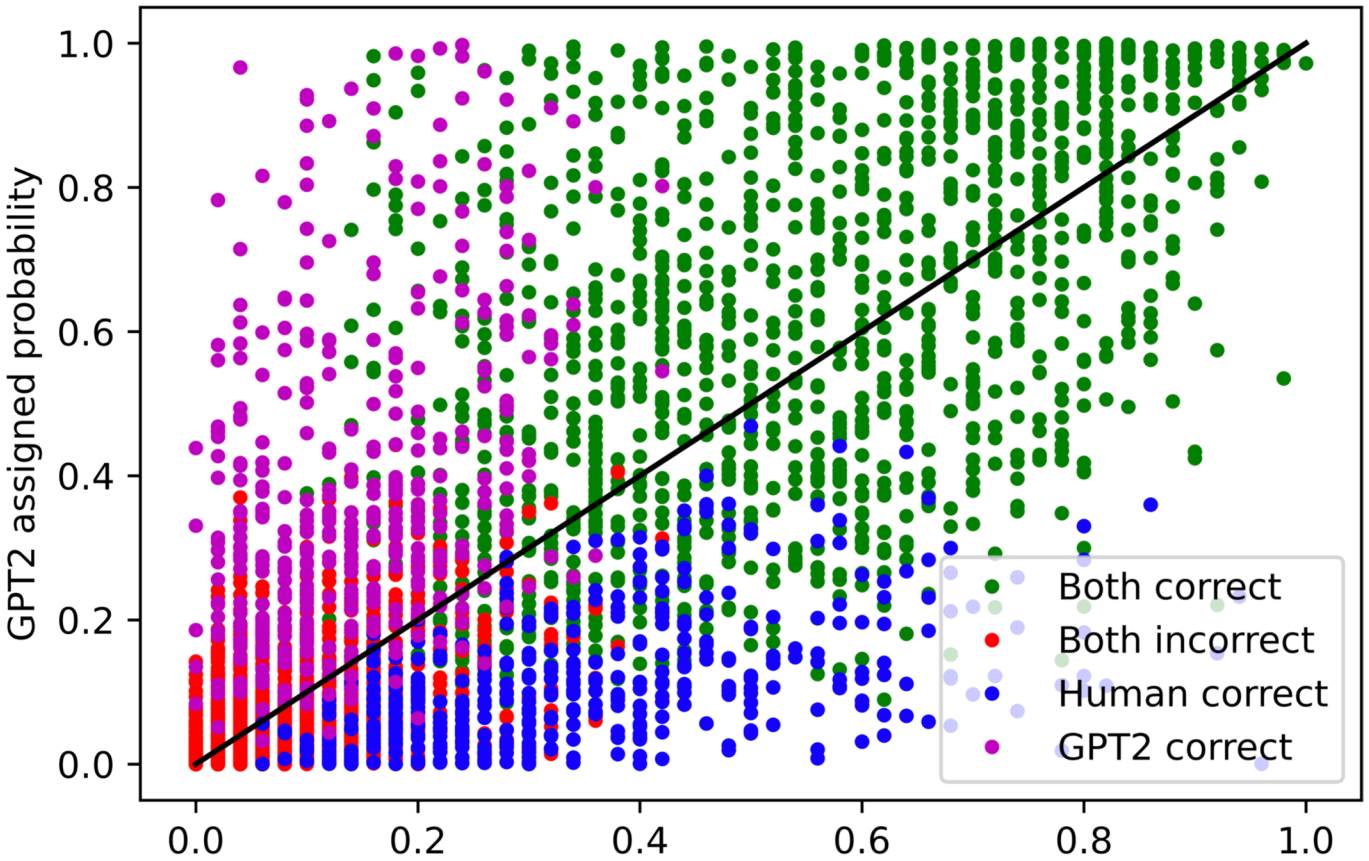
Correspondence and requests for materials should be addressed to Ariel Goldstein.

Peer review information *Nature Neuroscience* thanks Alona Fyshe and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

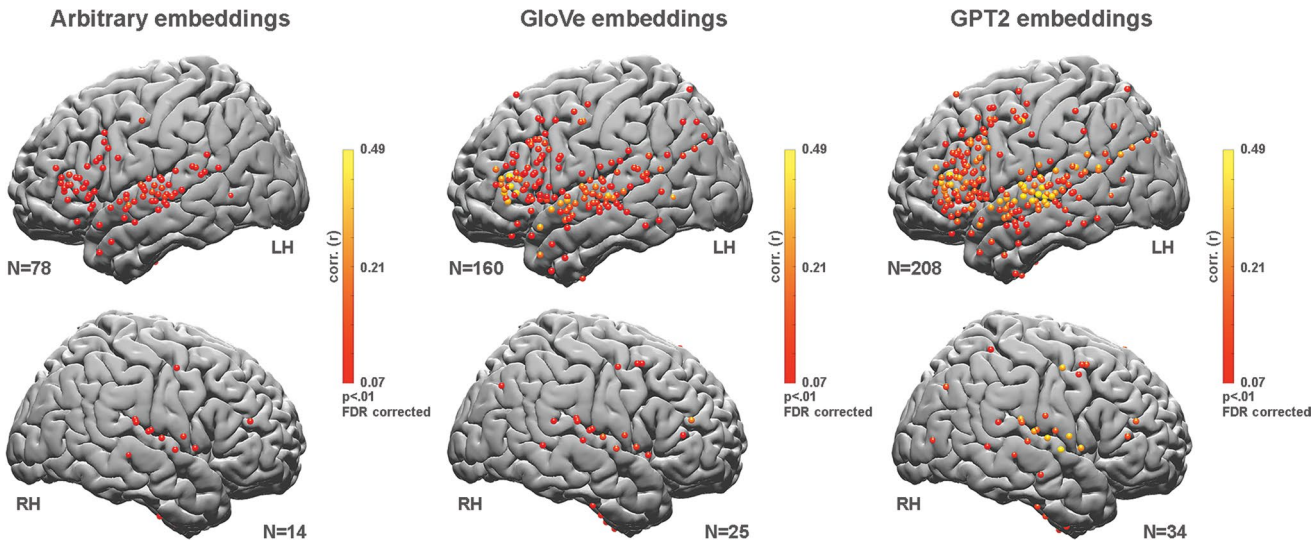
Reprints and permissions information is available at www.nature.com/reprints.



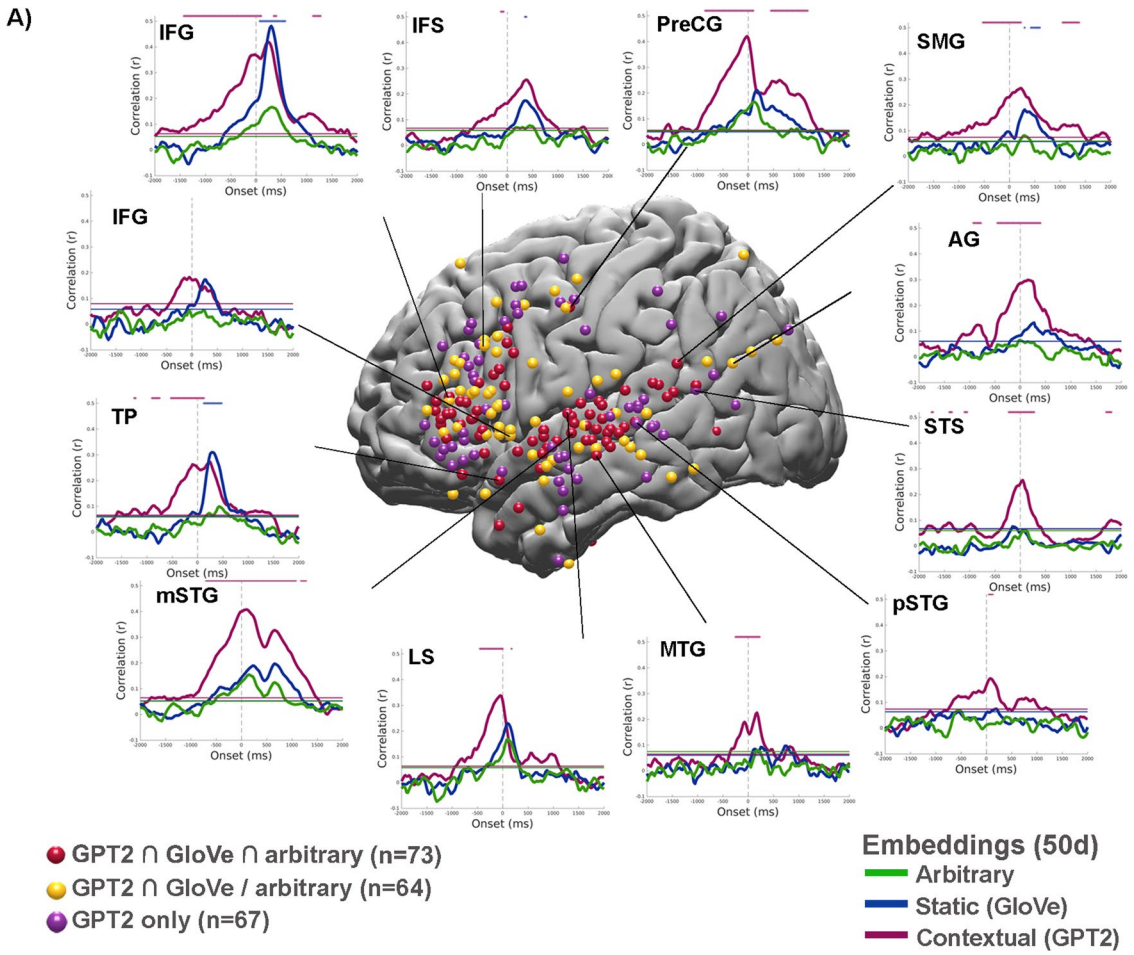
Extended Data Fig. 1 | Figure S1. Comparing agreement with human prediction between most probable predictions based on n-grams or GPT-2. The plots show higher agreement between human predictions and GPT-2's top-1 predictions than all the n-gram model predictions we trained.



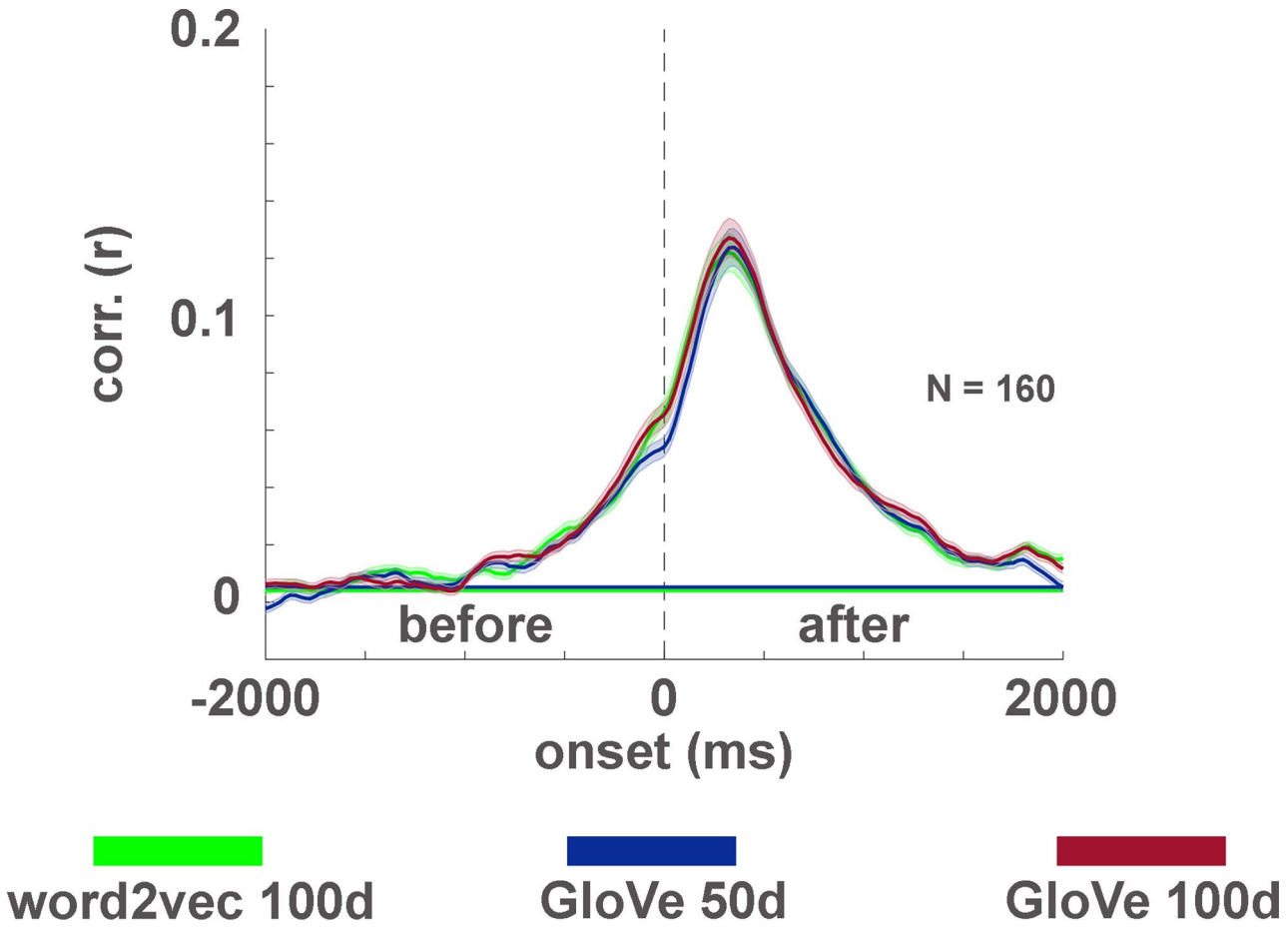
Extended Data Fig. 2 | Figure S2. Comparing GPT-2 predictions and human predictions. Coloring the scatter plot according to GPT-2/human accuracy. GPT-2 and humans jointly predicted correctly 27.6% of the words (green). GPT-2 and humans jointly predicted incorrectly and disagreed on the next word for 48.8% of the words (red). GPT-2 and humans jointly predicted incorrectly and agreed on the next word for 5.9% of the words (black) 9.2% of the words humans predicted correctly were not correctly predicted by GPT-2 (blue). 8.4% of the words correctly predicted by GPT-2 were not correctly predicted by humans.



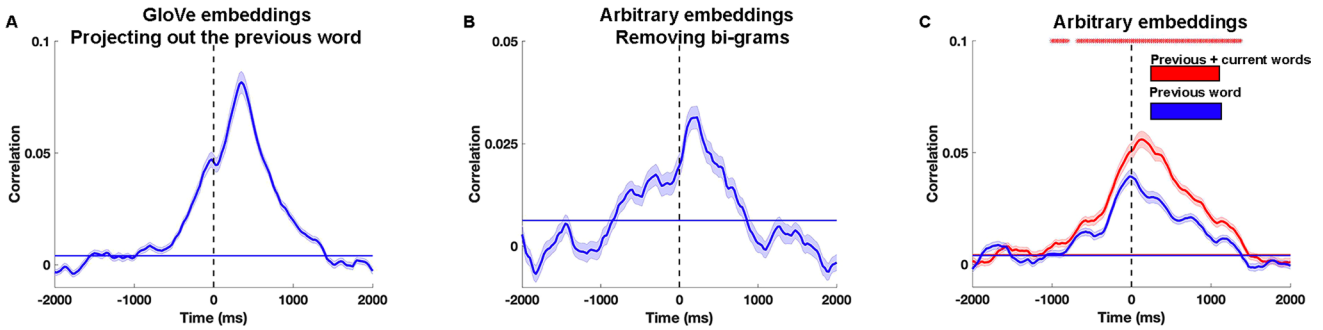
Extended Data Fig. 3 | Figure S3. Left and right hemisphere encoding results show an advantage for contextual (GPT-2) embeddings over static (GloVe) and arbitrary embeddings. Right Hemisphere maps for correlation between. **A**. Predicted and actual word responses for the arbitrary embeddings (nonparametric permutation test; $q < .01$, FDR corrected). **B**. Correlation between predicted and actual word responses for the static (GloVe) embeddings. **C**. Correlation between predicted and actual word responses for the contextual (GPT-2) embeddings. Using contextual embeddings significantly improved the encoding model's ability to predict the neural signals for unseen words across many electrodes. Given that we had fewer electrodes in the right hemisphere relative to the left hemisphere, this study is not set up to test differences in language lateralization across hemispheres.



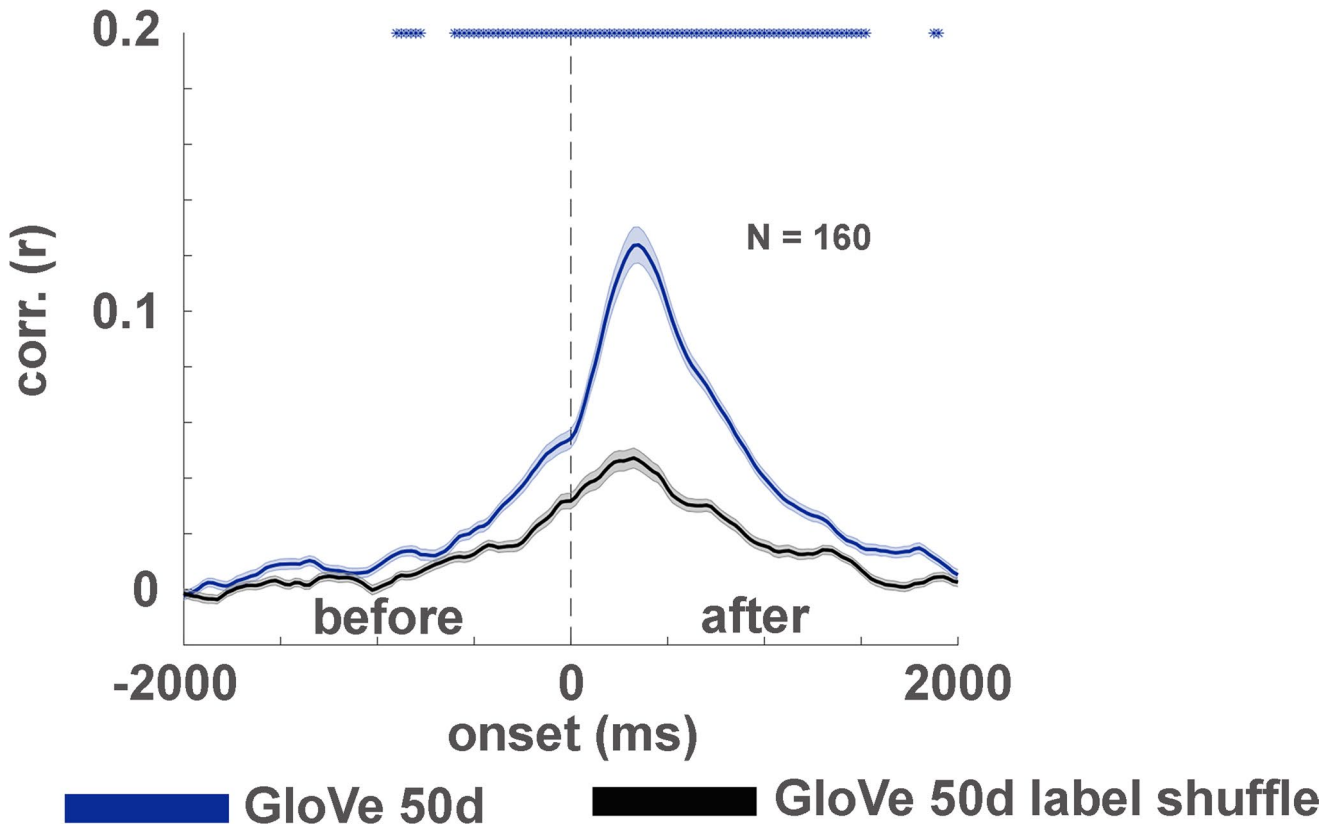
Extended Data Fig. 4 | Figure S4. Contextual embedding significantly improves the modeling of neural signals. Map of the electrodes in the left hemisphere with significant encoding for 1) all three types of embeddings (GPT-2 ∩ GloVe ∩ arbitrary, red); 2) for static and contextual embeddings (GPT-2 ∩ GloVe, but not arbitrary, yellow); 3) and contextual only (GPT-2, purple) embeddings. Note the three groups do not overlap. A sampling of encoding performance for selected individual electrodes across different brain areas: inferior frontal gyrus (IFG), temporal pole (TP), middle superior central gyrus (mSTG), superior temporal sulcus (STS), lateral sulcus (LS), middle temporal gyrus (MTG), posterior superior temporal gyrus (pSTG), angular gyrus (AG), post central gyrus (postCG), precentral gyrus (PreCG), and middle frontal sulcus (MFS). (Green - encoding for the arbitrary embeddings, blue - encoding for static (GloVe) embeddings; purple - encoding for contextual (GPT-2) embeddings).



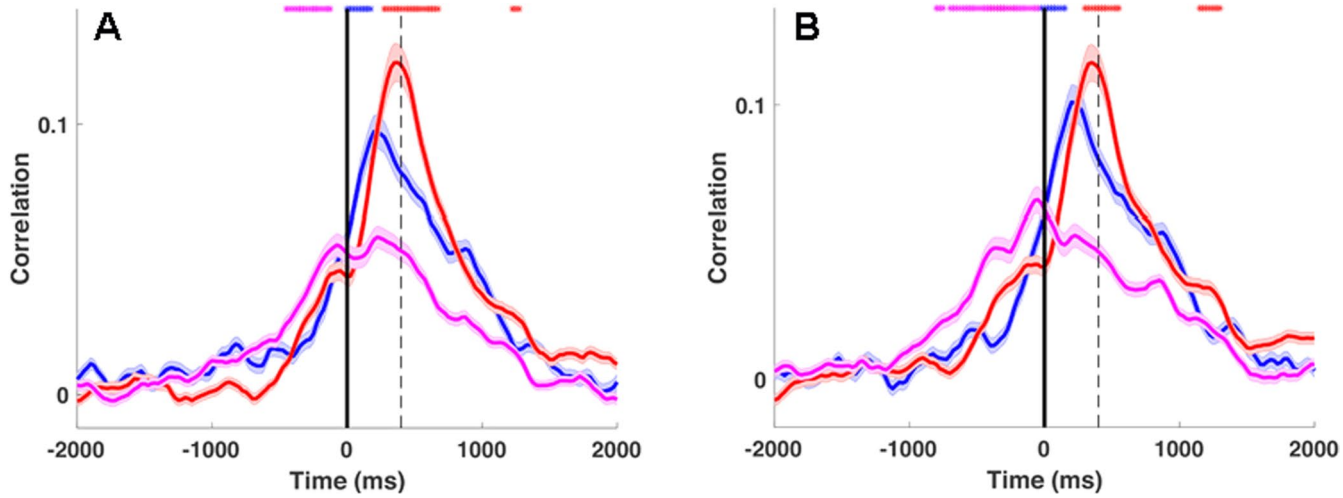
Extended Data Fig. 5 | Figure S5. Comparison of GloVe and word2vec-based static embeddings. The encoding procedure was repeated for two additional static embeddings using the electrodes that were found significant for GloVe-50 encoding on the left hemisphere (Fig. 3B). Each line indicates the encoding model performance averaged across electrodes for a given type of static embedding at lags from -2000 to 2000 ms relative to word onset. The error bars indicate the standard error of the mean across the electrodes at each lag. 100-dimensional word2vec and GloVe embeddings resulted in similar encoding results to the initial 50-dimensional GloVe embeddings. This suggests that results obtained with static embeddings are robust to the specific type of static embeddings used.



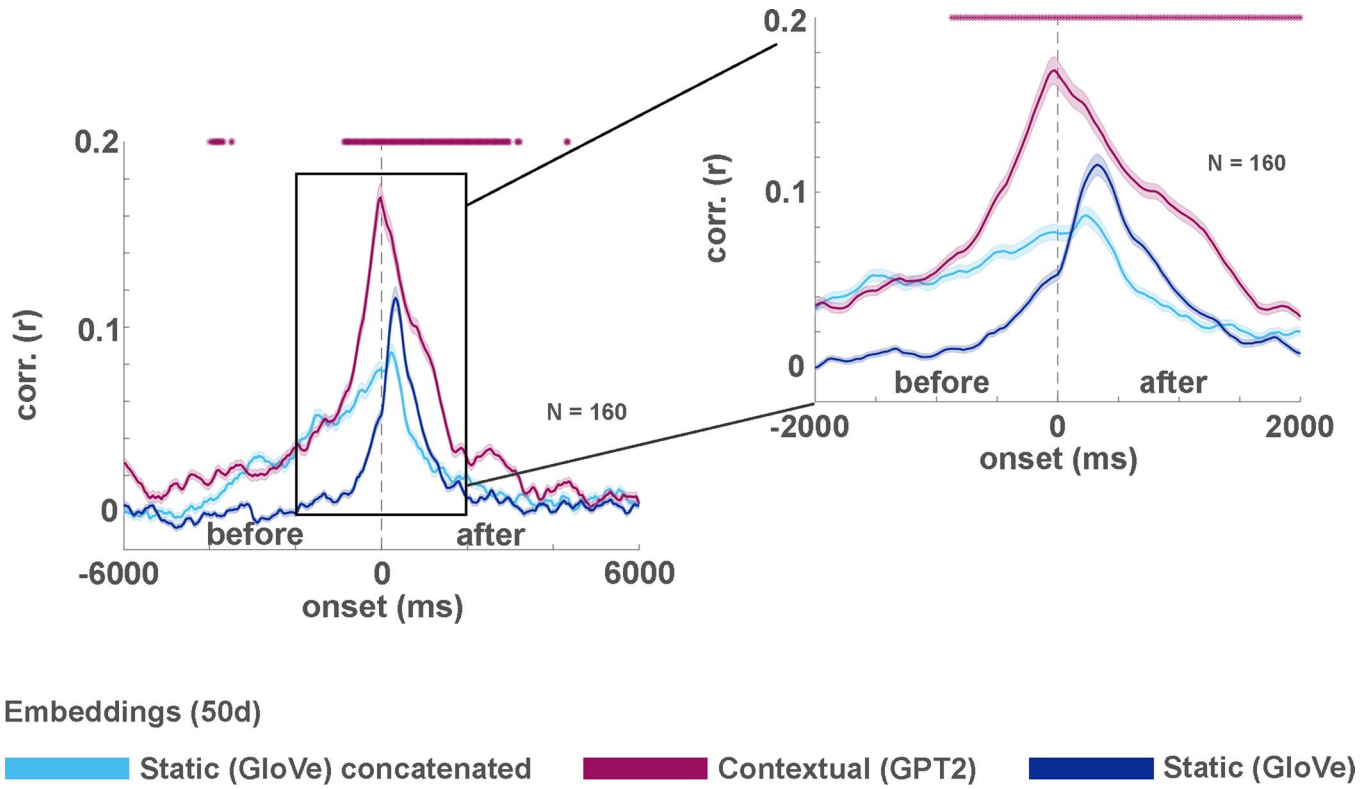
Extended Data Fig. 6 | Figure S6. Controlling for correlations among adjacent GloVe embeddings. To ensure that the signal predicted before word-onset is not a result of a correlation among adjacent GloVe embeddings we ran the following additional control analyses: **A**. We projected (by inner product) and then subtracted the GloVe embedding of the previous word from each word and re-ran the encoding analysis. The analysis demonstrates that the significant encoding before word onset holds even after removing local contextual dependencies in the GloVe embedding of consecutive words. The error bars indicate the standard error of the encoding models across electrodes. The horizontal line indicates the significance threshold calculated using a permutation test and FDR corrected for multiple comparisons ($q < .01$). **B**. We trained an encoding model using arbitrary embeddings on our dataset after removing all bi-grams that repeated more than once. The encoding before word onset remained significant after the removal of the bi-grams. The error bars indicate the standard error of the encoding models across electrodes. The horizontal line indicates the significance threshold calculated using a permutation test and FDR corrected for multiple comparisons ($q < .01$). **C**. We compared an encoding model based on arbitrary embeddings using the previous word embedding (blue line), to an encoding model where we concatenated previous and current word embeddings (red line). The error bars indicate the standard error of the encoding models across electrodes. Red asterisks mark significant differences using a permutation test and FDR correction ($q < .01$). The significant difference between these two models before word onset is another evidence that there is predictive information in the neural activity as to the upcoming word, above and beyond the contextual information embedded in the previous word. The horizontal line indicates the significance threshold calculated using permutation test and FDR corrected for multiple comparisons ($q < .01$).



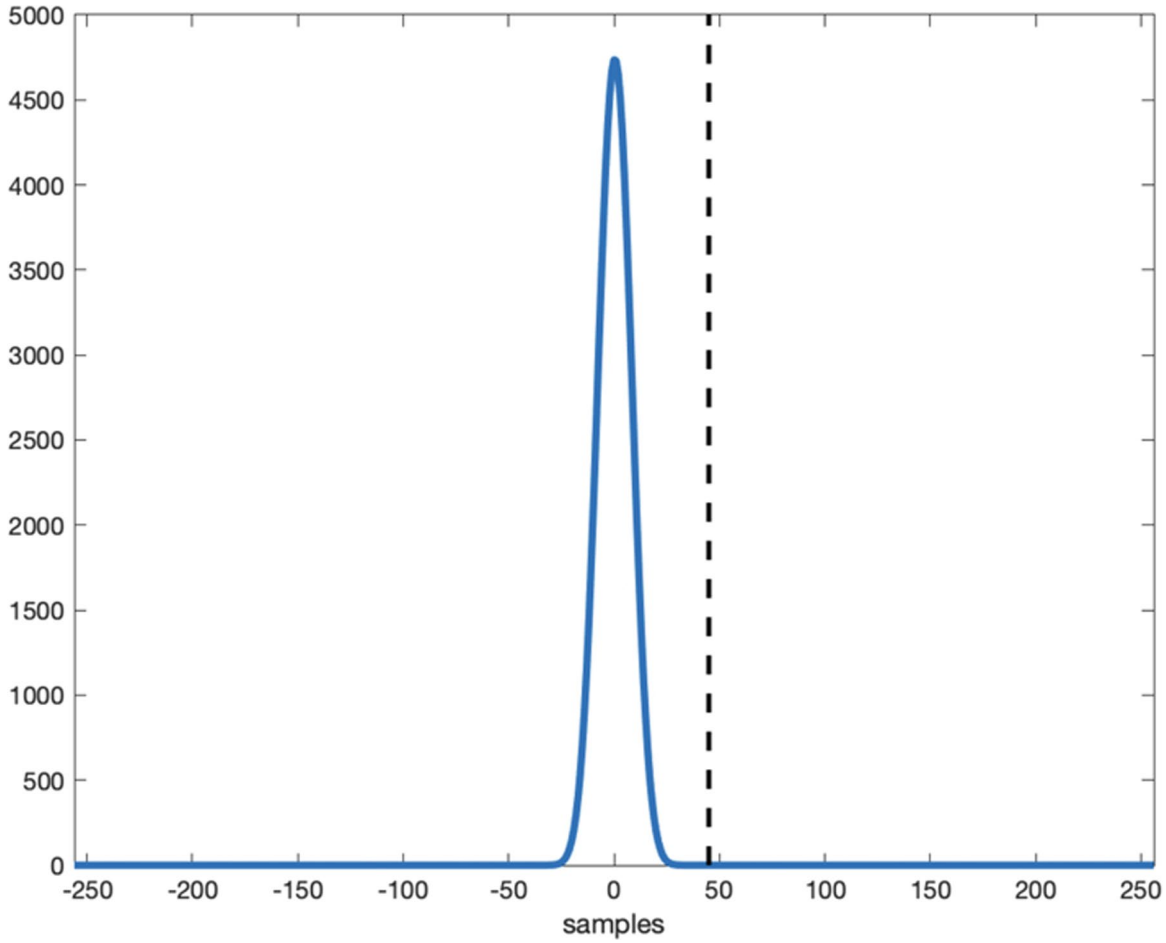
Extended Data Fig. 7 | Figure S7. GloVe's space embedding attributes. It can be argued that GloVe based encoding outperforms arbitrary-based encoding due to a general property of the space that GloVe embeddings induce (for example, they are closer / further away from each other). To control for this possible confound, we consistently mismatched the labels of the embeddings of GloVe and used the mismatched version for encoding. This means that each unique word was consistently matched with a specific vector that is actually an embedding of a different label (for example, matching each instance of the word 'David' with the embedding of the word 'court'). This manipulation uses the same embedding space that GloVe uses and also induces a consistent mapping of words to embeddings (as in the arbitrary-based encoding). The matched GloVe (blue) outperformed the mismatched GloVe (black), supporting the claim that GloVe embedding carries information about word statistics that is useful for predicting the brain signal.. The error bars indicate the standard error of the encoding models across electrodes.



Extended Data Fig. 8 | Figure S8. Encoding for correct / incorrect predictions. This is a variation of Fig. 4B where: A. We classify words as correctly predicted if they are the most predictable words by humans' ratings. The error bars indicate the standard error of the encoding models across electrodes. **B**. We classify words as correctly predicted if they are the most predictable by GPT-2 (instead of top-5). The error bars indicate the standard error of the encoding models across electrodes.



Extended Data Fig. 9 | Figure S9. Comparison of GPT-2 and concatenation of static embeddings. The increased performance of GPT-2 based contextual embeddings encoding may be attributed to the fact that it consists of information about the previous words' identity. To examine this possibility, we concatenated the GloVe embeddings of the 10 previous words and current word, and reduced their dimensionality to 50 features. GPT-2 based encoding outperformed mere concatenation before word onset, suggesting that GPT-2's ability to compress the contextual information improves the ability to model the neural signals before word onset. The error bars indicate the standard error of the encoding models across electrodes.



Extended Data Fig. 10 | Figure S10. Preprocessing procedure applied to an impulse response. The plot demonstrates the temporal uncertainty introduced by the preprocessing procedure (especially by the wavelet and smoothing procedures). At sample 45 after onset (dashed line) the value is back to zero, considering the 512 HZ sampling rate this means that the leak from the future is bounded by 93 ms.

nature research

Corresponding author(s): Ariel Goldstein

Last updated by author(s): 2021/12/19

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Behavioral data were collected Mechanical Turk.
Data analysis	Data were preprocessed using Matlab 2019b and The Fieldtrip toolbox. Data was analyzed using python packages are specified here https://github.com/hassonlab/247-main/blob/main/env.yml. Brain plots were done using toolbox for MATLAB available at (https://github.com/HughWXY/ntools_elec). All scripts for analyses are available at:All the scripts for analyses can be found at: https://github.com/orgs/hassonlab/repositories

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We will make the data available 6 month after paper publication. Pending anonymization process.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to pre-determine sample sizes but our final sample sizes are similar to those reported in previous publications. For exmaple, Michelmann et. al. (2021). Moment-by-moment tracking of naturalistic learning and its underlying hippocampo-cortical interactions. Nature communications.
Data exclusions	One patient was removed according pre-established exclusion criteria of excessive epileptic activity .
Replication	Pre-existing behavioral data that were analyzed in this study were replicated in a new sample. The neural results are replicated at the single patient level (which can be considered as replication). No failed replications have be found.
Randomization	Randomization was not applicable as patients freely hear a strory when their neural activity is measured. Behavior pilot for predictability scores also did not require randomization
Blinding	Blinding was not relevant in this study, because participants and experimenter do not directly interact in online-experiments. Patients were not assigned to conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☐	☒ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Population characteristics Nine epilleptic patients who volunteered to participate in the study (5 female; 20–48 years old); For the behavioral experiment we used Mechanical Turk, no details about age and gedner were collected.
Recruitment	Behavioral participants were recruited via Amazon Mechanical Turk. They were payed 10$ for their participation. Thier age and gender were not collected. The quality of behavior is assessed an no self selection criteria was found relevant. Patients were recruited via the Comprehensive Epilepsy Center of the New York University School of Medicine. Patients are asked whehte they want to listen to a podcast during their time in the hospital and are consenting that the neural data will be used for analyses. No relevent self selection is known with this regard.
Ethics oversight	Princeton University and New York University School of Medicine respective Institutional Review Boards approved the studies. Board/committee (Princeton university IRB) that approved the protocol: Daniel Notterman (Chair), Calvin Chin,Elizabeth Davis,Patricia Fernandez-Kelly,Michael Graziano, Kyle Jamieson, Jacqueline Kariithi, Vikki Lovvoll, Brandt McCabe, Chad Pettengill, Pascale Poussart, Naila Rahman, Rusty Reeves, Carrie Siegler, Jordan Taylor, Rory Truex

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	n.a.
Study protocol	n.a.
Data collection	n.a.
Outcomes	n.a.