eLife

Reviewed Preprint
v1 • April 22, 2026
Not revised

For correspondence:
sahel.azizpour@donders.ru.nl

† These authors contributed equally to this work

Funding: See page 22

Reviewing editor: Huan Luo, Peking University, China

Reassessing prediction in the brain: Pre-onset neural encoding during natural listening does not reflect pre-activation

Sahel Azizpour[1,2], **Britta U Westner**[1,3,†], **Jakub Szewczyk**[4,†], **Umut Güçlü**[1,2], **Linda Geerligs**[1,2]

[1]Donders Institute for Brain, Cognition, and Behaviour, Nijmegen, Netherlands • [2]Radboud University, Nijmegen, Netherlands • [3]Radboud University Medical Center, Nijmegen, Netherlands • [4]Jagiellonian University, Kraków, Poland

eLife Assessment

This study presents **valuable** findings by reanalyzing previously published MEG and ECoG datasets to challenge the predictive nature of pre-onset neural encoding effects. The evidence supporting the central conclusions remains **incomplete**, as additional details of the analyses are needed and alternative interpretations, such as the possibility that pre-onset predictive and sensory-evoked responses rely on distinct neural representations, have not been sufficiently addressed. The work may be of interest to researchers in language processing, predictive coding, and related fields.

https://doi.org/10.7554/eLife.110085.1.sa4

Abstract

Predictive processing theories propose that the brain continuously anticipates upcoming input. However, direct neural evidence for predictive pre-activation during natural language comprehension remains limited and debated. Previous studies using large language model (LLM)-based encoding models with fMRI and ECoG have reported pre-onset signals that appear to encode upcoming words, but these effects may instead reflect dependencies in the stimulus or autocorrelations in neural activity.

Here, we re-examined this question by aligning LLM-derived word embeddings with neural activity recorded during naturalistic listening using magnetoencephalography (MEG) and electrocorticography (ECoG). We replicated pre-onset encoding effects previously observed in ECoG across both modalities, and found that they persist even after controlling for stimulus correlations. Crucially, temporal generalization analyses revealed no stable overlap between pre- and post-onset representations, indicating that pre-onset activity does not reflect pre-activation of the next word. Consistent with this, long-range predictive effects previously reported in fMRI did not replicate in our higher-temporal-resolution data.

While we found no evidence for predictive pre-activation, we observed clear signatures of postdiction, with neural activity reflecting persistent encoding of prior words. These results suggest that reported apparent predictive signals do not reflect pre-activation of upcoming input. They call for caution in interpreting LLM-based encoding models and highlight the need for a more nuanced understanding of what constitutes “prediction” in language comprehension.

Introduction

Predictive processing is a foundational hypothesis in neuroscience, that proposes thatthe brain is continuously engaged in making predictions about upcoming stimuli (Clark, 2013; Friston, 2010). These predictions are shaped by the context in which the input is presented and are informed by the individual’s previous experiences with similar input sequences (De Lange et al., 2018). Language processing is one domain where predictions are thought to play a crucial role, particularly in facilitating comprehension by pre-activating linguistic representations (Federmeier, 2007; Kutas et al., 2011). Various approaches have been used to investigate how the availability of predictive information influences lexico-semantic processing. At the behavioral level, studies have shown that highly predictable words result in shorter reaction times and reading times (Smith and Levy, 2013; Traxler and Foss, 2000). At the neural level, numerous studies have demonstrated thatthe ERP response is modulated by the predictability of a word within a contextually constraining sentence (DeLong et al., 2005; Kutas et al., 2011). Compelling evidence from studies of natural conversations also shows that the absence of temporal gaps between exchanges requires listeners to predict the upcoming word and content with high accuracy (De Ruiter et al., 2006; Magyari and De Ruiter, 2012). Over the past five decades, a wealth of research has strongly suggested that language comprehension is inherently predictive (Kuperberg and Jaeger, 2016).

Despite broad agreement that language comprehension involves prediction, there is less agreement on what form these predictionstake and how they influence processing. Some accounts emphasize pre-activation, in which features of upcoming words (e.g., semantic, syntactic, or phonological) are actively represented before they appear (Federmeier, 2007; Antonello and Huth, 2024). Others argue that any facilitation in word processing necessarily implies some form of prediction (Kuperberg and Jaeger, 2016). However, such predictions may reflect a representation of the word already integrated with the context, rather than pre-activation of its features. Empirically disentangling these accounts remains challenging, as traditional neural measures such as the N400 can reflect both pre-activation of word-related features and post-hoc processing of input (Kutas and Federmeier, 2011). This ongoing debate highlights the need for methods that can directly test pre-activation accounts of prediction by capturing the representational content of prediction during continuous language comprehension.

Evidence for predictive pre-activation was provided by (Wang et al., 2020, 2024). They used representational similarity analysis (RSA) to show that EEG patterns just before a highly predictable target word could reveal whether the noun was animate. However, their approach required a one-second gap before the target. While this interval presumably improved the ability to decode the noun’s animacy, it may have encouraged participants to adopt strategic, unnatural predictive processes that are less likely in natural comprehension scenarios.

Recent advancements have created new opportunities to study predictive processing in naturalistic settings. High-quality datasets from MEG and ECoG now allow researchers to examine prediction across tens of thousands of words, such as in participants listening to entire audiobooks. These large-scale datasets offer greater statistical power to detect subtle predictive representations while exposing participants to richer linguistic contexts with higher ecological validity than traditional experimental designs (Hamilton and Huth, 2020; Willems et al., 2020). Crucially, in such conditions, a word’s predictability is shaped by long-range discourse rather than just the preceding word or sentence.

A potential challenge with using large datasets has been quantifying word predictability, as traditional cloze probability tests are labor-intensive and infeasible for datasets with tens of thousands of words. Fortunately, recent advances in LLMs overcome this limitation. Trained on vast text corpora, LLMs capture nuanced linguistic representations and can serve as proper linguistic theories (Baroni, 2022), and by extension, proxies for the brain’s language processes. Indeed, across various neuroimaging modalities, such as fMRI (Caucheteux and King, 2022; Caucheteux et al., 2023; Schrimpf et al., 2021), MEG (Caucheteux and King, 2022; Toneva et al., 2022), and ECoG (Goldstein et al., 2022b), it has been demonstrated that LLM-generated word embeddings can be linearly mapped onto brain activity, explaining a substantial portion of the variance in neural responses to language.

Leveraging the power of LLMs, only a few studies have examined neural signatures of predictive pre-activation during naturalistic narrative listening. The first, by Goldstein et al. (2022b), examined whether neural signals preceding word onset contain predictive information about upcoming words. Using high-precision ECoG recordings during natural speech listening, they found that the brain represents the upcoming word upto two seconds before the word onset. This finding has since been replicated and extended to examine various properties of predictive processing in natural language comprehension (Goldstein et al., 2022a; Zada et al., 2024).

A second line of evidence comes from Caucheteux et al. (2023), who demonstrated that fMRI responses during natural story listening were better predicted when models incorporated embeddings of both current and future words. This resultwas interpreted as reflecting the brain's ability to represent multiple upcoming words, providing potential evidence for long-range predictive processing.

Both studies suggest that the brain actively predicts upcoming linguistic input before it is perceived. However, their interpretations remain debated. Schönmann et al. (2025) cautioned that the pre-onset signals reported by Goldstein et al. (2022b) may not reflect genuine pre-activation of lexical representations. They showed that similar pre-onset patterns can emerge when models predict the acoustics of continuous speech rather than neural activity. This finding indicates that apparent pre-onset encoding may partly arise from correlations inherent in the stimulus, such as dependencies between adjacent word embeddings or co-occurring acoustic features, rather than from predictive neural processes.

Similarly, Caucheteux et al. (2023) relied on fMRI data with inherently low temporal resolution, averaging word embeddings across 1.5 s repetition times (TRs). This temporal smoothing could obscure word-level dynamics and introduce autocorrelations that inflate apparent prediction effects.

Here, we revisit these claims using a multimodal approach to test (1) whether previously reported pre-activation signals can be replicated across different datasets and modalities, and (2) whether such effects truly reflect predictive pre-activation rather than other sources of temporal correlation.

To this end, we mapped LLM-derived word embeddings onto neural activity from two complementary datasets: (1) broadband MEG recordings collected while participants listened to approximately 10 hours of narrative speech, and (2) high-gamma ECoG signals recorded as participants listened to 30 minutes of a podcast (the same dataset as reported in Goldstein et al. (2022b)).

Across both modalities, we replicated pre-onset encoding effects reported in prior work (Goldstein et al., 2022b) and confirmed that these effects persist even after controlling for statistical regularities in the stimulus, such as correlations among adjacent word embeddings. However, when applying Temporal Generalization (TG) analysis, we found no evidence for stable representational overlap between pre- and post-onset time points. This suggests that pre-onset encoding does not reflect the active pre-activation of specific upcoming word representations.

In line with this finding, when adapting the fMRI analysis of Caucheteux et al. (2023) to our MEG and ECoG data, we found no improvement in neural encoding when including embeddings of future words. Instead, we observed robust evidence for postdiction: incorporating embeddings of preceding words improved model performance, indicating that prior word representations persist after word offset.

Our results suggest that previously reported predictive signals do not reflect genuine pre-activation of upcoming input. Instead, they call for caution when using LLM-based encoding models to study predictive processing and highlight the need fora more nuanced discussion ofwhat constitutes “prediction” in language comprehension, and whether prediction in this context truly entails pre-activation.

Results

The encoder processed the words from the narrative and predicted corresponding brain signals within a time window around word onset (Fig. 1). It was trained to learn a linear mapping between the LLM's activations (word embeddings) and the MEG broadband signal (0.1-40 Hz) and the ECoG high gamma (70-200 Hz) broadband power recorded at each time point within this window. To quantify how well the LLM activations capture information in the brain activity, we computed the correlation between the predicted and actual brain responses. The resulting correlation at each time point, termed the 'brain score' $\mathscr{R}$, indicates the extent to which the word's features are represented in the brain signal.

Encoding models were trained separately for each participant, time point, and MEG source or ECoG electrode. In MEG, nearly all regions exhibited positive correlations between predicted and observed brain responses, with the strongest effects in the left temporal gyri and inferior frontal regions (Fig. 2.a). Similarly, all ECoG electrodes showed positive encoding of word embeddings, with maximum correlation values larger than MEG. The strongest effects were observed in the left inferior frontal gyrus and the temporo-parietal junction, areas classically associated with language processing (Fig. 2.a). To examine the temporal dynamics of the brain score around word onset, we first computed average brain scores across a predefined set of MEG sources and ECoG electrodes (see Methods) and then averaged within each set. As shown in Fig. 2.b, encoding profiles peaked approximately 300 ms afterword onset, with ECoG showing lower average values given the broad electrode selection compared to MEG. A session-by-session analysis (~ 1 hour per session) of individual participants in the MEG dataset revealed stable encoding across sessions, with brain scores comparable to those obtained when training on the full 10-hour dataset (Fig. S1). In contrast, analyses of individual participants in the ECoG dataset showed greater variability across participants (Fig. S2), consistent with differences in electrode coverage across subjects and the shorter session length (30 min).

Word embeddings explain brain responses before word onset

Notably, positive brain scores are observed up to one second before the word onset, indicating consistent preonset neural representations across modalities (Fig. 2.b), with these pre-onset effects being stronger for predictable words compared to unpredictable ones (Fig. S1). These findings align with those of Goldstein et al. (2022b), who used ECoG recordings to suggest that pre-onset encoding reflects predictions of the upcoming word. Moreover, as shown in Fig. S3, the amplitude of the FIR filter used in the preprocessing step reaches zero within 100 ms, indicating that the observed pre-onset encoding is unlikely to result from temporal smearing introduced by the Alter.

A potential concern is that pre-onset encoding might partly reflect residual processing of preceding words, since embeddings of neighboring words are highly correlated (see Fig. S4). To address this, we decorrelated the embeddings of nearby words using an approach similar to those used by Goldstein et al. (2022b) and Schönmann et al. (2025). However, rather than removing the effect of just one preceding word, we removed the influence of the eight previous words in the narrative (corresponding to ~ 2 seconds before word onset). This procedure eliminates potential contributions from past-word encoding to the pre-onset encoding. As shown in Fig. 2.b, pre-onset encoding remains largely unchanged, suggesting that correlations between neighboring word embeddings cannot account for the observed pre-onset encoding. Notably, this decorrelation of embeddings also eliminates the baseline offset observed in the ECoG encoding, suggesting that this offset reflects encoding of shared representations among adjacent embeddings.

To further investigate whether this pre-onset encoding genuinely reflects predictive pre-activation, we considered the possibility that the part of the pre-onset encoding might result from the encoding model learning word co-occurrences, particularly repeated bi-grams in the story. To test this, we re-ran the analysis after removing all but the first instance of each bi-gram. As shown in Fig. 2.b, this manipulation led to a slight reduction in preonset encoding, but the effect persisted. This finding suggests that pre-onset encoding is not merely an artifact of repeated co-occurrences in the stimulus set.

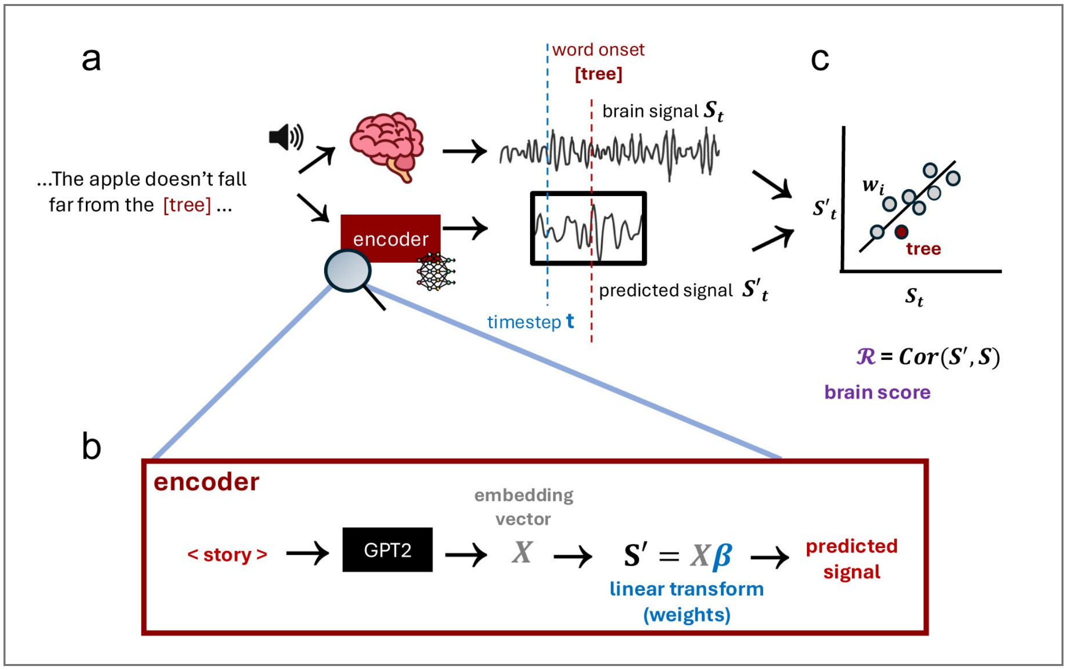

Figure 1. **The encoding model finds a linear mapping between words in the narrative and corresponding brain responses.**

a. The encoding model mimics the brain by taking each word and its context and learning to generate a brain-like response within a time window around the word's onset. **b.** Word embeddings are extracted from the GPT-2 model and used to predict brain responses at each time point *t* through linear regression. **c.** The brain score represents the correlation between actual and predicted brain responses across multiple words, calculated at each time point *t*.

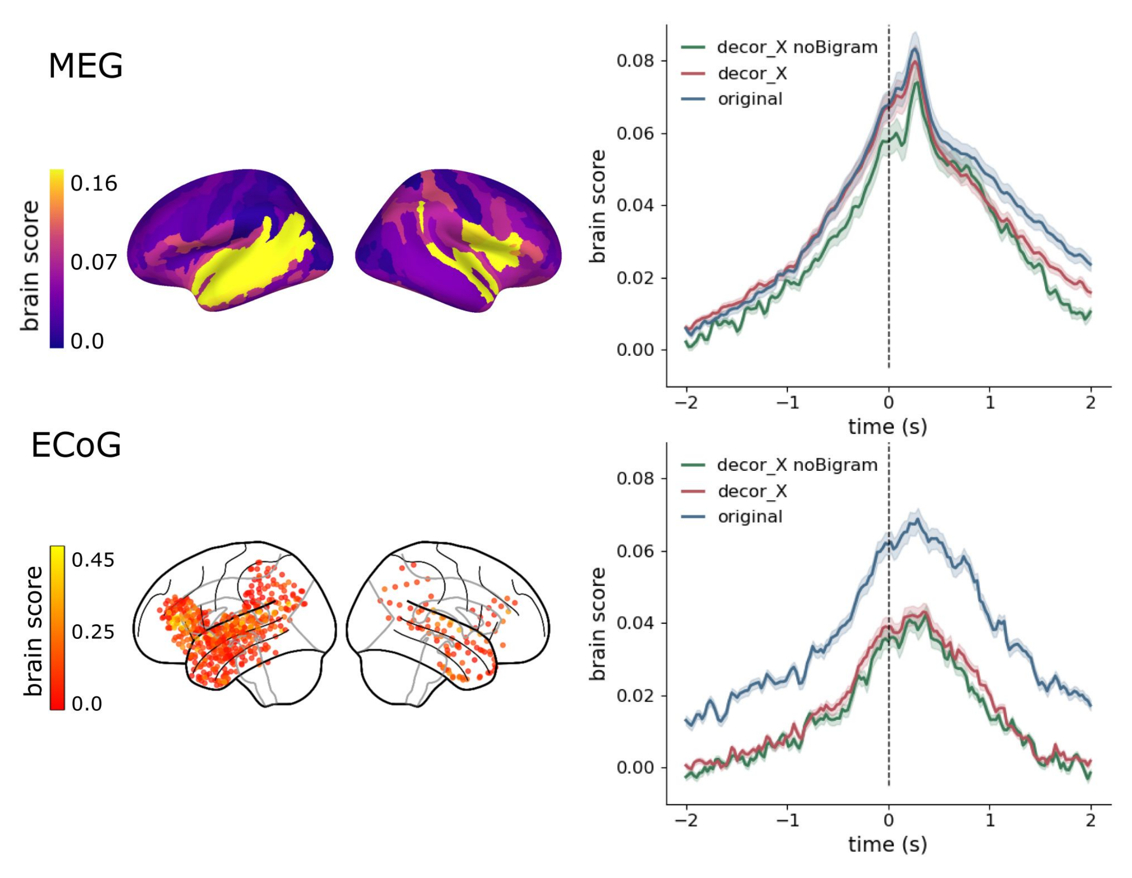

Figure 2. **Word embeddings explain brain responses before word onset**

left column: All brain regions show encoding of the word embedding, with peak values in the left hemisphere, especially in the temporal cortex, inferior frontal areas and tempoparietal junction associated with language processing. **right column:** A comparison of encoding models shows that removing correlations between neighboring word embeddings (decor_X) does not remove pre-onset encoding. Similarly, eliminating bi-grams in the narrative (decor_X noBigram) slightly reduces overall encoding performance, but the pre-onset effect persists. The shaded regions in the line plots indicate the standard error of the mean (SEM) across aggregated ECoG electrodes/MEG sources of all participants.

Pre- and post-onset encoding do not reflect the same neural representations

To assess whether pre-onset encoding truly reflects pre-activation of the same neural representations engaged after word onset, we compared the representations captured by the encoding model before and after word onset in more detail. Because separate linear regression models are fitted at each time pointwithin the [-2,2] s window in the encoding analyses above, it is not guaranteed that the model learns the same mapping between embeddings and brain activity across time. To directly test whether pre-onset encoding reflects predictive pre-activation, we evaluated whether the mapping between embeddings and the neural signal learned after word onset generalize to pre-onset activity using temporal generalization analysis (King and Dehaene, 2014). In this analysis, an encoder is trained on the neural signal at one time point and tested at another, relative to word onset.

As shown in Fig. 3.a, post-onset representations evolved dynamically over time, consistent with findings from Gwilliams et al. (2022, 2025). TG matrices revealed limited generalization between pre- and post-onset periods: models trained at the peak of word representation (~ 300 ms) predicted pre-onset activity only within a narrow window immediately preceding word onset (Fig. 3.b). This overlap accounts for at most ~ 400 ms of the pre-onset encoding observed previously, indicating that most of the pre-onset signal cannot be explained by pre-activation of the upcoming word.

As an additional test of whether pre-onset encoding reflects pre-activation, we decorrelated pre-onset from post-onset activity (see Fig. S5 for the correlation profile). This step should remove any residual trace of genuine pre-activation of the features typically evoked by the word. After decorrelation (Fig. 4), pre-onset encoding remained largely intact, while the small temporal generalization observed before onset disappeared. Together, these results indicate that pre-onset encoding and post-onset processing rely on distinct neural representations, ruling out pre-activation as a shared source.

Interestingly, the temporal generalization (TG) matrix revealed negative generalization areas approximately one second from the diagonal (Fig. 3, Fig. 4), suggesting an inversion of the neural code overtime. Similar negative-positive patterns have been observed in MEG studies of both visual and auditory processing (Carlson et al., 2013; King et al., 2014; King and Dehaene, 2014) and have been interpreted as reflecting reversals in neural activity patterns. However, in our data, the negative values do not simply correspond to a sign reversal in the neural signal (see Supplementary Material, Fig. S6, for details).

Future word embeddings do not improve encoding, while past word embeddings do

Caucheteux et al. (2023) demonstrated that incorporating future word embeddings alongside the current word in the encoding model improves the model's performance in explaining fMRI BOLD signals, consistent with the idea that the brain pre-activates representations of upcoming words. They further observed a cortical hierarchy of predictive timescales, where lower-level areas have shorter prediction windows, while frontal areas anticipate further into the future.

Here we applied a similar approach as Caucheteux et al. (2023) to test whether including future word embeddings enhances model performance. Moreover, given the superior temporal resolution of these modalities, we can track this effect over time. We used the decorrelated embeddings to make sure there are no confounding effects from the nearby word embeddings sharing information (although the same pattern of results was observed with the original embeddings Fig. S7).

If the findings of Caucheteux et al. (2023) generalize to MEG and ECoG data, we would expect to observe an increased brain score ($\Delta R > 0$) for models that include future word embeddings ($d > 0$) around the onset of the current word. However, as shown in Fig. 5.a, the inclusion of the next word embedding ($d = 1$) did not improve the brain score at the current word onset; rather, the

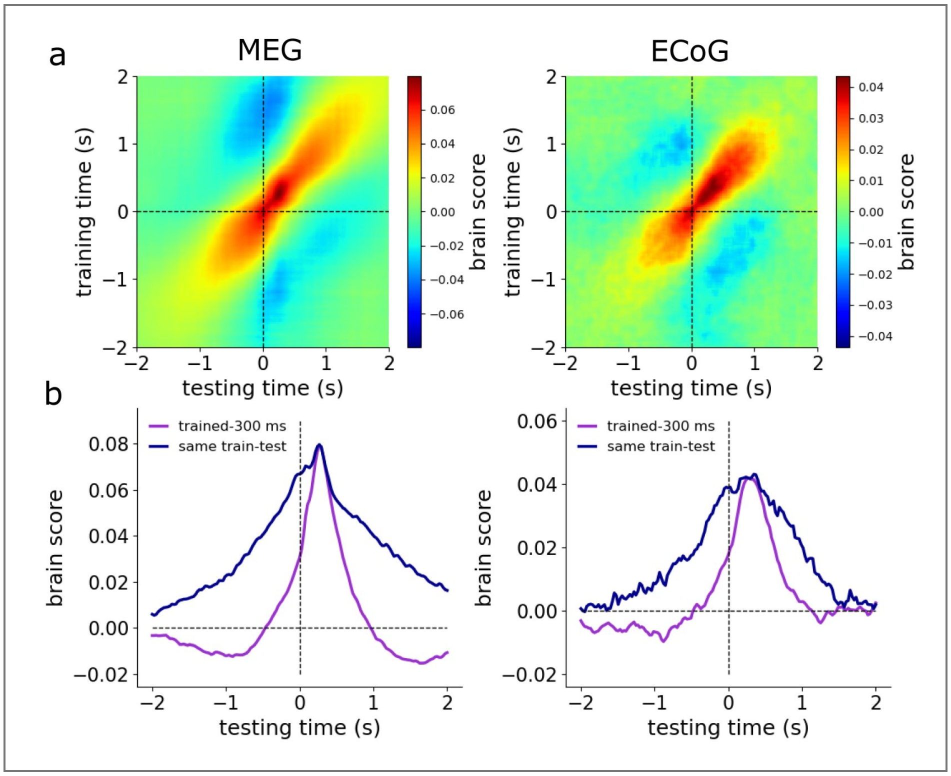

Figure 3. **Temporal generalization of representations captured by the encoding model differs before and after word onset.**

a. Temporal generalization (TG) matrix computed by training the encoding model with decorrelated embeddings at one time point and testing it at another. Positive values indicate successful generalization of representations across time. The diagonal pattern reflects temporally dynamic rather than stable representations. **b.** Generalization profiles for models trained at the peak encoding response (~ 300 ms; purple) and along the diagonal (blue). The divergence between the two curves indicates that pre-onset encoding does not reflect the same representations engaged during word processing.

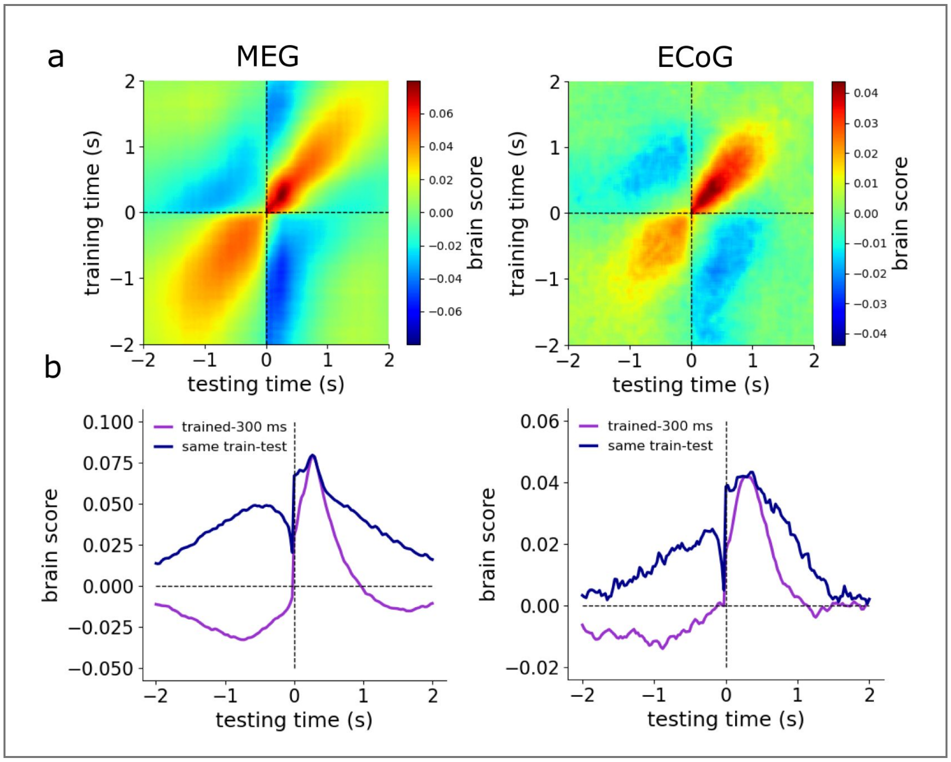

Figure 4. **Removing autocorrelation between pre- and post-onset activity does not eliminate pre-onset encoding.**

a. Temporal generalization (TG) matrix computed using decorrelated embeddings after regressing out post-onset brain activity from the pre-onset signal. Removing correlations in the neural signal should, in principle, eliminate any trace of predictive pre-activation. Following this procedure, the small pre-onset generalization observed near word onset in Fig. 3 disappears. **b.** Generalization profiles for models trained at the peak encoding response (~ 300 ms; purple) and along the diagonal (blue). The persistence of pre-onset encoding (blue curve) despite the absence of pre-activation indicates that pre-onset encoding is not necessarily a signature of prediction.

increase appeared only about 250 ms after the current word onset. Given that the average inter-word interval is 230-240 ms, this increase likely reflects the neural response to the next word once it has been heard.

Consistent with this, Fig. 5.c shows that increases in brain score emerge after approximately one inter-word interval, and curves for larger positive values of d diverge at subsequent multiples of this interval. Together, these results indicate that future-word embeddings improve model performance only after the corresponding words are presented, arguing against pre-activation of upcoming word representations.

In contrast, including previous word embeddings enhances encoding performance at the current word onset (Fig. 5.b). This increase is most pronounced for the immediate previous word and diminishes with the inclusion of words further back in the past. This suggests that representations of previous words remain active for a period of time after they have been processed. Fig. 5.b and c clearly show an asymmetry between the representations of past and future words, with past words still being encoded in the MEG brain responses around the current word onset, while future words are not.

To further examine whether this absence of effect could be attributed to our source selection method, we repeated the analysis using only MEG sources that, on average, exhibited enhanced encoding across the [0, 230] ms window for $d > 0$. Even with this strongly biased selection, encoding does not appear to improve noticeably within the [0,230] ms interval (Fig. S8). However, including previous word embeddings led to a measurable increase in encoding performance. The same is true when we specifically focus on the IFG electrodes in the ECoG data Fig. S9, which have been previously found to robustly show pre-onset encoding (Goldstein et al., 2022b). Together these findings further support the idea that past words remain active while future words are not pre-encoded in MEG signals.

Discussion

Recent studies have leveraged the representational richness of LLMs to investigate predictive processing during natural speech comprehension (Goldstein et al., 2022b; Caucheteux et al., 2023). Here, we sought to replicate these findings across MEG and ECoG datasets and determine to what extent they may reflect predictive preactivations of the upcoming words.

Consistent with previous ECoG results (Goldstein et al., 2022b), we found that embeddings of upcoming words explained neural activity in language-related regions. Importantly, we ruled out several potential confounds. The pre-onset effects could not be attributed to residual encoding of the preceding word via similarity between adjacent embeddings, as suggested by Schönmann et al. (2025): orthogonalizing consecutive embeddings within the narrative did not diminish the pre-onset encoding. Moreover, similarly to Goldstein et al. (2022b), removing all bigrams from the stimulus did not abolish the effect. Of course, removing n-grams does not remove all regularities inherent to natural speech, and fully removing all such dependencies (e.g., words that co-occur at longer distances) remains challenging as these issues are intrinsic to studying language in naturalistic contexts.

Given the many confounding factors that cannot be fully controlled, we directly examined the shared representations captured by the model before and afterword onset. The TG analysis suggests that the observed pre-onset encoding does not primarily reflect predictive pre-activation of upcoming words. Although a small portion of the pre-onset signal shares representational similarity with post-onset activity, this overlap disappears once pre- and post-onset correlations are removed, while the overall pre-onset encoding remains. Thus, pre-onset encoding reflects a distinct mapping between embedding of the word W_i and neural activity before its onset, rather than shared predictive representations.

Notably, the strongly diagonal structure of the TG matrix indicates that lexical representations are highly dynamic, with the mapping between embeddings and brain activity continuously transforming over time. This pattern is consistent with previous MEG findings during story listening (Gwilliams et al., 2022, 2025) and has been proposed to reflect a neural coding

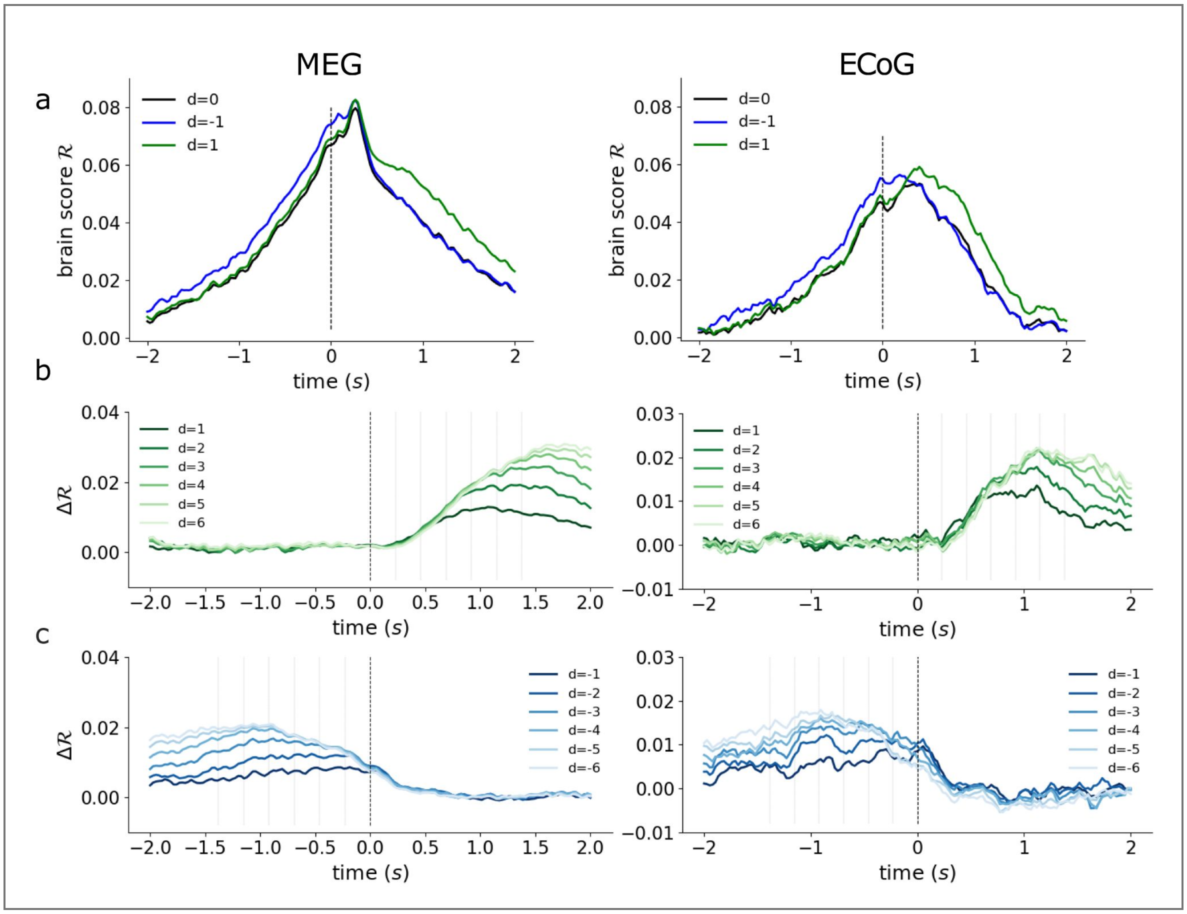

Figure 5. **Encoding of the future and past words.**

All curves represent averages across participants. The embedding vector is constructed by concatenating d future word embeddings ($d > 0$) or $|d|$ past word embeddings ($d < 0$) along with the embedding of the current word w_i. **a** Including the next word embedding in the encoding model ($d = 1$) enhances encoding only after that word is heard in the story, while including the previous word ($d = -1$) improves encoding even after the current word's onset. Encoding enhancement, $\Delta\mathscr{R}$, is shown for **b** positive and **c** negative values of d. Vertical gray lines mark the median inter-word interval values. Adding each successive future word embedding improves encoding only after that word is heard in the narrative, while including previous words consistently improves encoding beyond their offset.

strategy that allows the brain to process multiple stimuli simultaneously while avoiding interference between representations. This temporal variability underscores the need for a more nuanced view of what constitutes predictive pre-activation, as no stable representational state appears to persist afterword presentation that could serve as its target.

To further assess whether previously reported effects truly reflect predictive pre-activation, we tested whether incorporating future word embeddings would enhance the encoding of neural responses to the current word (w_i). Unlike the fMRI findings of Caucheteux et al. (2023), adding embeddings of upcoming words, even the immediate next word (w_{i+1}), did not improve encoding performance. This result held after controlling for embedding correlations and when using static GloVe embeddings, which capture semantic content but lack contextual information. These findings are consistent with our earlier results from the TG analysis and argue against a predictive pre-activation account in which specific word features are the target of prediction.

Why, then, can encoding of future words be captured with fMRI but not with MEG or ECoG? fMRI measures slower hemodynamic responses associated with neuronal metabolic activity (Logothetis, 2008), integrating neural signals over relatively long temporal windows. As a result, it likely reflects integrative semantic processes operating across extended contexts. Moreover, to match fMRI's temporal resolution, Caucheteux et al. (2023) averaged word embeddings within each repetition time (TR; approximately 1.5 s, covering about eight words). This temporal averaging likely produced embeddings that capture broader contextual meaning, resembling more stable, sentence-level representations rather than transient, word-specific ones. Consequently, the predictive signals observed in fMRI may primarily reflect pre-activation of contextual or semantic representations, rather than the prediction of specific upcoming words, a possibility that requires further investigation.

Unlike fMRI, MEG and ECoG rely on synchronized neural activity and with their high temporal resolution, directly capture rapid neural dynamics with millisecond precision. These modalities therefore primarily reflect instantaneous lexical processing rather than slower, integrative semantic processes. This interpretation aligns with findings from Toneva et al. (2022), who reported robust encoding of supra-word meaning (information derived from combinations of words beyond individual lexical items) in fMRI signals, whereas MEG predominantly captured lexical-level information.

In our analysis inspired by Caucheteux et al. (2023), we also tested whether incorporating embeddings of preceding words enhances MEG and ECoG encoding performance. Including past word embeddings improved model performance, indicating that representations of prior words remain active over time (Fig. 5.b). This effect was robust across both GPT embeddings (before and after decorrelation) and static GloVe embeddings (Fig. S10). These findings are consistent with those of Toneva et al. (2022), who reported that MEG signals encode both current and prior word meanings. Our replication of this pattern in ECoG further supports the view that electrophysiological recordings predominantly reflect representations of individual words rather than extended contextual predictions.

More broadly, our results support theories of postdictive processing, which propose that the brain buffers evidence overtime, refining interpretations of earlier stimuli based on incoming information (Gwilliams et al., 2018; Hogendoorn, 2022; Szewczyk et al., 2022). These findings emphasize the need to clarify how electrophysiological measures encode both individual word representations and context-dependent information.

In summary, our findings challenge the interpretation that pre-onset encoding in electrophysiological signals reflects predictive pre-activation of upcoming words. Instead, they suggest that MEG and ECoG primarily capture transient lexical dynamics and the reverberation of prior information, consistent with postdictive processing. More broadly, our results call for a re-evaluation of what current encoding models actually capture and for a sharper definition of what counts as evidence for "prediction" in the brain.

Methods and Materials

MEG

Dataset

We used an openly available dataset from Armeni et al. (2022). This dataset contains MEG recordings from a 275-channel axial gradiometer CTF system (sampling rate: 1200 Hz) collected while three native English-speaking participants (1 female; aged 35, 30, and 28 years) passively listened to 10 stories from the Adventures of Sherlock Holmes. All participants were right-handed with no reported neurological, developmental or language deficits. This dataset provides long MEG recordings (~ 10 sessions of ~ 1 hour each) per participant, enabling single-subject analysis. Additionally, participants wore MEG-compatible head casts to immobilize the head position during recording, yielding a high signal-to-noise ratio.

Preprocessing

All data were preprocessed using the MNE-Python open-source software package, v. 1.6.1 (Gramfort et al., 2013; Larson et al., 2024). Because our analysis approach required continuous data, we did not remove any noisy segments from the data; instead, we pre-processed the data using ICA, separately for each session and participant, removing components that related to eye movement and heartbeat artifacts. The components were identified from raw data that was filtered between 0.5 and 40 Hz using an FIR filter and then resampled to 150 Hz. After identifying the components, we removed them from the data filtered between 0.1-40 Hz using a zero-phase non-causal FIR filter.

Source reconstruction

Using the FreeSurfer software (Dale et al., 1999), we computed a single layer boundary element model for each participant using the 3T MRI anatomical scans provided in the dataset. For each subject, an individual single-shell boundary element model (BEM) was made and individual forward models were calculated. Source reconstruction was performed using a linearly constrained minimum variance (LCMV) beamformer with unit-noise-gain, optimizing source orientation to maximize output power (Sekihara and Nagarajan, 2008; Van Veen et al., 1997; Westner et al., 2022).

The LCMV adaptive filter was computed using a data covariance matrix within the [0,150] ms window after the stimulus onset (each word in the story). We regularized the covariance using Tikhonov regularization with 10% of overall sensor power. Finally, the filter was applied to the broadband raw time series. We computed 150 source regions based on the Destrieux Atlas (available through FreeSurfer; (Dale et al., 1999; Destrieux et al., 2010; Fischl et al., 2004)) averaging source points within regions for each participant.

ECoG dataset and preprocessing

Here we use on the “Podcast” ECoG dataset introduced by Zada etal. (2025). The dataset contains ECoG recordings from 9 epilepsy patients with a total of 1,330 electrodes (grid, strip, and depth) as they listened to a 30-min of “This American Life” podcast. The open dataset provides preprocessed data with bad-channel exclusion, artifact removal, and high-gamma (70-200 Hz) broadband power extraction. We used the provided high-gamma signal directly, applying no additional preprocessing except for downsampling from 512 Hz to 150 Hz to match the MEG sampling frequency.

Word embeddings

To extract contextual word embeddings, we used the pre-trained GPT2-small model with 12 hidden layers parameters (Radford et al., 2019). We first tokenized the text (without punctuation and capitalization) using the model tokenizer. To obtain the GPT embedding for each token w_i in the story, we fed chunks of the tokenized story in windows of size 50 tokens $[w_{i-49}, w_{i-48}, \ldots, w_i]$ (with the first 50 words having a smaller context window) to the GPT2-small tokenizer model and extracted the hidden state of the 8th out of 12 layers of the model $\mathbf{x} \in \mathbb{R}^{768}$ corresponding to the last word. This was because the 8th layer showed slightly higher encoding than the 12th layer. Finally, words with multiple tokens assigned by the model were excluded to remove within-word effects. This left us with a total of 80472 words split into 10 similar-sized sessions for the MEG dataset and 4708 in a single session for the ECoG dataset. For certain conditions, bi-grams were removed from the story which reduced the MEG dataset to 2,5 h and the ECoG dataset 19 min. For non-contextual embeddings, we used GloVe (Wikipedia 2014) embeddings (Pennington et al., 2014) with vectors of size 300. To reduce computational load, the embeddings were dimensionally reduced to 50 using PCA. This was done on the training set and applied to the test set to prevent data leakage. This resulted in $\mathbf{X} \in \mathbb{R}^{N\times 50}$ as predictors.

Encoding model and brain score

To preprocess the MEG data for encoding model training, time series data were smoothed using a 100 ms rectangular window (as in Goldstein et al., 2022b). The data were then epoched into time windows from [-2 to +2 sec] around word onset and downsampled to 151 time points within this range, yielding a target matrix $\mathbf{Y} \in \mathbb{R}^{N\times 151}$, where N is the number of words.

The encoding model was built using linear ridge-regression with word embeddings $\mathbf{X}$ as predictors and brain response $\mathbf{Y}$ as targets. A separate model was fit for each time point $\mathbf{t}$ within the [-2 to +2 sec] window. To find the optimal regularization parameter λ of the ridge-regression, we used grid search with generalized crossvalidation, which approximates leave-one-out cross-validation (Hastie et al., 2017). We defined a grid to search by first sampling the effective degrees of freedom of the ridge-regression fit from [1,N], since its parameter space is bounded from above. To find degrees of freedom of the ridge regression we used singular value decomposition $\mathbf{X} = \mathbf{USV}^{\boldsymbol{T}}$ where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and $\mathbf{S}$ diagonal matrices. The degrees of freedom is then defined as $df\left(\lambda_i\right) = \sum_{j=1}^{N} \frac{s_j^2}{s_j^2 + \lambda_i}$. We then used Newton's method to solve df for λ_i. Once the grid was defined, we chose the optimal value λ^* that minimizes the generalized cross-validation error using Leave-One-Out Cross-Validation implementation of scikit-learn library in python (Pedregosa et al., 2011). As λ^* remained highly stable across timepoints, electrode/sources, sessions and participants, a fixed value was used per dataset in subsequent computations to improve efficiency.

We assessed model performance using 5-fold cross-validation, training on 80% of the story and testing on the remaining 20% Training data was normalized to a mean of 0 and variance of 1, and the same transformation applied to the test set. To prevent overfitting due to temporal autocorrelation in brain activity, training and test sets were temporally separated (Feghhi et al., 2024). This produced a predicted brain signal for each word at each time point within the [-2 to +2 sec] window around word onset. To evaluate model fit, we computed the Pearson correlation between the predicted and actual brain signals at each time point t around word onset across all words to compute the brain score $\mathscr{R}$.

Temporal generalization analysis

To examine how similar representations captured by the encoding model are pre- and post-onset, similarly to before, we trained an encoder to predict brain responses at time point t_i. However this time, we tested the model by predicting brain responses at time point t_{i+n} and calculated Pearson correlation between this prediction and actual brain response (i.e. brain score). Doing so for all 151 time points yielded a temporal generalization matrix of size 151 × 151.

Analysis with expanded future/past window

To test whether future/past words are represented in the brain response, we constructed embedding vectors by concatenating additional word embeddings to the current word embedding. For a given window size d, we defined the embedding as $\mathrm{z}_i^d = \mathrm{x}_{i+1} \oplus \mathrm{x}_{i+d-1} \oplus \ldots \oplus \mathrm{x}_i$ where $\mathbf{x}_{\boldsymbol{i}}$ is the word embedding vector of word w_i. Negative and positive d values correspond to windows spanning the words in past or future respectively.

Finally, we reduced dimensions of both z_i^d and $\mathbf{x_i}$ to 50 and concatenated the two as $\hat{z}_i^d \oplus \hat{x}_i$, where $\hat{z}$ and $\hat{x}$ are the vectors after applying PCA. To see how much encoding improves, we computed the change in the brain score when including future/past word embeddings as $\Delta\mathscr{R} = \mathscr{R}\left(\widehat{Z}^d \oplus \widehat{X}\right) - \mathscr{R}\left(\widehat{X}\right)$, where $\mathscr{R}(.)$ denotes the brain score computed given the respective embeddings.

MEG Source selection

To compute the brain-wide brain score, we selected a subset of 30 sources for each participant. To avoid double dipping, we ensured that the source selection for each participant was not based on their own data. To do this, we averaged the brain score of two participants for each MEG source and chose the top 30 sources with the highest values for the third participant.

Supplementary Information

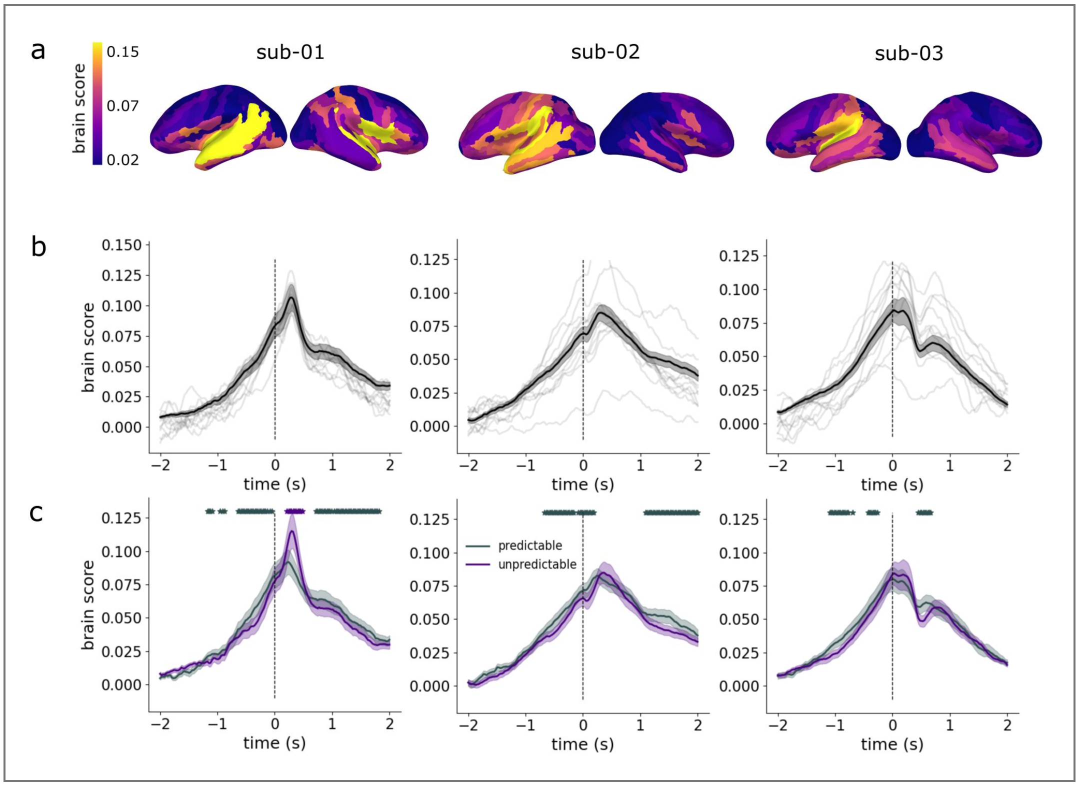

Figure S1. **Encoding results for individual subjects in the MEG dataset** To examine whether predictable words are better encoded before word onset, we categorized words as predictable or unpredictable based on whether they appeared in the top-5 predictions of GPT-2 for that position (Goldstein et al., 2022b). **a.** Brain score values peak in the left hemisphere, especially in the temporal cortex and inferior frontal areas associated with language processing. **b.** Brain scores from single (gray) and all 10 (black) sessions show encoding before word onset, peaking at or shortly after it (see Methods for MEG source selection). **c.** The encoding model was trained separately to compute brain scores for predictable and unpredictable words. The figure indicates that predictable words show stronger encoding up to 1 second before word onset. The encoding of unpredictable words appears stronger around 400 ms post-onset, though these effects did not reach statistical significance for subjects 2 and 3. Star symbols mark significant differences between predictable and unpredictable words calculated using a dependent t-test for paired samples on the brain scores of the two groups across the same MEG sources and accounted for multiple hypothesis testing using the Benjamini-Hochberg correction. we considered *q* values smaller than 0.05 as significant. A smoothing window of 200ms was used in this analysis. All error bars were computed as the standard error across the MEG sources.

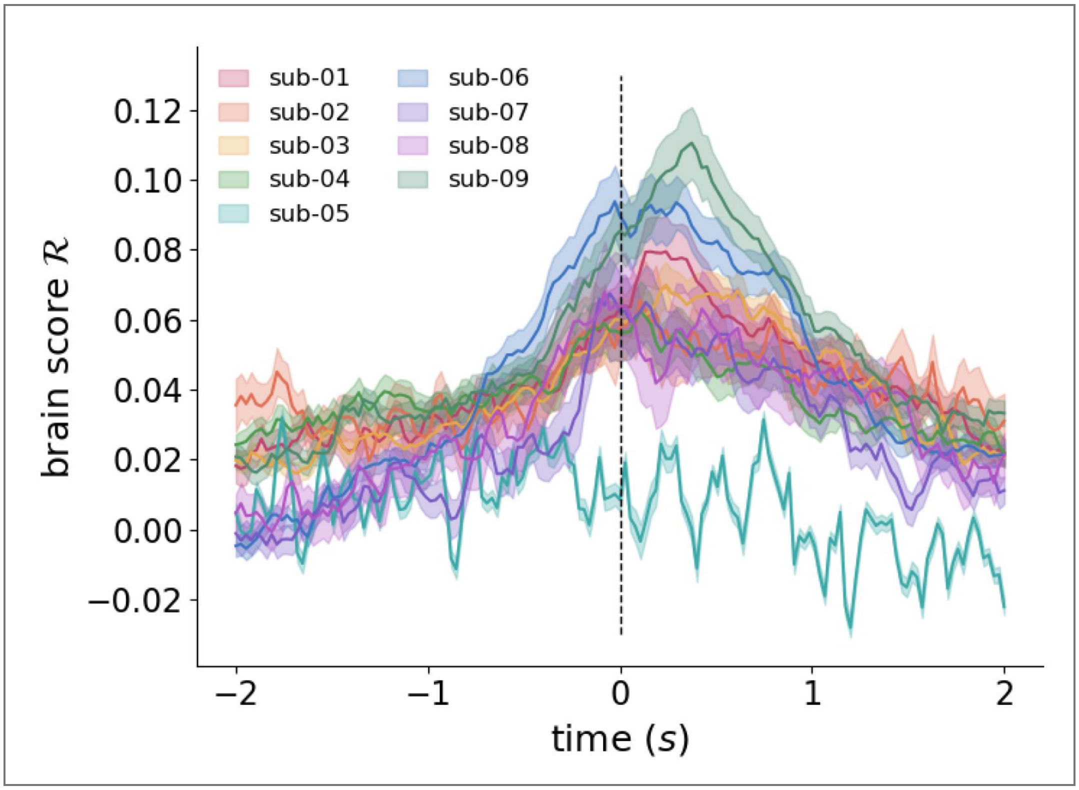

Figure S2. Encoding values for individual subjects in the ECoG dataset. Brain score values and temporal profiles vary across subjects in the ECoG dataset, partly due to differences in the number and locations of electrodes selected for each participant. Shaded regions in the line plots indicate the standard error of the mean (SEM) across ECoG electrodes included for each subject.

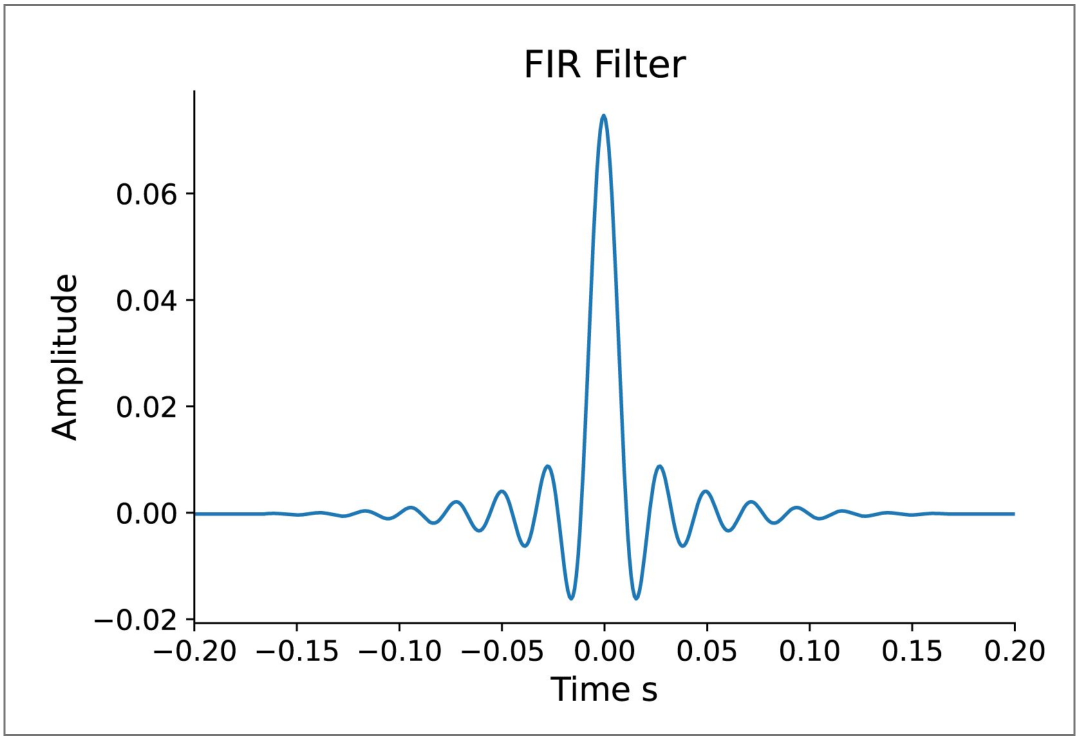

Figure S3. Shape of the FIR filter used for band-pass filtering MEG data between 0.1-40 Hz. Shown is the flnite impulse response (FIR) Alter implemented in MNE and used for MEG preprocessing. Because this Alter is non-causal and applied in both forward and reverse directions, it can, in principle, introduce temporal leakage.

However, the Alter amplitude converges to zero within the first 100 ms, making any temporal leakage negligible for the effects examined in this study.

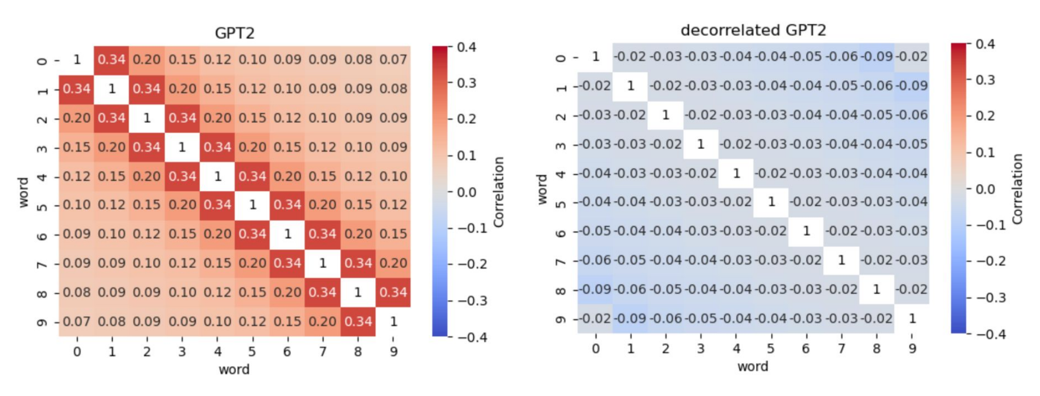

Figure S4. Average correlation between nearbyword embeddings in the 10h MEG dataset. (left) GPT2 embeddings have large, symmetric correlations between nearbywords. (right) After removing the correlation of each word embedding in the story with its previous 8 word embeddings, most of the correlations between nearbyword embeddings disappear. These figures were obtained by calculating the correlation matrix given each 10 consecutive words (in non-overlapping windows), and then averaging all those values across the 10h MEG dataset.

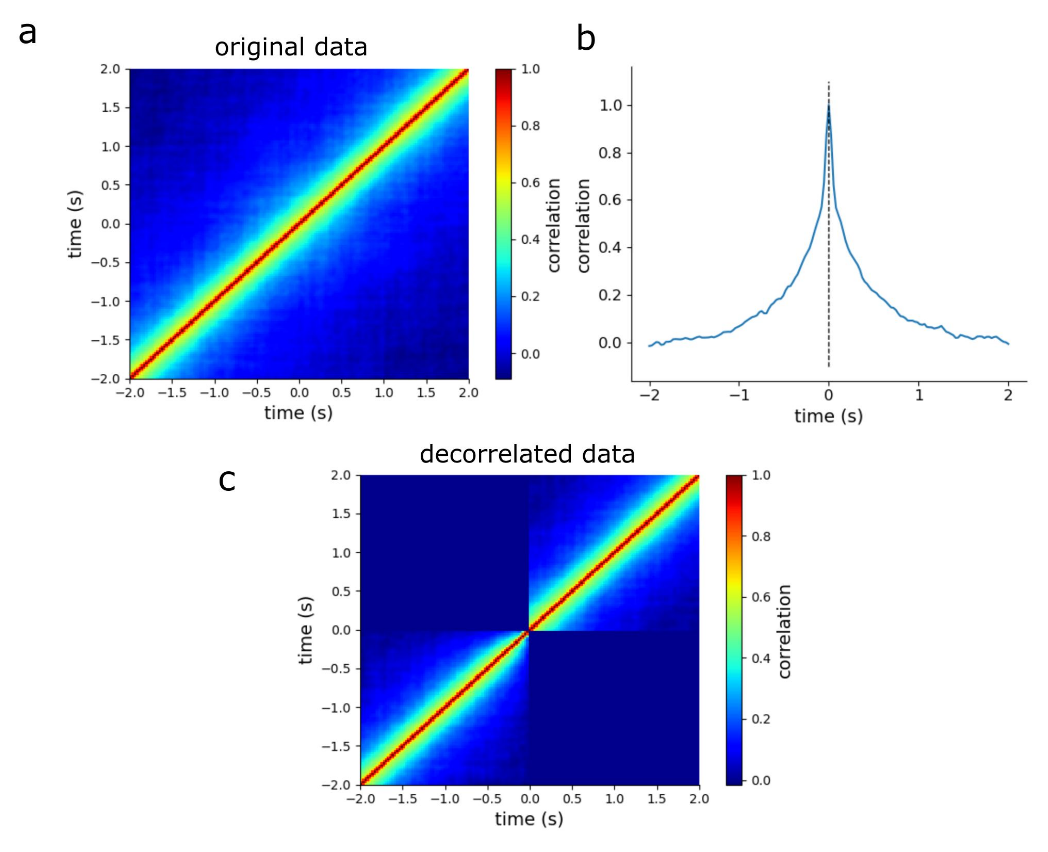

Figure S5. Temporal correlation within the epochs in the MEG dataset. a. shows the time-by-time correlation matrix of the MEG signal across the epochs. **b.** shows a slice at t=0 of the correlation matrix. The autocorrelation decays rapidly and reaches zero within ~ 1 second. **c** shows the correlation matrix after regressing out all post-onset time-points from each pre-onset time-point.

The temporal generalization (TG) matrix revealed negative generalization areas approximately one second from the diagonal (Fig. 3 and Fig. 4), suggesting an inversion of the neural code over time. Similar negative-positive patterns have been reported in MEG studies of visual and auditory processing (Carlson et al., 2013; King et al., 2014; King and Dehaene, 2014) and are often attributed to reversals in neural activity patterns.

In our case, because the encoding model is linear and maps embeddings to individual MEG time points, the negative generalization might reflect a sign flip in the underlying neural response from pre- to post-onset, potentially arising from the oscillatory nature of the MEG signal. Although we do not observe such a sign flip in the correlation matrix computed across the full dataset (Fig. S5), it could still exist within the subset of words that most strongly drive the encoding model.

To test this, we examined encoding model predictions for each word in the story and selected those for which the model predicts a sign reversal across word onset (Fig. S6.a). If the negative TG values were caused by a true sign flip in the MEG signal, this subset should exhibit negative correlations between pre- and post-onset neural responses. However, Fig. S6.b shows that the recorded MEG activity for these words does not switch sign across onset.

Together, these findings suggest that the negative values in the TG analysis are not simply driven by a corresponding reversal in the neural signal itself. In fact, Gwilliams et al. (2025) showed in a simulation that this type of negative-positive diagonal pattern can emerge from a spatio-temporal dynamic neural code. They suggest that the neural representation of an item in a temporal sequence, such as a word in a sentence, moves through neural representational space over time. This effectively creates room for the next item to be encoded while still maintaining the earlier item in memory, avoiding interference between them.

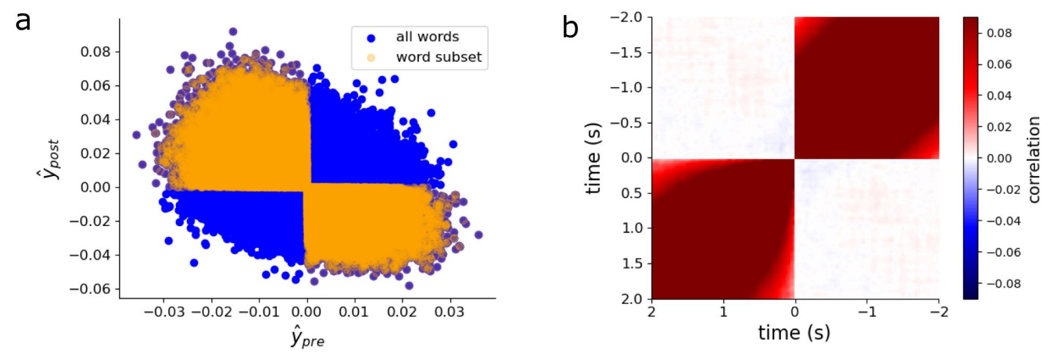

Figure S6. **Negative areas in the TG matrix are not caused by a sign flip in brain activity. a.** Average model predictions for pre- and post-onset activity (averaged over [-600, 0] ms and [0, 600] ms, respectively). For each MEG source, a subset of words showing anticorrelated model predictions is highlighted in orange. This analysis was performed on data from the first participant in the MEG dataset after decorrelating pre- and post-onset signals, similar to Fig. 4. **b.** Correlation matrix computed for the selected subset of words at each MEG source and averaged across the 30 MEG sources used in the main analyses. The resulting matrix shows that even for this subset, there is no systematic sign reversal in brain activity before and after word onset; instead, the two signals remain largely uncorrelated.

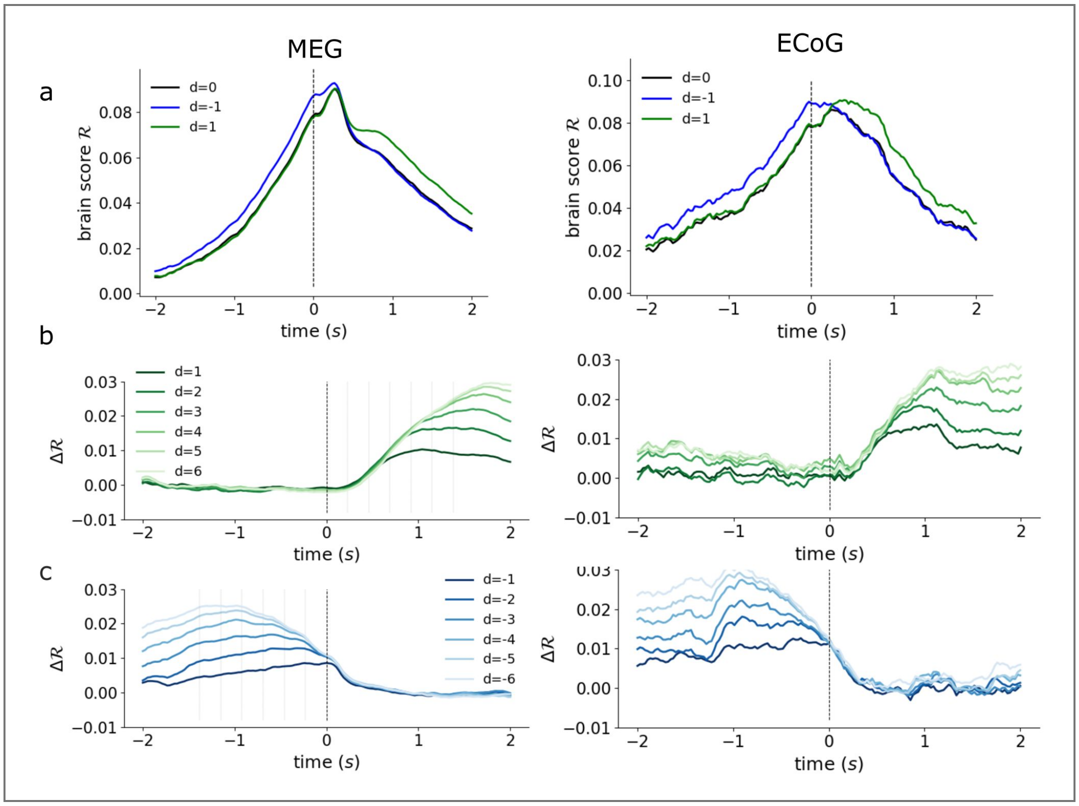

Figure S7. Analysis with original gpt2 embeddings also shows that future word embeddings do not improve encoding, while past word embeddings do. This figure shows was obtained similar to Fig. 5 but using original gpt2 embeddings, instead of decorrelated embeddings. All curves represent averages across participants. The embedding vector is constructed by concatenating d future word embeddings ($d > 0$) or $|d|$ past word embeddings ($d < 0$) along with the embedding of the current word w_i. **a** Including the next word embedding in the encoding model ($d = 1$) enhances encoding only after that word is heard in the story, while including the previous word ($d = -1$) improves encoding even after the current word's onset. Encoding enhancement, $\Delta\mathscr{R}$, is shown for **b** negative and **c** positive values of d. Vertical gray lines mark the median inter-word interval values. Adding each successive future word embedding improves encoding only after that word is heard in the narrative, while including previous words consistently improves encoding beyond their offset.

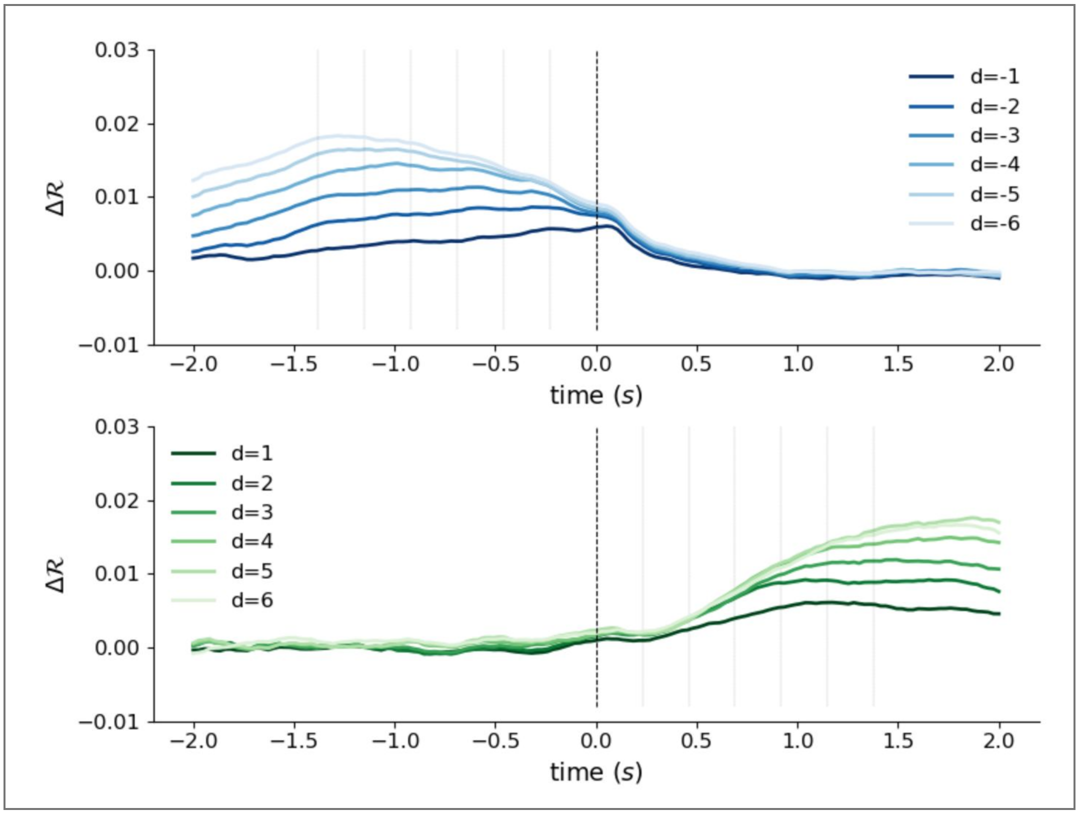

Figure S8. **Lack of evidence for future-word encoding is not due to MEG source selection.** To test whether the absence of an effect observed in Fig. 5 could be explained by our MEG source selection procedure, we repeated the analysis using only sources that, on average, showed enhanced encoding within the [0, 230] ms window for $d > 0$. Even with this intentionally biased selection, encoding did not noticeably improve within the [0, 230] ms interval. In contrast, including embeddings of previous words resulted in a measurable increase in encoding performance.

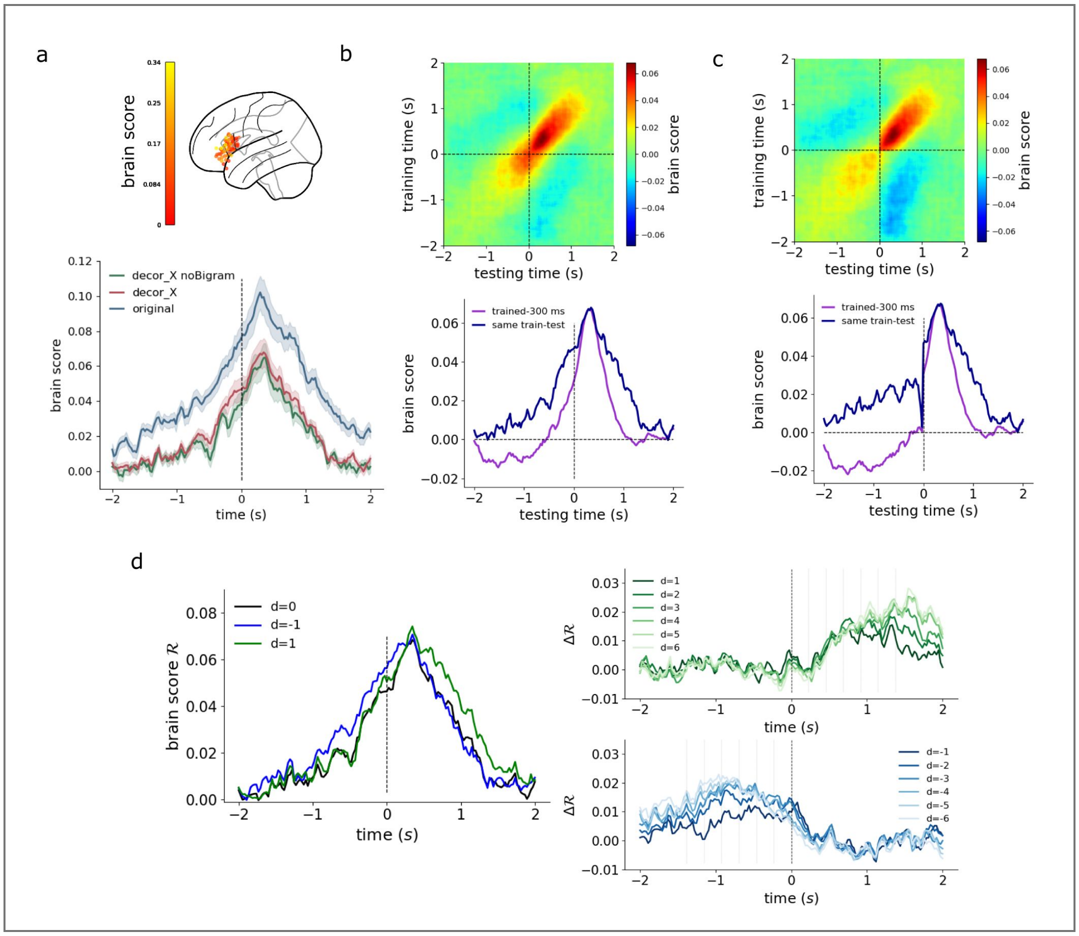

Figure S9. **No evidence for predictive pre-activation in the IFG region of the ECoG dataset.** To test whether the absence of effects was due to the broad selection of ECoG electrodes, we repeated the analysis using only electrodes located in the inferior frontal gyrus (IFG), previously reported to show robust pre-onset encoding of upcoming stimuli (Goldstein et al., 2022b). The same qualitative patterns were observed in the IFG as in the borader selection of electrodes in the main text.**a. top**: location of the IFG electrodes included in this analysis.**bottom**: A comparison of encoding models shows that removing correlations between neighboring word embeddings (decor_X) doesn’t eliminate pre-onset encoding. Similarly, eliminating bi-grams in the narrative (decor_X noBigram) slightly reduces overall encoding performance, but the pre-onset effect persists. The shaded regions in the line plots indicate the standard error of the mean (SEM) across aggregated ECoG electrodes/MEG sources of all participants. **b.** Temporal generalization of representations captured by the encoding model differs before and after word onset. **top**: Temporal generalization (TG) matrix computed with decorrelated embeddings . Positive values indicate successful generalization of representations across time. **bottom**: Generalization profiles for models trained at the peak encoding response (~300 ms;purple) and along the diagonal (blue). The divergence between the two curves indicates that pre-onset encoding does not reflect the same representations engaged during word processing., **c.** Removing autocorrelation between pre- and post-onset activity does not eliminate pre-onset encoding. **top** Temporal generalization (TG) matrix computed using decorrelated embeddings after regressing out post-onset brain activity from the pre-onset signal. Removing correlations in the neural signal should, in principle, eliminate any trace of predictive pre-activation. Following this procedure, the small pre-onset generalization observed near word onset in panel b. disappears. **bottom** Generalization profiles for models trained at the peak encoding response (~300 ms;purple) and along the diagonal (blue). The persistence of pre-onset encoding (blue curve) despite the absence of pre-activation indicates that pre-onset encoding is not necessarily a signature of prediction. .**d.** Encoding of the future and past words. All curves represent averages across participants. The embedding vector is constructed by concatenating d future word embeddings ($d > 0$) or $|d|$ past word embeddings ($d < 0$) along with the embedding of the current word w_i. **left** Including the next word embedding in the encoding model ($d = 1$) enhances encoding only after that word is heard in the story, while including the previous word ($d = -1$) improves encoding even after the current word’s onset. Encoding enhancement, $\Delta\mathscr{R}$, is shown for **top right** positive and **bottom right** positive values of d. Vertical gray lines mark the median inter-word interval values. Adding each successive future word embedding improves encoding only after that word is heard in the narrative, while including previous words consistently improves encoding beyond their offset.

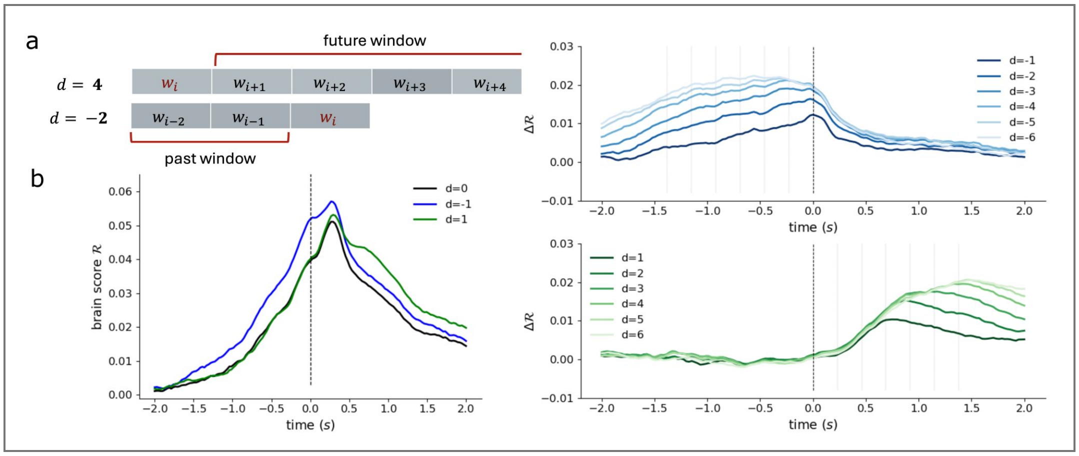

Figure S10. **No evidence for future-word encoding replicated using GloVe embeddings in MEG data.** To test whether the absence of future-word encoding observed in Fig. 5 could be attributed to the contextual nature of GPT embeddings, we repeated the analysis using static GloVe embeddings. **a.** The variable *d* represents the number of word embeddings concatenated with the current word embedding to form the final vector. Positive values of *d* signify the inclusion of future word embeddings and negative *d* signify the inclusion of past word embeddings. **b. and c.** The same qualitative pattern was observed: including future-word embeddings did not enhance model performance within the [0, 230] ms window after word onset, whereas including past-word embeddings clearly improved it.

Data availability

All analyses were performed using previously published, publicly available datasets (Dataset 1: *https://data.ru.nl/collections/di/dccn/DSC_3011085.05_995*; Dataset 2: *https://openneuro.org/datasets/ds005574*). No new data were generated in this study. Analysis code has been made available at *https://github.com/sahelazizpour/Linguistic_Preactivation*.

Acknowledgements

We would like to thank Lucia Rust for contributing to initial explorations of certain analyses; Nasir Ahmad and Dora Gözükara for the valuable and insightful discussions that greatly enhanced this work.

Additional information

Funding

Funder	Grant reference number	Author
NOW	VI.Vidi.201.150	Linda Geerligs

References

Antonello R, Huth A (2024) Predictive coding or just feature discovery? An alternative account of why language models fit brain data. *Neurobiology of Language* **5**:64-79
https://doi.org/10.1162/nol_a_00087 | PubMed

Armeni K, Güçlü U, van Gerven M, Schoffelen JM (2022) A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. *Scientific Data* **9**:278 https://doi.org/10.1038/s41597-022-01382-7 | PubMed

Baroni M (2022) On the proper role of linguistically oriented deep net analysis in linguistic theorising. In: Lappin S, Bernardy J-P (Eds). *Algebraic structures in natural language* CRC Press. https://doi.org/10.1201/9781003205388-1

Carlson T, Tovar DA, Alink A, Kriegeskorte N (2013) Representational dynamics of object vision: the first 1000 ms. *Journal of vision* **13**:1-1 https://doi.org/10.1167/13.10.1 | PubMed

Caucheteux C, Gramfort A, King JR (2023) Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour* **7**:430-441 https://doi.org/10.1038/s41562-022-01516-2 | PubMed

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

Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences* **36**:181-204 https://doi.org/10.1017/s0140525x12000477 | PubMed

Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage* **9**:179-194 https://doi.org/10.1006/nimg.1998.0395 | PubMed

De Lange FP, Heilbron M, Kok P (2018) How do expectations shape perception?. *Trends in cognitive sciences* **22**:764-779 https://doi.org/10.1016/j.tics.2018.06.002 | PubMed

De Ruiter JP, Mitterer H, Enfield NJ (2006) Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language* **82**:515-535 https://doi.org/10.1353/lan.2006.0130

DeLong KA, Urbach TP, Kutas M (2005) Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience* **8**:1117-1121 https://doi.org/10.1038/nn1504 | PubMed

Destrieux C, Fischl B, Dale A, Halgren E (2010) Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* **53**:1-15 https://doi.org/10.1016/j.neuroimage.2010.06.010 | PubMed

Federmeier KD (2007) Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology* **44**:491-505 https://doi.org/10.1111/j.1469-8986.2007.00531.x | PubMed

Feghhi E, Hadidi N, Song B, Blank IA, Kao JC (2024) What Are Large Language Models Mapping to in the Brain? A Case Against OverReliance on Brain Scores. *arXiv* https://doi.org/10.48550/arXiv.2406.01538

Fischl B, Van Der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, Busa E, Seidman LJ, Goldstein J, Kennedy D, *et al.* (2004) Automatically parcellating the human cerebral cortex. *Cerebral cortex* **14**:11-22 https://doi.org/10.1093/cercor/bhg087 | PubMed

Friston K (2010) The free-energy principle: a unifled brain theory?. *Nature reviews neuroscience* **11**:127-138 https://doi.org/10.1038/nrn2787 | PubMed

Goldstein A, Ham E, Nastase SA, Zada Z, Grinstein-Dabus A, Aubrey B, Schain M, Gazula H, Feder A, Doyle W, *et al.* (2022) Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain. *BbioRxiv* https://doi.org/10.1101/2022.07.11.499562

Goldstein A, Zada Z, Buchnik E, Schain M, Price A, Aubrey B, Nastase SA, Feder A, Emanuel D, Cohen A, *et al.* (2022) Shared computational principles for language processing in humans and deep language models. *Nature neuroscience* **25**:369-380 https://doi.org/10.1038/s41593-022-01026-4 | PubMed

Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L, *et al.* (2013) MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroinformatics* **7**:267 https://doi.org/10.3389/fnins.2013.00267 | PubMed

Gwilliams L, King JR, Marantz A, Poeppel D (2022) Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature communications* **13**:6606 https://doi.org/10.1038/s41467-022-34326-1 | PubMed

Gwilliams L, Linzen T, Poeppel D, Marantz A (2018) In spoken word recognition, the future predicts the past. *Journal of Neuroscience* **38**:7585-7599 https://doi.org/10.1523/jneurosci.0065-18.2018 | PubMed

Gwilliams L, Marantz A, Poeppel D, King JR (2025) Hierarchical dynamic coding coordinates speech comprehension in the human brain. *bioRxiv* https://doi.org/10.1073/pnas.2422097122 | PubMed

Hamilton LS, Huth AG (2020) The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, cognition and neuroscience* **35**:573-582 https://doi.org/10.1080/23273798.2018.1499946 | PubMed

Hastie T, Tibshirani R, Friedman J (2017) *The elements of statistical learning: data mining, inference, and prediction* Springer.

Hogendoorn H (2022) Perception in real-time: predicting the present, reconstructing the past. *Trends in Cognitive Sciences* **26**:128-141 https://doi.org/10.1016/j.tics.2021.11.003 | PubMed

King JR, Dehaene S (2014) Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences* **18**:203-210 https://doi.org/10.1016/j.tics.2014.01.002 | PubMed

King JR, Gramfort A, Schurger A, Naccache L, Dehaene S (2014) Two distinct dynamic modes subtend the detection of unexpected sounds. *PloS one* **9**:e85791 https://doi.org/10.1371/journal.pone.0085791 | PubMed

Kuperberg GR, Jaeger TF (2016) What do we mean by prediction in language comprehension?. *Language, cognition and neuroscience* **31**:32-59 https://doi.org/10.1080/23273798.2015.1102299 | PubMed

Kutas M, DeLong KA, Smith NJ (2011) A look around at what lies ahead: Prediction and predictability in language processing. *Predictions in the brain: Using our past to generate a future* 190207(10.1093) https://doi.org/10.1093/acprof:oso/9780195395518.003.0065

Kutas M, Federmeier KD (2011) Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology* **62**:621-647 https://doi.org/10.1146/annurev.psych.093008.131123 | PubMed

Larson E, Gramfort A, Engemann DA, Leppakangas J, Brodbeck C, Jas M, Brooks T, Sassenhagen J, Luessi M, McCloy D, *et al.* (2024) MNE-Python. https://doi.org/10.5281/zenodo.10519948

Logothetis NK (2008) What we can do and what we cannot do with fMRI. *Nature* **453**:869-878 https://doi.org/10.1038/nature06976 | PubMed

Magyari L, De Ruiter JP (2012) Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in psychology* **3**:376 https://doi.org/10.3389/fpsyg.2012.00376 | PubMed

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, *et al.* (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**:2825-2830

Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532-1543 https://doi.org/10.3115/v1/d14-1162

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. (2019) Language models are unsupervised multitask learners. *OpenAI blog* **1**:9

Schönmann I, Szewczyk J, de Lange FP, Heilbron M (2025) Stimulus dependencies—rather than next-word prediction—can explain pre-onset brain encoding during natural listening. *bioRxiv* https://doi.org/10.1101/2025.03.08.642140

Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, Tenenbaum JB, Fedorenko E (2021) The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* **118**:e2105646118 https://doi.org/10.1073/pnas.2105646118 | PubMed

Sekihara K, Nagarajan SS (2008) *Adaptive spatial filters for electromagnetic brain imaging* Springer Science & Business Media.

Smith NJ, Levy R (2013) The effect of word predictability on reading time is logarithmic. *Cognition* **128**:302-319 https://doi.org/10.1016/j.cognition.2013.02.013 | PubMed

Szewczyk JM, Mech EN, Federmeier KD (2022) The power of "good": Can adjectives rapidly decrease as well as increase the availability of the upcoming noun?. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **48**:856 https://doi.org/10.1037/xlm0001091 | PubMed

Toneva M, Mitchell TM, Wehbe L (2022) Combining computational controls with natural text reveals aspects of meaning composition. *Nature computational science* **2**:745-757 https://doi.org/10.1038/s43588-022-00354-6 | PubMed

Traxler MJ, Foss DJ (2000) Effects of sentence constraint on priming in natural language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **26**:1266 https://doi.org/10.1037//0278-7393.26.5.1266 | PubMed

Van Veen BD, Van Drongelen W, Yuchtman M, Suzuki A (1997) Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on Biomedical Engineering* **44**:867-880 https://doi.org/10.1109/10.623056 | PubMed

Wang L, Brothers T, Jensen O, Kuperberg GR (2024) Dissociating the pre-activation of word meaning and form during sentence comprehension: Evidence from EEG representational similarity analysis. *Psychonomic Bulletin & Review* **31**:862-873 https://doi.org/10.3758/s13423-023-02385-0 | PubMed

Wang L, Wlotko E, Alexander E, Schoot L, Kim M, Warnke L, Kuperberg GR (2020) Neural evidence for the prediction of animacy features during language comprehension: evidence from MEG and EEG representational similarity analysis. *Journal of Neuroscience* **40**:3278-3291 https://doi.org/10.1523/jneurosci.1733-19.2020 | PubMed

Westner BU, Dalal SS, Gramfort A, Litvak V, Mosher JC, Oostenveld R, Schoffelen JM (2022) A unified view on beamformers for M/EEG source reconstruction. *NeuroImage* **246**:118789 https://doi.org/10.1016/j.neuroimage.2021.118789 | PubMed

Willems RM, Nastase SA, Milivojevic B (2020) Narratives for neuroscience. *Trends in neurosciences* **43**:271-273 https://doi.org/10.1016/j.tins.2020.03.003 | PubMed

Zada Z, Goldstein A, Michelmann S, Simony E, Price A, Hasenfratz L, Barham E, Zadbood A, Doyle W, Friedman D, *et al.* (2024) A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. *Neuron* **112**:3211-3222 https://doi.org/10.1016/j.neuron.2024.06.025 | PubMed

Zada Z, Nastase SA, Aubrey B, Jalon I, Michelmann S, Wang H, Hasenfratz L, Doyle W, Friedman D, Dugan P, *et al.* (2025) The "Podcast" ECoG dataset for modeling neural activity during natural language comprehension. *Scientific Data* **12**:1135 https://doi.org/10.1038/s41597-025-05462-2 | PubMed

Armeni K., Schoffelen J.M. (2022) A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. Radboud university. https://doi.org/10.34973/5rpw-rn92

Zaid Zada, Samuel A.Nastase, Bobbi Aubrey, Itamar Jalon, Ariel Goldstein, Sebastian Michelmann, Haocheng Wang, Liat Hasenfratz, Werner Doyle, Daniel Friedman, *et al.* (2025) The "Podcast" ECoG dataset for modeling neural activity during natural story listening. OpenNeuro. ID ds005574 https://openneuro.org/datasets/ds005574

Peer reviews

Reviewer #1 (Public review):

The manuscript analyzes previously published MEG and ECoG datasets to examine pre-onset neural encoding effects during language processing, replicating effects that have been reported in earlier work and demonstrating that they persist even after controlling for correlations in the stimulus sequence. Replication of these effects across recording modalities and datasets is a valuable contribution, as it strengthens confidence in the robustness of anticipatory neural activity related to upcoming linguistic input. However, I have significant concerns regarding the interpretation of these findings, particularly the conclusion that the absence of temporal generalization between pre- and post-onset activity implies that pre-onset activity does not reflect predictive pre-activation of the upcoming word.

The central inferential step in this argument relies on an implicit assumption: that if the brain were predicting an upcoming word, the neural representation prior to word onset should resemble, or generalize to, the representation observed after word onset. This assumption is not theoretically necessary and is not supported by a substantial body of work on predictive processing. Many contemporary models posit that predictions are represented in abstract, compressed, or probabilistic formats that differ from sensory-evoked representations, particularly in hierarchical systems such as language (e.g., Rao & Ballard, 1999; Friston, 2005; Federmeier, 2007; Kuperberg & Jaeger, 2016; de Lange et al., 2018). Under such accounts, predictive representations may encode expectations over latent semantic features or probability distributions rather than reinstating the neural code associated with perceptual input.

In this context, the temporal generalization analyses presented here convincingly demonstrate that pre-onset and post-onset activity do not share a stable representational code. However, this result does not rule out predictive processing per se. Rather, it rules out a specific and relatively strong hypothesis: that prediction takes the form of early reinstatement of the same neural representation used during post-onset word processing. The data are equally consistent with the interpretation that pre-onset activity reflects predictive information expressed in a different representational format that is transformed upon stimulus onset.

I therefore recommend that the authors substantially soften and clarify their conclusions regarding prediction. Statements suggesting that pre-onset activity does not reflect prediction should be revised to more precisely reflect what is directly supported by the analyses, namely, the absence of representational identity or stable overlap between pre- and post-onset activity. Explicit acknowledgement of alternative interpretations grounded in established predictive processing frameworks would improve theoretical alignment and avoid overstating the implications of the temporal generalization results.

Overall, the empirical analyses are carefully executed, and the replication across datasets is a strength. However, the current framing risks over-interpreting what the data can rule out about prediction. A clearer distinction between representational equivalence and predictive processing would significantly strengthen the manuscript's theoretical contribution.

https://doi.org/10.7554/eLife.110085.1.sa3

Reviewer #2 (Public review):

Summary:

The authors show that pre-onset neural encoding is likely not a product of predictive processing. They demonstrate this primarily through two analyses:

(1) They decorrelate the neural responses between pre- and post-word onset and show that this does not eliminate pre-onset neural encoding. This suggests that this pre-onset neural encoding is not a result of pre-activation driven by an underlying predictive process.

(2) They show that the future word improvement to encoding performance shown in Caucheteux et al. is likely a result deriving from the low temporal resolution in fMRI, as it does not reproduce in MEG or ECoG data, modalities that have a higher temporal resolution better suited to this kind of analysis.

Strengths:

Both of the paper's arguments are overall very compelling and point to potential problems in the underlying literature that may require reevaluation. The paper does not make any unreasonable claims. The limitations of the study are appropriately addressed. The paper is well-reasoned and well-written. Overall, I believe the paper is a worthy addition to the literature on this subject.

Weaknesses:

One concern is that I wonder about the degree to which the residualization/decorrelation that the authors employ in Figure 4 is truly forcing the model to unlearn all the interactions between pre- and post-word onset when referencing the neural activity. This point is explicitly noted in Schonmann et al. (which the authors cite): "While residualised word embeddings no longer contain temporal stimulus dependencies, these dependencies are still represented in the neural data, and can hence be 're-learned' when fitting the regression model." I imagine the inverse of this could be true here - the dependencies are still represented in the stimulus and so can be relearned when mapping to the neural data. It is possible that the small positive onset correlation that occurs after decorrelation can be entirely explained by this. This is not a bad thing per se (as it aligns with the overall point of the article), but it is a potential methodological oversight. A clear description of the decorrelation process is necessary in the methods section.

The paper correctly notes that their removal of bigram/n-gram information does not entirely exclude all stimulus dependencies. However, removing this fully would be extremely difficult, and the small reduction in performance of the bigram-ablated model does not point to this being a major issue.

Separately, some of the figures are a little rough. Suggestions have been provided to the authors.

https://doi.org/10.7554/eLife.110085.1.sa2

Reviewer #3 (Public review):

Previous studies have shown that language model embeddings of future words can predict brain responses to language. This has been interpreted as evidence for predictive representations in the brain. The primary finding of the present study is that this index of predictive processing is not consistent with a pre-activation account, but instead suggests continuously evolving representations. A strength of the manuscript is that it uses methods that build on previous studies and shows that previous results replicate in the current datasets, before testing new hypotheses. Addressing some minor weaknesses would further strengthen the results and ascertains that the conclusions are justified:

(1) When analyzing neural data, "words with multiple tokens assigned by the model were excluded" (11). I am wondering whether this could have had an influence on the results. I suspect that using only single token words would bias the dataset towards semantically light high frequency and function words. Pre-activation may be different for those words from more semantically rich, longer words.

(2) The study only used a context window of 50 tokens for language model predictions (11). This is less than in previous studies, and may constitute a confound when comparing results across studies. This may be particularly relevant in comparison to Caucheteux et al. (2003), whose results suggested more extensive predictions (9), which may require more extensive context.

(3) The manuscript is largely missing data on the reliability of the results. Some form of significance test, and indication of variability and/or the noise floor in the figures would be helpful.

A primary concern when analyzing naturalistic speech data is that different speech features are highly correlated across linguistic levels and across time. The manuscript makes a reasonable effort to control for stimulus autocorrelations. It is encouraging that the effect survived this correction. As the manuscript concedes, control is not perfect and controlling for "all regularities inherent to natural speech" remains a challenge (9). This should be kept in mind when interpreting the results.

Finally, the manuscript also argues that "we observed clear signatures of postdiction, with neural activity reflecting persistent encoding of prior words" (abstract). I did not follow this reasoning. The ostensible evidence for this is that "including the previous word ... improves encoding even after the current word's onset" (Figure 5). However, this is not further surprising, because the previous word can often only be recognized around the end of the word, corresponding to the time of the current word onset. Language model embeddings reflect a contextual semantic interpretation of the word, which likely requires further processing after word recognition. I would thus expect that the initial contextual interpretation of a word occurs during presentation of the subsequent word. Evidence for "persistent encoding" should include encoding beyond this point, i.e., over the course of several subsequent words. Contrary to this, Figure 5 a (left) suggests that the predictive effect of the previous word (*d-1*) stops around the offset of the current word (*d*). This suggests to me that, once controlling for subsequent embeddings, the embedding of a word disappears from the neural activity soon after word recognition.

https://doi.org/10.7554/eLife.110085.1.sa1

Author response:

Reviewer 1:

We thank the reviewer for bringing a critical theoretical distinction to our attention. We agree that the Temporal Generalization (TG) results specifically rule out the reinstatement of post-onset neural codes, the idea that the brain pre-activates the same neural representation evoked by the stimulus. In fact, we mention in the discussion: "This temporal variability underscores the need for a more nuanced view of what constitutes predictive pre-activation, as no stable representational state appears to persist after word presentation that could serve as its target.".

To our understanding, prediction is rarely explicitly defined in the literature, and the distinction between predictive pre-activation and other forms of prediction is seldom made. Moreover, the idea of compressed or abstract forms of pre-activated representations has not, to our knowledge, been explicitly articulated in the literature. Our TG findings therefore, put meaningful constraints on theories of prediction. In the revisions we will expand on this more and include a broader description of potential forms of pre-activation. We will emphasize that the TG results specifically rule out that the brain pre-activates the same neural code used for sensory-evoked processing.

Moreover, although TG analysis does not rule out alternative notions of predictive pre-activation, we believe our second analysis (the inclusion of future word embeddings) provides independent evidence that argues against more abstract forms of prediction. Unlike the TG analysis, this encoding approach is not constrained to a specific neural code; if the brain represented upcoming words in any linearizable format (abstract, probabilistic, or latent) incorporating those embeddings should have improved the brain score at the current word's onset. We found no such improvement until the word was actually heard. In the revised manuscript, we will reformulate the narrative to clarify that while TG alone rejects a specific form of pre-activation, the combined evidence from both analyses suggests there is a broader lack of predictive pre-activation.

Reviewer 2:

We thank the reviewer for their constructive feedback and for bringing to our attention the missing information in our Methods section. We realized that the final two sections were inadvertently omitted during formatting changes before submission. These will be restored in the revised version.

We appreciate the reviewer's careful reading of this analysis and agree that the concern whether the decorrelation in figure 4 forces the model to unlearn the associations between pre- and post-onset activity is a valid one. To clarify, this is not what we intended to claim. Rather, our argument follows a different logic: if we assume that pre-onset encoding is purely a signature of predictive pre-activation, then decorrelating the pre- and post-onset brain responses should effectively remove that signature. The fact that pre-onset encoding remains largely intact after this procedure suggests that our initial premise was false; the observed pre-onset encoding is likely not a signature of pre-activation. We would also like to note that in this analysis, we use both residualized neural data and we use decorrelated embeddings. Therefore, the majority of stimulus dependencies are removed. Nevertheless, as the reviewer notes, some dependencies such as bi-grams and other word-co-occurrences, inevitably remain. These dependencies might explain the remaining pre-onset encoding we observed. This aligns with our main message of the paper. In the revisions, will provide a detailed description of the decorrelation process and we will make this interpretive logic more explicit in the main text.

Reviewer 3:

We are grateful for the reviewer’s detailed comments and for raising several points that will significantly improve the clarity and comparability of our study. Specifically, the reviewer’s feedback helped us realize that our evidence for postdiction required further clarification. While the encoding of the immediate preceding word ($d-1$) may involve recognition lags, we observe that word $d-2$ further improves the brain score even after the current word's onset, beyond what is explained by word $d-1$ alone. This may extend beyond simple recognition delays. To address this we will visualize this effect further in the upcoming version and expand the manuscript to include alternative explanations for this observation, such as extended lexical processing or integration delays.

To ensure our results are not biased toward high-frequency or function words, we will re-run our analyses including multi-token words. Given that these words constitute a small part of the datasets, we expect our core findings to remain stable.

In line with our response to reviewer 2, we will more clearly emphasize that despite our extensive controls, we cannot be sure that we accounted for all regularities inherent to natural speech.

Additionally, we will increase the context windows of the LLM to match the larger windows used in previous literature and add significance tests, error bars, and noise floor indications to our figures to ensure the reliability and variability of our findings are clearly communicated.

https://doi.org/10.7554/eLife.110085.1.sa0