Stimulus dependencies—rather than next-word prediction—can explain pre-onset brain encoding during natural listening

Inés Schönmann[1*], Jakub Szewczyk[1], Floris P de Lange[1], Micha Heilbron[1,2]

***For correspondence: ines.schoenmann@gmail.com (IS)**

[1]Donders Institute for Brain, Cognition and Behaviour; [2]Amsterdam Brain and Cognition, University of Amsterdam

Abstract The human brain is thought to constantly predict future words during language processing. Recently, a new approach to investigating linguistic predictions emerged which aims to capture predictive pre-activation directly by using neural network representations of words to predict brain activity prior to word onset. However, it is unclear what exactly is driving the predictability of pre-stimulus brain activity. Here we show, across two datasets, that both proposed hallmarks of neural pre-activation—i.e. (i) pre-onset brain response predictability and (ii) its modulation by word expectedness—is not only observed in brain responses, but also in representations of the stimulus material itself. We show that various structural and incidental dependencies existing in natural language can explain previously reported hallmarks of prediction without assuming any pre-activation in the neural data. This suggests that pre-onset prediction of brain activity might only reflect dependencies within the stimulus material rather than predictive computations, and questions the extent to which this new prediction-based method can be used to study prediction in the brain.

Introduction

In the past years, the field of natural language processing (NLP) has made great advances in developing computational systems that can generate, classify and interpret language. Much of this progress has been driven by large language models (LLMs): neural networks trained in a self-supervised manner to predict the next word or token (***Minaee et al., 2024***). Surprisingly, this simple training objective is sufficient for models to learn about language more broadly, making models develop a human-like knowledge of syntax (***Manning et al., 2020***; ***Linzen and Baroni, 2021***) and enabling them to solve almost any NLP task (***Minaee et al., 2024***; ***Manning, 2022***). Furthermore, internal model representations ("embeddings") also constitute the highest performing encoding models for predicting brain responses to linguistic stimuli (***Caucheteux and King, 2022***; ***Jain and Huth, 2018***; ***Schrimpf et al., 2021***)—indicating that LLMs' internal representations might capture some aspects of human language representations (***Tuckute et al., 2024***).

Two lines of research suggest that predicting the upcoming linguistic input also plays an important role in *human* language comprehension. The first line of research focuses on neural and behavioural responses occurring *after* the onset of the stimulus in question. The reasoning being that if the brain is engaged in predicting upcoming linguistic input, brain responses and reading times should vary as a function of linguis-

tic predictability (***Smith and Levy, 2013***; ***Frank et al., 2015***; ***Willems et al., 2016***; ***Shain et al., 2024***; ***Heilbron et al., 2023***). Many studies following this line of reasoning have demonstrated that both neural responses and reading times are sensitive to even subtle fluctuations in predictability at several levels of linguistic analysis, e.g. phonemes, words or semantics (***Boston et al., 2011***; ***Brodbeck et al., 2022***; ***Heilbron et al., 2022***; ***Smith and Levy, 2013***; ***Szewczyk and Federmeier, 2022***). This approach can be seen as an extension of a long-standing tradition of research into linguistic expectations which relied on carefully constructed sentences that violate linguistic expectations and use cloze probabilities to quantify unexpectedness (***Kutas and Hillyard, 1980a***,b, ***1984***).

Recently, a new, alternative approach emerged that aims to probe linguistic predictions directly by capturing their neural signature *prior* to the onset of a word (***Wang et al., 2018***; ***Goldstein et al., 2022b***). Predicting a word is thought to involve a pre-activation of the representation of that word. Hence, finding a trace of a representation of a word in the neural signal *prior* to its onset, is interpreted as direct evidence for a word's pre-activation, and therefore, next-word prediction. Capturing the neural signature of the prediction itself is appealing, since it potentially circumvents interpretational challenges of more indirect, post-stimulus predictability effects which have been suggested by some to reflect related but distinct downstream processes—such as semantic integration difficulty or 'post-diction'—rather than prediction *per se* (***Pickering and Gambi, 2018***; ***Huettig, 2015***; ***Huettig and Mani, 2016***). Perhaps the most influential demonstration of predictive pre-activation during language comprehension was presented by ***Goldstein et al.*** (***2022b***). Using both encoding (predicting brain responses to a stimulus based on a model's representation of that stimulus) and decoding analyses (determining the model representation of a stimulus from the brain's response to that stimulus) on electrocortographic (ECoG) recordings of participants listening to naturalistic narratives, the authors showed (i) that brain responses could be predicted significantly better than chance using word embeddings of the upcoming word, as early as two seconds prior to onset. Furthermore, they showed (ii) that this effect was modulated by word predictability: words that were highly predictable in context showed higher pre-stimulus brain-prediction performance—in line with the idea that their representation was pre-activated more strongly. More recently, a similar pattern of results was found in non-invasive MEG recordings (***Azizpour et al., 2024***).

Here, we build on these results and ask to what extent, and under which conditions, these two proposed hallmarks of prediction can unequivocally be interpreted as reflecting neural pre-activation. This question is motivated by the fact that, in natural language, some inherent correlations or dependencies exist between neighbouring words. Often they share semantic aspects—as in the sequences "pine tree" or "green leaves"—or (morpho)syntactic information, as in the phrase "he goes" in which both words indicate the use of the third person singular. At other times dependencies result from words that frequently co-occur in a discourse, like "Sherlock Holmes". Irrespective of the exact nature of the shared information between neighbouring words, natural language is full of such structural and incidental regularities. A priori, then, these inherent *stimulus dependencies* might already explain why representations of future words can predict brain responses to preceding words, without having to assume an additional pre-activation by the brain. The goal of this study is two-fold: (i) to assess to what extent the hallmarks introduced by ***Goldstein et al.*** (***2022b***) might be explained by stimulus dependencies, and (ii) to explore potential methods to correct for this alternative explanation, and propose a more stringent test for pre-activation within the encoding modelling framework introduced by ***Goldstein et al.*** (***2022b***).

To foreshadow our findings, we show here that ostensible hallmarks of next-word prediction based on the analysis of neural activity before word-onset are also present in the stimulus material itself, namely when applying the same analysis to word embeddings and acoustic representations of the stimuli. Furthermore, we find that these dependencies can be corrected for only within the same, but not between

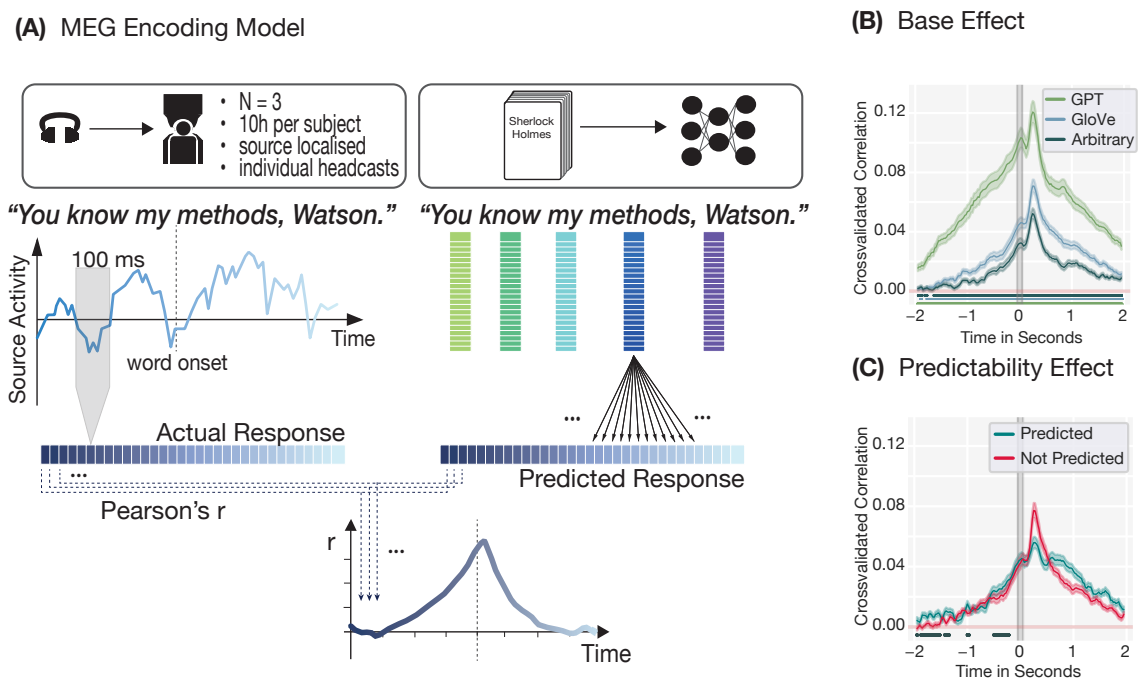

Figure 1. A) MEG Encoding Model. MEG data was epoched to word onset and averaged over a sliding window of $100ms$, moving with a step size of $25ms$. The model representation (GPT-2, GloVe or arbitrary) of the word at $t = 0$ was then used to predict the neural response for each channel and time point in a separate cross-validated Ridge regression. The actual and predicted response were then correlated time point by time point, resulting in a time-resolved encoding plot. B) Positive pre-onset encoding (subject 1) for GPT-2 (green), GloVe (blue) and arbitrary (grey) embeddings shows that it possible to find ostensible neural signatures of pre-activation in MEG data. Lines show clusters of time points that are significantly different from zero ($p < 0.05$ under the permutation distribution). C) Encoding using GloVe embeddings demonstrate a slight advantage of the predictability of a word (top-1 prediction by GPT-2-XL) for pre-onset encoding. Line indicates clusters of time points prior to word onset during which predictable words are significantly better encoded ($p < 0.05$ under the permutation distribution).

representational spaces—a necessity for encoding neural data from model representations. We conclude that pre-onset encoding results simply reflect correlational patterns found in naturalistic language. This poses a serious challenge to the direct approach of probing linguistic pre-activations prior to word onset through encoding modelling, and questions to what extent such findings can demonstrate that brains, like LLMs, actually perform next-word prediction.

Results

Hallmarks of prediction replicate in MEG data

Drawing on two different, publicly available MEG datasets in which participants listened to narratives, we analysed the data following the approach put forth by ***Goldstein et al.*** (***2022b***) and ***Azizpour et al.*** (***2024***). Hence, MEG data were epoched with respect to the onset of each word between $-2s$ and $+2s$, and brain activity was averaged over a sliding window of $100ms$. Subsequently, the word representation (word embedding) corresponding to the word to which each epoch was time-locked, was used to predict brain activity within the epoch, i.e., within the time window of $-2s$ to $+2s$ (see Figure 1a). Correlating the predicted with the actual brain response, results in a time-resolved prediction accuracy curve which allows us to test the two hallmarks of prediction proposed by ***Goldstein et al.*** (***2022b***), namely i) whether brain activity prior to the onset of a word can be predicted from that word's embedding and ii) whether pre-onset prediction accuracy is higher for predictable words than for unpredictable ones. These two aspects are proposed as

evidence of prediction, since pre-onset encoding is thought to capture the pre-activation of a word's neural representation.

Using representations from three different models (GPT-2, GloVe and arbitrary 300-dimensional word embeddings) to encode brain activity, we found that it is possible to replicate both hallmarks of predictions described by ***Goldstein et al.*** (***2022b***). We found both i) positive encoding prior to word onset for all three models (see Figure 1B) and ii) a slight encoding advantage prior to word onset for highly predictable (i.e., GPT-2's top-1 prediction) as opposed to less predictable words (see Figure 1C). For subject 1 this advantage started as early as $575ms$ prior to word onset, and predictable words lead to a continuously higher encoding performance ($M = .004$, $SD = .001$ for significant time points, all *p's* < 0.031, TFCE FWE-corrected), corresponding to an average improvement in encoding performance within this pre-onset time window of 17% with respect to unpredictable words (with improvements of $M = .006$ ($SD = .002$, all *p's* < 0.035) corresponding to 25% starting $525ms$ prior to onset, and $M = .011$ ($SD = .006$, all *p's* < 0.033) corresponding to 46% during the time window of $1\,400 - 300ms$ prior to word onset for subject 2 and 3, respectively). In accordance with previous findings, using GPT-2's contextualised word representations allowed for earlier and better brain encoding than using non-contextualised GloVe embeddings (***Goldstein et al., 2022b***; ***Schrimpf et al., 2020***). In turn, these non-contextualized word embeddings predicted brain response better than arbitrary, 300-dimensional vectors. However, arbitrary embeddings still performed remarkably well, given that they contain no structured information besides the identity of a word (see Discussion).

These results replicated for all three participants (see Figure S1A and S1B) in the few-subject dataset, indicating the robustness of the results given sufficient amounts of data (see Figure S1C for results from a more conventional multi-subject dataset). This demonstrates that pre-onset encoding is a tremendously robust phenomenon, replicating across different word embeddings (GPT-2, GloVe and even arbitrary embeddings), data sets (single-subject as well as multi-subject) and even different forms of MEG spaces (source as well as sensor data).

Stimulus dependencies mirror brain encoding results

In order to assess to what extent brain encoding results prior to word onset might be driven by dependencies between neighbouring words in the dataset, we propose a *self-predictability analysis* in which we use the same word-embedding-based encoding model in order to predict word embeddings of preceding words (see Figure 2A). Thus, instead of using word embeddings to predict brain activity prior to word onset, the encoding model is using word embeddings to predict *preceding word embeddings*, prior to word onset. This was done separately for each type of word embedding: GPT-2's word representations were used to predict preceding GPT-2 embeddings, GloVe embeddings were used to predict preceding GloVe embeddings and arbitrary word embeddings were used to predict preceding arbitrary embeddings. We refer to the encoding performance of each model's representations as its *self-predictability* (see Figure 2A).

Strikingly, this approach revealed that—for each type of word embedding—self-predictability curves exhibited the same two hallmarks of prediction as observed in the neural data. In line with the first hallmark, we found significant self-predictability prior to word onset for all three models (see Figure 2B and S2A), with highest self-predictability for GPT-2 and lowest for arbitrary embeddings. Furthermore, and in line with the second hallmark, models exhibited higher self-predictability in epochs time-locked to a highly predictable word as opposed to those time-locked to less predictable words (see Figure 2C and S2B and C). Additionally, in the time window prior to word onset, self-predictability showed a similar time course and similar differences between models as observed in the brain encoding (see Figure 2B and S2A). Hence, in naturalistic texts, both ostensible hallmarks of prediction—encoding prior to word-onset and sensitivity of pre-onset encoding to word predictability—can be the result of the correlational structure of neighbouring model representations. Therefore, these proposed hallmarks do not necessarily reflect a neural pre-activation of the

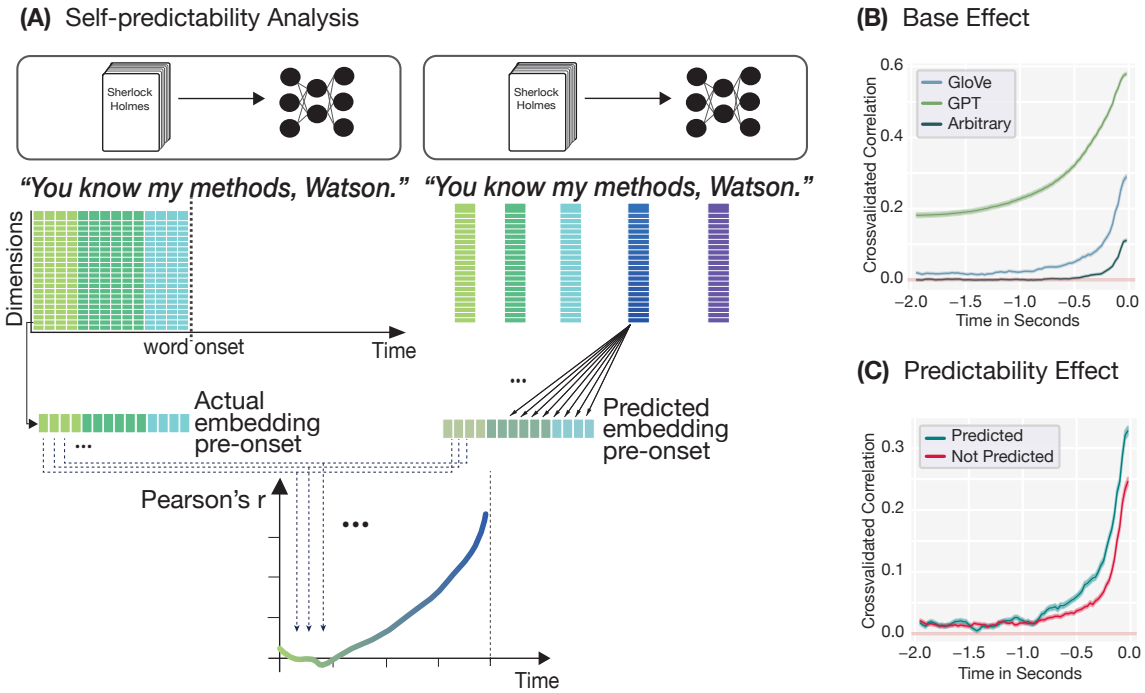

Figure 2. A) Self-predictability of Model Representations. In order to assess the predictability of neighbouring model representations in each MEG epoch, we replaced the neural data at each time point with the model representation of the word presented during that time point. The model representation of the word at $t = 0$ was then used to predict the previous model representations for each dimension and time point in a separate cross-validated Ridge regression. The actual and predicted values were then correlated time point by time point, resulting in a time-resolved self-predictability plot. **B)** Self-predictability for GPT-2(green), GloVe (blue) and arbitrary (grey) embeddings prior to word onset mirrors patterns observed when encoding the neural data. **C)** Self-predictability of GloVe embeddings demonstrate a clear advantage of the predictability of a word (top-1 prediction by GPT-2-XL), and therefore the same qualitative difference as observed when encoding neural data.

representation of that word. Instead, they can, in principle, be explained through stimulus dependencies alone, without assuming any predictive pre-activations in the neural data.

Removing stimulus dependencies from word embeddings does not remove stimulus dependencies across representational domains

Given that these stimulus dependencies *within* word embeddings can be quantified so straightforwardly, we then asked whether they can be removed using a similar approach, in order to test for pre-activation while correcting for stimulus dependencies. A general and powerful approach is to use residualisation or "regressing out". For instance, in the sentence "You know my methods.", "You" can be regressed out of "know", "know" out of "my", and "my" out of "methods" (see Figure 3A). Note that this is a generalised version of a control analysis performed by ***Goldstein et al.*** (***2022b***), which removed neighbouring words from word embeddings through projection. When we residualised the embeddings in this way and then repeated the self-predictability analysis with these residualised embeddings, results confirmed that this procedure reduces pre-onset self-predictability to zero—suggesting that this may be an easy and effective way to remove stimulus dependencies from the word embeddings themselves (see Figure 3B).

However, one important limitation of this approach lies in the fact that it only accounts for the correlations between neighbouring word embeddings within this specific representational space, and not for temporal stimulus dependencies more broadly. When encoding brain activity, however, we are perform-

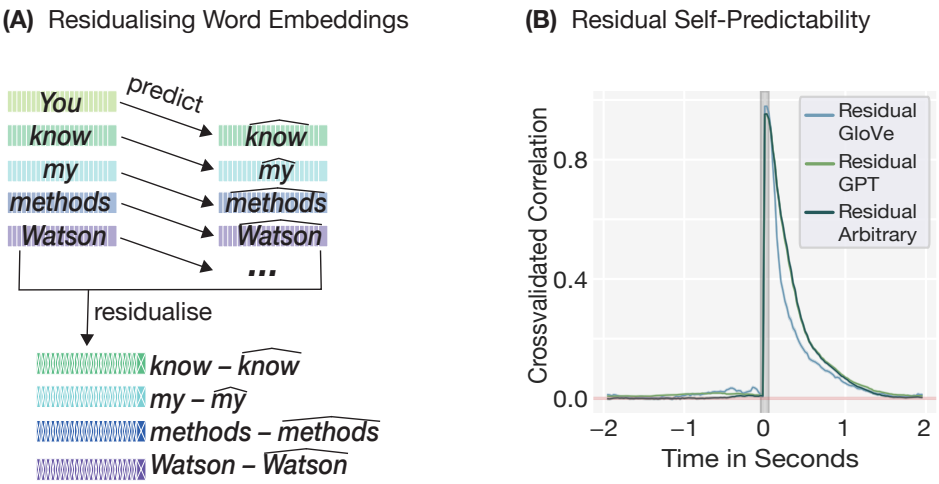

Figure 3. A) In order to remove shared information between a word and its predecessor in the text, we residualised word embeddings by first fitting an OLS regression to predict the next word based on the previous word's embedding, i.e. predicting "know" based on "You". This resulted in a predicted embedding $\hat{x}$, e.g. "$\widehat{know}$", which contained the shared information between the two words. Finally, the predicted embedding, e.g. "$\widehat{know}$", was removed from the original embedding, e.g. "know", to generate word representations for which the dependency between neighbouring words was removed. **B)** Self-predictability after regressing out the previous embedding from the embedding at $t = 0$ shows that it is possible to successfully remove the correlations between neighbouring model representations. For brain encoding results when using these residualisded embeddings, see Figure S3.

ing a mapping *between* spaces, i.e. from word embeddings to the neural response. 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. Therefore, it may be insufficient to remove the temporal dependencies in only one specific representational format, i.e. the embedding space, in order to account for *all* the shared information between neighbouring words that may be exploited by the regression-based encoding analysis. Hence, in order to show that removing self-predictability can account for all stimulus dependencies, one would need to show that it is no longer possible to encode the stimulus itself prior to its onset on the basis of residualised embeddings.

To evaluate this possibility, we tested whether residualised word embeddings can be used to predict a *different* representation of the stimulus material prior to onset—namely the speech acoustics. Hence, instead of predicting the word embeddings or the brain signal, we used the residualised embeddings to predict the speech acoustics the participants listened to. We chose the acoustics as they represent a straightforward format to quantify the 'predictability of the stimulus itself', and since they are known to have a strong impact on the neural signal. Since this analysis involves a mapping between spaces (from embeddings to acoustics) it is more similar to the brain encoding analysis (mapping from embeddings to brain activity), and therefore, can be used to test whether operations that modify the embeddings are sufficient to correct for stimulus dependencies more broadly in brain encoding analyses.

Using this cross-representational approach, we find that the same two hallmarks of prediction observed in the neural data and in the self-predictability of our models, are also present in the acoustics of our speech signal. Worryingly, we find this even when using *residualised* word embeddings—i.e., word embeddings for which information about neighbouring word embeddings was *removed*. Specifically, this means that, just like the neural signal, the acoustic signal itself exhibits temporal dependencies that allow for i) significant

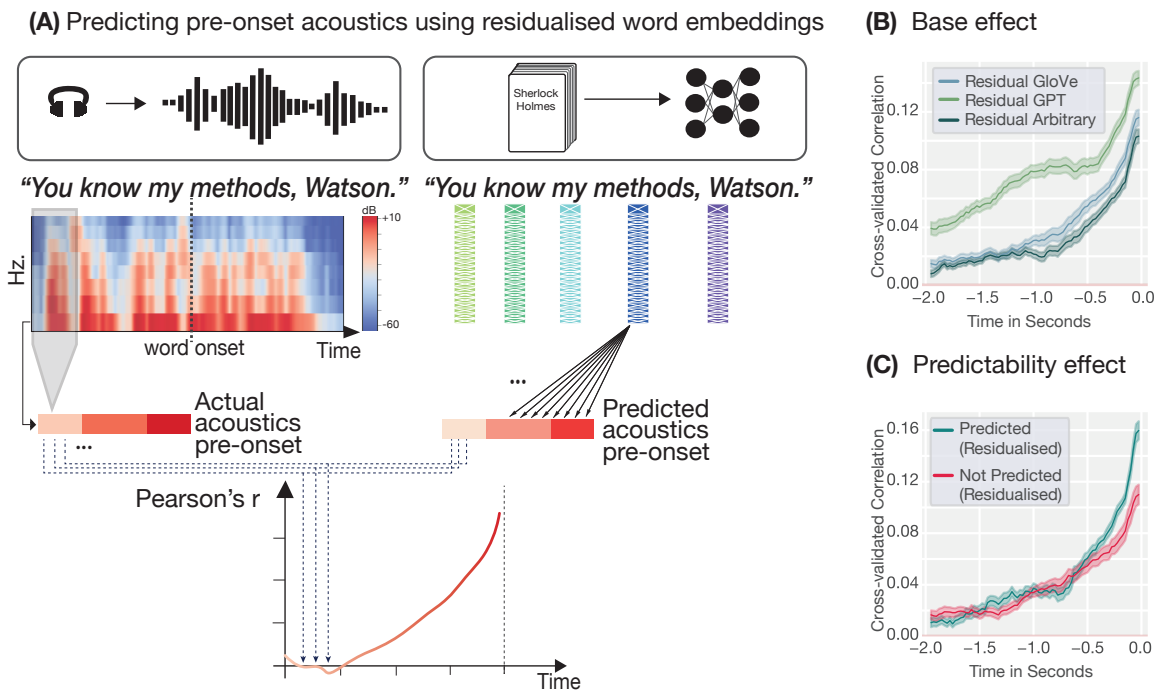

Figure 4. A) To test whether removing self-predictability (as in Fig.3) can correct for stimulus dependencies more generally, we investigated the predictability of pre-onset acoustics using residualised word embeddings. We computed an 8-Mel spectrogram and the envelope for each word and replaced the neural data at each time point with the respective acoustic representation. The residualised GPT-2, GloVe or arbitrary embedding of the word at $t = 0$ was then used to predict previous acoustic representations. The actual and predicted values were then correlated, resulting in a time-resolved correlation plot. **B)** Predictability of prior word acoustics when using residualised GPT-2 (green), GloVe (blue) and arbitrary (grey) embeddings prior to word onset. Patterns closely mirror those observed in each model's self-predictability or in residual brain encoding results. **C)** Predictability of prior word acoustics using residualised GloVe embeddings demonstrate a clear advantage of the predictability of a word (top-1 prediction by GPT-2-XL) for predicting its prior acoustic representations, and therefore, the same qualitative difference as observed when encoding neural data.

encoding prior to word onset (see Figure 4B and S4A), as well as ii) higher encoding in epochs time-locked to highly predictable words as opposed to those time-locked to less predictable words (see Figure 4C but also see S4B and C for the reversed effect). In other words, the same 'hallmarks' that are found in brain encoding studies—and that are interpreted as 'representational pre-activation'—are found in the acoustics of the stimulus material itself, and may thus, more likely, reflect the inherent predictability of the stimulus material. The fact that this effect persists with residualised embeddings shows that adjusting the embeddings is insufficient to fully correct for stimulus dependencies within brain encoding analyses. Hence, even when using modified or 'corrected' word embeddings (see Figure S3), these ostensible hallmarks of next-word prediction may be due to stimulus dependencies exploited by the regression model rather than predictive pre-activation in the brain.

Discussion

We evaluated an encoding modelling paradigm that used neural network representations of words to capture neural signatures of linguistic pre-activations during naturalistic listening across two publicly available MEG datasets. Specifically, we asked to what extent, and under which conditions, two proposed hallmarks of prediction—namely (i) positive encoding performance prior to word onset, and (ii) sensitivity to the predictability of a word—can be interpreted as reflecting neural pre-activation. We found that the two proposed

hallmarks could not only be found in the neural data but also in the stimulus material itself, specifically in the acoustics and the various word embeddings. These results demonstrate that such effects can be explained by correlations between neighbouring word representations alone, without assuming any predictive pre-activations in the neural data. We further showed that trying to account for these dependencies through adjusting (residualising) word embeddings, remains insufficient when performing a mapping between representational spaces, as is the case in any brain encoding paradigm. Together, our results demonstrate that various structural and incidental correlations existing in natural language can explain ostensible hallmarks of prediction when using a neural encoding approach. Accounting for these correlations poses a non-trivial problem in naturalistic designs.

We observed that the first hallmark of prediction—pre-onset encoding—is tremendously robust, replicating not only in non-invasive MEG data (which has a lower signal-to-noise ratio compared to the ECoG data used by ***Goldstein et al.*** (***2022b***)), but also across various types of word embeddings (GPT-2, GloVe and even arbitrary embeddings), data sets (single-subject as well as multi-subject), MEG spaces (source as well as sensor data) and type of linguistic representation (neural, artificial or acoustic). By contrast, the second hallmark—i.e., sensitivity of pre-onset encoding to next-word predictability—could only be replicated reliably in the few-subject dataset (Figure 1C, 3C, S1, S3), which is the same dataset analysed by ***Azizpour et al.*** (***2024***). We see two potential explanations for this discrepancy. First, it may be simply because the single-subject dataset comprised ten times more data per subject than the multi-subject dataset. Alternatively, it could reflect a difference in the experimental designs: the multi-subject dataset did not use natural speech but carefully manipulated computer-generated speech, to minimise acoustic confounds such as co-articulation. Indeed, the second hallmark was not just absent in the neural data, but critically also in the acoustics of the multi-subject dataset (see Figure S4 and S5). Hence, in both datasets, neural encoding closely mirrored acoustic encoding results, suggesting that ostensible hallmarks of prediction reflect the correlation structure of the stimulus material rather than neural pre-activation in the brain.

Another important finding of the present paper is the difficulty of removing or correcting for such correlations in the stimulus material. Natural language is rife with temporal structure that is useful for predicting neighbouring words—whether such structure may be semantic, syntactic, acoustic, or due to n-gram statistics. While the self-predictability analysis (Fig. 2) showed that removing neighbouring words within one representational format can be achieved through residualisation, brain encoding involves two representational spaces: the neural and the word embedding space. Removing temporal correlations from only one of these two spaces, still allows for the regression model to capitalise on the regularities and correlations in the second representational space. This is exemplified by our embedding-based acoustic analysis which demonstrated that even residualised word embeddings can be used to predict the acoustics of preceding words. As a result, removing neighbouring words from the word embeddings is not sufficient to correct for the inherent predictabilities in the stimulus material that can be learnt by the regression model.

Critically, we do not want to suggest that our results question the role of prediction during language processing itself. Indeed, there is a large body of work suggesting that human language processing is inherently predictive. For instance, readers and listeners are highly sensitive to even subtle fluctuations in linguistic predictability (***Smith and Levy, 2013***; ***Frank et al., 2015***; ***Willems et al., 2016***; ***Shain et al., 2024***; ***Heilbron et al., 2023***; ***Brodbeck et al., 2022***; ***Heilbron et al., 2022***; ***Szewczyk and Federmeier, 2022***). However, such predictability effects on brain responses to language are by definition post-stimulus (and hence indirect), while pre-stimulus evidence has, for some researchers, been considered the 'gold standard' in evidence for linguistic prediction (***Kuperberg and Jaeger, 2016***; ***Pickering and Gambi, 2018***; ***Nieuwland, 2019***). On first consideration, encoding modelling provides a new and more direct line of evidence that can asess pre-stimulus prediction (***Schrimpf et al., 2021***; ***Goldstein et al., 2022b***; ***Caucheteux et al., 2023***; ***Azizpour et al.,***

2024). However, due to the opacity of the word embeddings and regression models involved, interpreting these results as evidence for prediction in the brain is challenging (***Antonello and Huth, 2024***; ***Azizpour et al., 2024***), rendering the evidence less direct than it may initially appear, even when it concerns evidence of pre-stimulus brain activity.

While these results taken together pose challenges to using pre-stimulus brain encoding to test for neural pre-activation, we yet see two possible future applications to use this analytical framework for this purpose. First, the predictability of the stimulus could be used as a threshold or benchmark: if brain encoding prior to word onset is quantitatively higher than the predictability of the stimulus, this might indicate that a predictive representation adds to the encoding performance stemming from temporal correlations alone. Though this is not the case in this current study, MEG is limited in terms of signal-to-noise ratio, and less noisy data, such as ECoG, might be able to fulfil this criterion. Secondly, stimulus predictability could be used as a tool in order to pre-select trials in which stimulus correlations do not allow for pre-onset encoding or favour a different trend (as in the case of our multi-subject dataset).

Encoding-based approaches represent an enticing new approach to studying linguistic processing. However, their reliance on regression-based prediction of brain activity renders it, paradoxically, difficult to use encoding to test for predictions in brain activity. Namely, since the same regularities that the brain may use to predict natural language can also be exploited by the regression model, it is ultimately difficult to know which system is performing the prediction—the brain or the regression model used by the brain scientist.

Methods

Data

To test whether previously observed evidence for encoding a word's pre-activation in neural data can be can be replicated in non-invasive MEG data, we used a publicly available, high quality MEG data set (***Armeni et al., 2022***) in which three subjects listened to the audiobook version of the entire Sherlock Holmes corpus in ten one-hour-long recording sessions. To minimise noise, head motion was restricted using individual 3D-printed head casts (for details on the data set and stimuli see ***Armeni et al.***, 2022). Brain responses were source localised and minimally filtered between $0.1 - 40$Hz. For the analysis, the MEG data was time-locked to word onset from $-2000ms$ to $+2000ms$ without applying any baseline correction. Subsequently—akin to the analysis performed by ***Goldstein et al.***—neural activity was averaged over a sliding window of $100ms$ which moved with a step size of $25ms$. This resulted in $85\,719$ trials of 157 time points within the window of $4s$.

To examine whether our findings replicate in a more conventional multi-subject data set, we repeated our analyses in another publicly available MEG data set with 27 participants (see ***Gwilliams et al.***, 2022 for more details). Crucially, the data set differed in several important aspects from the main data set used in this analysis. Firstly, participants listened to 4 different stories within their one hour-long recording session, resulting in a total of $7\,745$ trials per participant. Secondly, listening was interrupted by a word list or question every 3 minutes. Thirdly, the speech rate varied every $5 - 20$ sentences between 145 and 205 words per minute and silences between sentences varied from $0 - 1000ms$. And lastly, MEG data was not source-localised, but all analyses were performed on channel data.

MEG Encoding Model

The neural response to any given word was predicted based on word embeddings by means of a ridge regression—separately for each source and time point—within a 10-fold cross-validation. Within each fold, the features of the models were standard scaled before running the regression. For each time point, the predicted response was then correlated with the actual response (see Figure 1 A for a visualisation). As features in the regression model, embeddings were obtained for each word.

For the non-contextualised analysis, we used 300-dimensional GloVe vectors (***Pennington et al., 2014***) obtained from the *spacy* package, version $3.4.3$ (***Honnibal and Montani, 2017***). For the contextualised analysis, 768-dimensional word embeddings were extracted from the 8th layer of Huggingface's (version $4.23.1$) pre-trained GPT-2-S (***Radford et al., 2019***), as middle layers have been shown to result in the best brain encoding performance (***Goldstein et al., 2022a***; ***Caucheteux and King, 2020***; ***Schrimpf et al., 2020***). For words which consisted of multiple byte-pair encoded tokens (BPEs), such as "Sherlock" which is broken down into "Sher-lock", the embedding of the last BPE was used. For computational reasons, the contextual window for each word ranged from 512 to 1024. For the arbitrary model, 300-dimensional word-specific vectors were drawn from a Gaussian distribution ($M = 0.1$, $SD = 1.1$). Hence, arbitrary vectors contained no systematic information other than word identity.

In order to test the second hallmark, i.e., the sensitivity of the encoding performance to the predictability of a word, we split the data into easily predictable and less predictable words, ran a separate encoding model for each split and compared their encoding performance. Since GPT-2's internal word representations are a combination of previous word representations, encoding results are difficult to interpret temporally. Consequently, this analysis was performed for non-contextualised GloVe embeddings only. In order to ascertain whether the encoding performance for predictable words was significantly larger than for unpredictable words, we performed cluster-based permutation testing using threshold-free cluster enhancement (TFCE) with $10\ 000$ permutations as implemented in mne stats' permutation_cluster_1samp_test function. Differences were considered significant if computed t-value exceeded the 95-th percentile under the permutation distribution.

Words were defined as easily predictable as opposed to less predictable words based on GPT-2-XL's top-one prediction. This resulted in $27\ 035$ predicted and $58\ 684$ unpredicted words for the few-subject dataset and in $2\ 124$ predicted and $5\ 621$ unpredicted words for the few-subject dataset. Given that GPT-2-XL's top-one prediction might constitute a conservative estimate of whether or not a word was predictable in context, we repeated this analysis but defined words as easily predictable if they were among GPT-2-XL's top-five predicted words. This resulted in $50\ 666$ predicted and $35\ 053$ unpredicted words for the few-subject dataset and in $3\ 703$ predicted and $4\ 042$ unpredicted words for the few-subject dataset. Results for these different splits are shown in the supplement.

Source Selection

Given that the purpose of this study was to encode neural responses related to linguistic predictions we aimed to restrain our model to sources related to language processing. Hence, sources were selected for each subject individually according to a 2-step procedure. First, prior to the encoding modelling, we pre-selected sources based on whether they were located in the bilateral language network (see ***Heilbron et al.***, 2022). This resulted in 100 sources, which were used for the encoding model. The data matrix, therefore, had the shape $100 \times 85\ 719 \times 157$. Out of the resulting 100 encoded sources, we retained only those per subject which proved to allow for good encoding performance post word onset.

For this purpose, we determined for each subject the peak encoding performance in the post-word-onset window of $0 - 500ms$. We subsequently defined a cut-off threshold of what constitutes a "good encoding performance" at least 30% of that peak value. A source was then considered to allow for good encoding if it reached this threshold within the post-word-onset window of $0 - 500ms$. We chose this time-window for our selection process, since encoding related to the word itself (as opposed to other spurious elements in the data) should be highest while the word is perceived and processed, i.e. during a time-window when well known components such as the N400 or P600 are usually observed. Additionally, both hallmarks of

prediction concern the encoding performance prior to word onset since they are supposed to reflect an encoding of the pre-activation of the representation of a word. Hence, selecting sources based on post-onset encoding performance avoids double dipping.

Source selection was based on GloVe encoding results. This procedure resulted in 32 sources for subject 1 (max = 0.150, threshold = 0.045), 33 sources for subject 2 (max = 0.103, threshold = 0.031), 25 sources for subject 3 (max = 0.187, threshold = 0.056). To ensure that source selection was stable across neural network representations, we performed the same procedure based on GPT-2 encoding results. This resulted in fewer sources (26, 22, and 20 sources for subject $1, 2$ and 3 respectively), all of which were a subset of the sources obtained from the GloVe based selection. Hence, all analyses were performed with the GloVe based selection in order to ensure greater inclusivity.

Since analyses in the multi-subject data were performed on channel and not source-localised data, and since encoding was performed on group-level, channel selection was based on a simple common cut-off threshold. Hence, the encoding model was run on all 208 channels for each subject separately, resulting in 27 data matrices of the size of $208 \times 7\,745 \times 157$. For each participant in the dataset, channels were retained for plotting if the channel reached the threshold in the post-word-onset window of $0 - 500ms$. This threshold was selected based on our results from the few-subject analysis (threshold = 0.031). This resulted in the exclusion of subject 12 for whom no channel surpassed the threshold.

Self-Predictability Analysis

As mentioned above, due to the inherent structure present in natural language, neighbouring words can share information, and therefore, neighbouring word representations can be correlated. For instance, nouns are frequently preceded by articles or prepositions and neighbouring words belong to a similar semantic field ("pine tree", "driving a car", etc.) Hence, a positive encoding performance prior to word onset might result from neighbouring embeddings being correlated and each embedding encoding the neural representation of the corresponding word, not due to a pre-activation in the neural signal. In other words, if a word's representation x_0 and the representation of the preceding word x_{-1} are correlated ($\rho(x_{-1}, x_0) > 0$) and each representation (x_{-1}, x_0) successfully encodes it's corresponding neural activity (y_{-1}, y_0), then encoding prior to word onset is possible since x_0 can be used to approximate x_{-1}, leading to a lower, but positive encoding performance.

In order to investigate this possibility, we constructed an encoding model in which the dependent variable, i.e. the y-matrix for each trial, did not the neural data, but consisted of the embedding vectors of the words that were presented at each time point. For instance, given the sentence "You know my methods, Watson." time-locked to the onset of the word "methods", we computed the onset and offset times of each of our 157 sliding time points, determined which word was presented at that time point and filled that data point with the vector for that word. For example, if we assume that each word above had a duration of $500ms$, the 20 time points between -1500 and $-1000ms$ were filled with the vector for "You", the next 20 time points with the vector for "know" etcetera (see Figure 2A). Due to computational reasons self-predictability was only computed for the first session in the few-subject dataset, i.e. approximately 10% of the available data ($8\,622$ words). We deemed this sufficient since the stimulus material in the few-subject dataset consisted of one text corpus, namely the full Sherlock Holmes corpus, and therefore the correlational structure in 10% of the data might reasonably be representative for the whole text. Additionally, unlike the MEG data in our brain encoding model, Thus for GPT-2, the dependent variable was a $768 \times 8\,622 \times 157$-dimensional matrix, and for GloVe and arbitrary vectors it was $300 \times 8\,622 \times 157$-dimensional.

Akin to the brain encoding, modelling was performed by means of a 10-fold cross-validated ridge regression in order to predict previous embeddings from the embedding at time point zero, and correlated the pre-

dicted and actual embeddings. This regression was performed for each feature and time point separately and both the y- and the X-matrix were standard scaled. The resulting correlation will be referred to as the *self-predictability* of a model.

In order to test whether model self-predictability would also be able to account for the second proposed hallmark of prediction, namely sensitivity to predictability, we repeated the same procedure as for the neural data. To compare the self-predictability of GloVe vectors of predictable as opposed to less predictable words, we split the data again based on GPT-2-XL's top-one prediction (see supplement for results from splits based on GPT-2-XL's top-five prediction). This split resulted in $3\ 075$ correctly and $5\ 547$ incorrectly predicted trials ($5\ 054$ and $3\ 568$ for correct and incorrect predictions in the top-five split), thereby closely mirroring the percentages from the whole data set.

Accounting for Model Self-Predictability

As mentioned above, correlations between neighbouring word embeddings might explain positive pre-onset encoding results, since the word embedding of the word in question x_0 can be used to predict the preceding word x_{-1}, and can therefore, be used in order to predict brain activity prior to word onset. We, therefore, aimed to determine whether this effect can be corrected for quantitatively, such that we could test for pre-activation while correcting for self-predictability. For this purpose, word embeddings were residualised. For each word in the text, we regressed-out the previous embedding from the the embedding at time point zero. Hence, given the sentence "You know my methods.", "You" is regressed out of "know", "know" is regressed out of "my", and "my" out of "methods". Subsequently, we re-ran the self-predictability to check that the correlations between neighbouring word-embeddings were indeed removed. Note that self-predictability is not the same as auto-correlation—in fact, it is the time-resolved auto-correlation under arbitrary linear transformations at each timepoint, making it more suitable as a control for an encoding analysis that is based on learning such arbitrary linear mappings. For the same reason, the analysis can be seen as a generalised version of the control analysis based on performed by ***Goldstein et al.*** (***2022b***), based on directly projecting out neighbouring word embeddings.

Acoustic Analysis

In order to ascertain to what extent correlations in the stimulus material might still be able to explain ostensible hallmarks of prediction, even in the absence of pre-activations, we analysed the predictability of the acoustics of each word. We obtained an acoustic embedding of a word by computing the average 8-Mel spectrogram and envelop of each word. Like for the self-predictability analysis, we constructed acoustic epochs by computing the onset and offset times of each of our 157 time points within the $4s$ intervall, determined which word was presented at that time point and filled that data point with the acoustic representation of that word. This resulted in a y-matrix of $85\ 719 \times 9$ for the few-subject dataset and $7\ 745 \times 9$ for the multi-subject dataset. We then tested the predictability of this representation of our stimulus material using the same encoding model as used for the brain encoding (see Figure 4A).

Crucially, however, the encoding model relied on residualised word embeddings in order to predict acoustic representations. Hence, any positive pre-onset encoding could no longer be due to model self-predictability, but would be an effect of correlations present in the acoustics of neighbouring words. The acoustic encoding analysis, therefore, probes the predictability of the stimulus itself, while mapping between different representational formats—from residualised word embeddings to acoustics—which is arguably more similar to the neural encoding analysis. This approach allows us to determine whether residual brain encoding could result from temporal correlations in the neural data which are not accounted for when controlling for a model's self-predictability.

Code availability

The code used for modelling analyses and plotting is available at https://github.com/InesSchoenmann/Lingpred

References

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

Armeni K, Güçlü U, van Gerven M, Schoffelen JM. A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. Scientific Data. 2022; 9(1):278.

Azizpour S, Westner BU, Szewczyk J, Güçlü U, Geerligs L. Signatures of prediction during natural listening in MEG data? arXiv preprint arXiv:241219622. 2024; .

Boston MF, Hale JT, Vasishth S, Kliegl R. Parallel processing and sentence comprehension difficulty. Language and Cognitive Processes. 2011; 26(3):301–349.

Brodbeck C, Bhattasali S, Heredia AAC, Resnik P, Simon JZ, Lau E. Parallel processing in speech perception with local and global representations of linguistic context. ELife. 2022; 11:e72056.

Caucheteux C, Gramfort A, King JR. Evidence of a predictive coding hierarchy in the human brain listening to speech. Nature Human Behaviour. 2023; p. 1–12.

Caucheteux C, King JR. Language processing in brains and deep neural networks: computational convergence and its limits. BioRxiv. 2020; p. 2020–07.

Caucheteux C, King JR. Brains and algorithms partially converge in natural language processing. Communications biology. 2022; 5(1):134.

Frank SL, Otten LJ, Galli G, Vigliocco G. The ERP response to the amount of information conveyed by words in sentences. Brain and language. 2015; 140:1–11.

Goldstein A, Ham E, Nastase SA, Zada Z, Grinstein-Dabus A, Aubrey B, Schain M, Gazula H, Feder A, Doyle W, et al. Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain. bioRxiv. 2022; p. 2022–07.

Goldstein A, Zada Z, Buchnik E, Schain M, Price A, Aubrey B, Nastase SA, Feder A, Emanuel D, Cohen A, et al. Shared computational principles for language processing in humans and deep language models. Nature Neuroscience. 2022; 25(3):369–380.

Gwilliams L, Flick G, Marantz A, Pylkkanen L, Poeppel D, King JR. MEG-MASC: a high-quality magneto-encephalography dataset for evaluating natural speech processing. arXiv preprint arXiv:220811488. 2022; .

Heilbron M, Armeni K, Schoffelen JM, Hagoort P, De Lange FP. A hierarchy of linguistic predictions during natural language comprehension. Proceedings of the National Academy of Sciences. 2022; 119(32):e2201968119.

Heilbron M, van Haren J, Hagoort P, de Lange FP. Lexical processing strongly affects reading times but not skipping during natural reading. Open Mind. 2023; 7:757–783.

Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing; 2017, to appear.

Huettig F. Four central questions about prediction in language processing. Brain research. 2015; 1626:118–135.

Huettig F, Mani N. Is prediction necessary to understand language? Probably not. Language, Cognition and Neuroscience. 2016; 31(1):19–31.

Jain S, Huth A. Incorporating context into language encoding models for fMRI. Advances in neural information processing systems. 2018; 31.

Kuperberg GR, Jaeger TF. What do we mean by prediction in language comprehension? Language, cognition and neuroscience. 2016; 31(1):32–59.

Kutas M, Hillyard SA. Event-related brain potentials to semantically inappropriate and surprisingly large words. Biological psychology. 1980; 11(2):99–116.

Kutas M, Hillyard SA. Reading senseless sentences: Brain potentials reflect semantic incongruity. Science. 1980; 207(4427):203–205.

Kutas M, Hillyard SA. Brain potentials during reading reflect word expectancy and semantic association. Nature. 1984; 307(5947):161–163.

Linzen T, Baroni M. Syntactic structure from deep learning. Annual Review of Linguistics. 2021; 7(1):195–212.

Manning CD. Human language understanding & reasoning. Daedalus. 2022; 151(2):127–138.

Manning CD, Clark K, Hewitt J, Khandelwal U, Levy O. Emergent linguistic structure in artificial neural networks trained by self-supervision. Proceedings of the National Academy of Sciences. 2020; 117(48):30046–30054.

Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, Gao J. Large language models: A survey. arXiv preprint arXiv:240206196. 2024; .

Nieuwland MS. Do 'early'brain responses reveal word form prediction during language comprehension? A critical review. Neuroscience & Biobehavioral Reviews. 2019; 96:367–400.

Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*; 2014. p. 1532–1543.

Pickering MJ, Gambi C. Predicting while comprehending language: A theory and review. Psychological bulletin. 2018; 144(10):1002.

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI blog. 2019; 1(8):9.

Schrimpf M, Blank I, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, Tenenbaum J, Fedorenko E. Artificial neural networks accurately predict language processing in the brain. BioRxiv. 2020; p. 2020–06.

Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, Tenenbaum JB, Fedorenko E. The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences. 2021; 118(45):e2105646118.

Shain C, Meister C, Pimentel T, Cotterell R, Levy R. Large-scale evidence for logarithmic effects of word predictability on reading time. Proceedings of the National Academy of Sciences. 2024; 121(10):e2307876121.

Smith NJ, Levy R. The effect of word predictability on reading time is logarithmic. Cognition. 2013; 128(3):302–319.

Szewczyk JM, Federmeier KD. Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. Journal of memory and language. 2022; 123:104311.

Tuckute G, Kanwisher N, Fedorenko E. Language in brains, minds, and machines. Annual Review of Neuroscience. 2024; 47.

Wang L, Kuperberg G, Jensen O. Specific lexico-semantic predictions are associated with unique spatial and temporal patterns of neural activity. elife. 2018; 7:e39061.

Willems RM, Frank SL, Nijhof AD, Hagoort P, Van den Bosch A. Prediction during natural language comprehension. Cerebral Cortex. 2016; 26(6):2506–2516.

Supplementary Figures

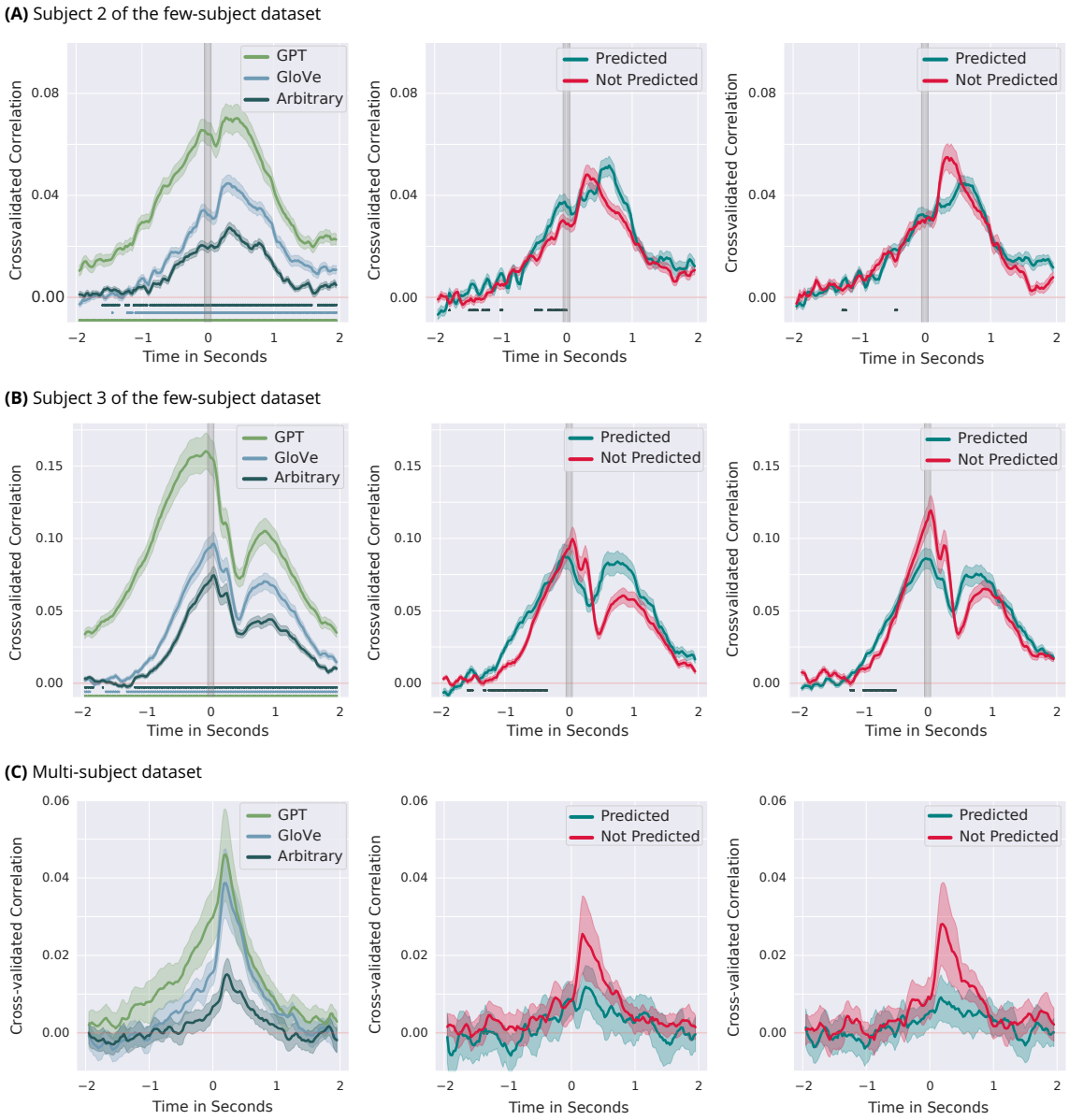

Figure S1. Hallmarks of prediction in the remaining two subjects of the few-subject dataset and the Multi-subject dataset. First panel from the left shows the overall encoding performance of GPT-2, GloVe and arbitrary vectors. Lines show clusters of time points for which encoding performance is significantly different from zero ($p < 0.05$ under the permutation distribution). Middle panel shows the sensitivity to the predictability of the word for GPT-2-XL's top-1 prediction and the right panel shows the the sensitivity to GPT-2-XL's top-5 prediction. Lines show clusters of time points for which encoding performance is significantly larger for the predictable as opposed to unpredictable words prior to word onset ($p < 0.05$ under the permutation distribution). Shaded areas show 95-% confidence intervals computed over sources and cross-validation splits in the single-subject analyses and over subjects in the multi-subject analysis. For the multi-subject dataset we find no evidence for the second hallmark of prediction, i.e. no sensitivity to the predictability of a word for pre-onset encoding performance.

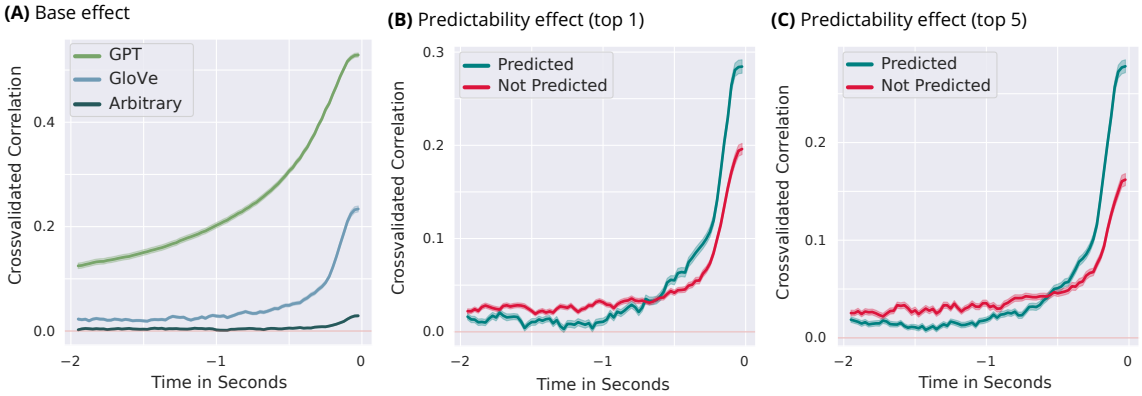

Figure S2. Self-Predictability in the Multi-subject dataset. First panel from the left shows self-predictability of GPT-2, GloVe and arbitrary models. Middle panel shows the sensitivity of self-predictability of GloVe to the predictability of the word as defined by GPT-2-XL's top-1 prediction and the right panel shows the the sensitivity as defined by GPT-2-XL's top-5 prediction. Shaded areas show 95-% confidence intervals computed over model dimensions. Self-predictability results are identical to those observed in the few-subject dataset. We find both ostensible hallmarks of prediction in the stimulus material, namely the word embeddings of the material.

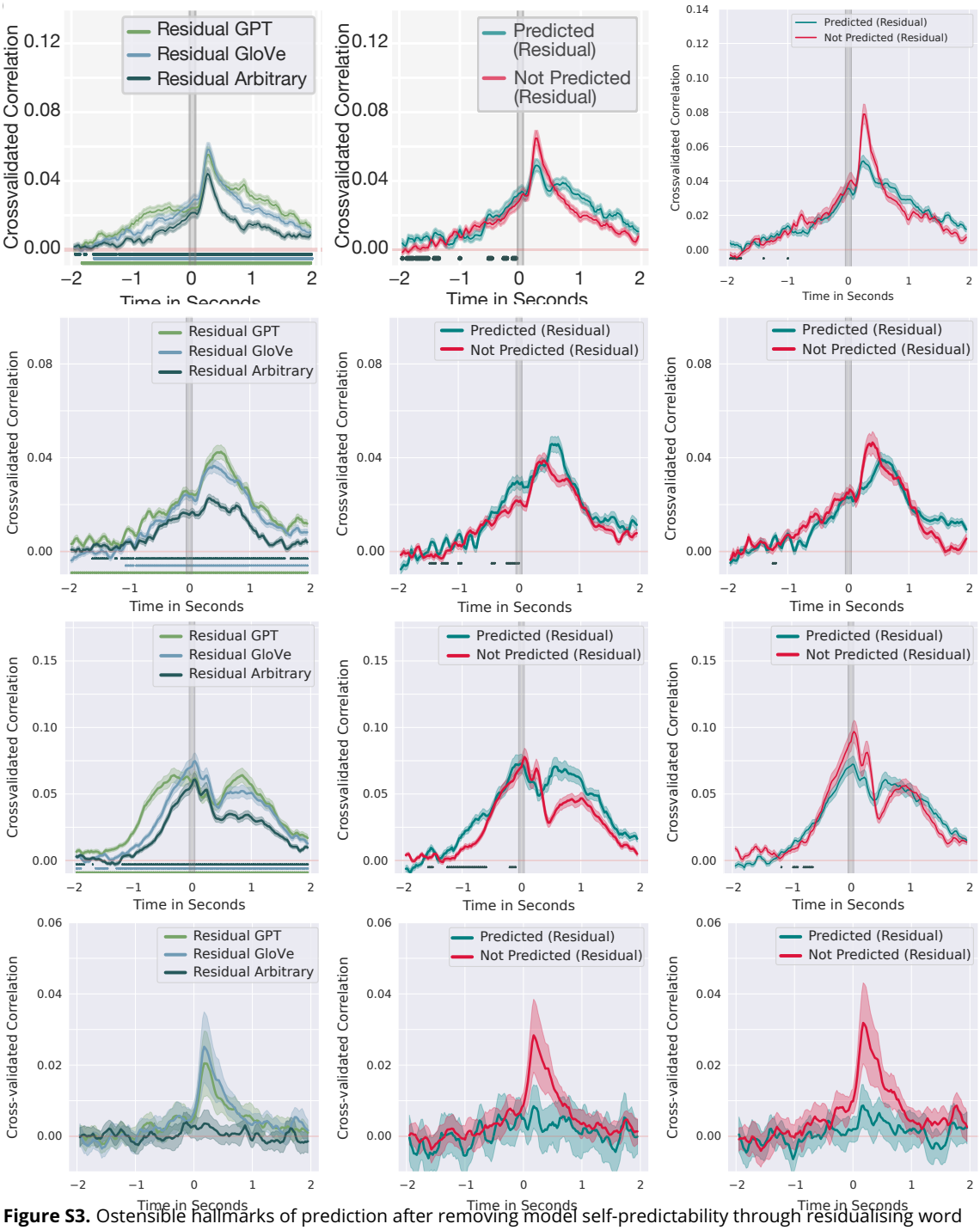

Figure S3. Ostensible hallmarks of prediction after removing model self-predictability through residualising word embeddings. **Left panels:** Overall encoding performance of residualised GPT-2, GloVe and arbitrary vectors. Lines show clusters of significant time points ($p < 0.05$). **Middle panels:** Sensitivity to GPT-2-XL's top-1 prediction. **Right panels:** Sensitivity to GPT-2-XL's top-5 prediction. Lines show clusters with significantly larger encoding for predictable vs. unpredictable words prior to onset ($p < 0.05$). Shaded areas: 95% confidence intervals. For the multi-subject dataset, no evidence of the second hallmark of prediction is found.

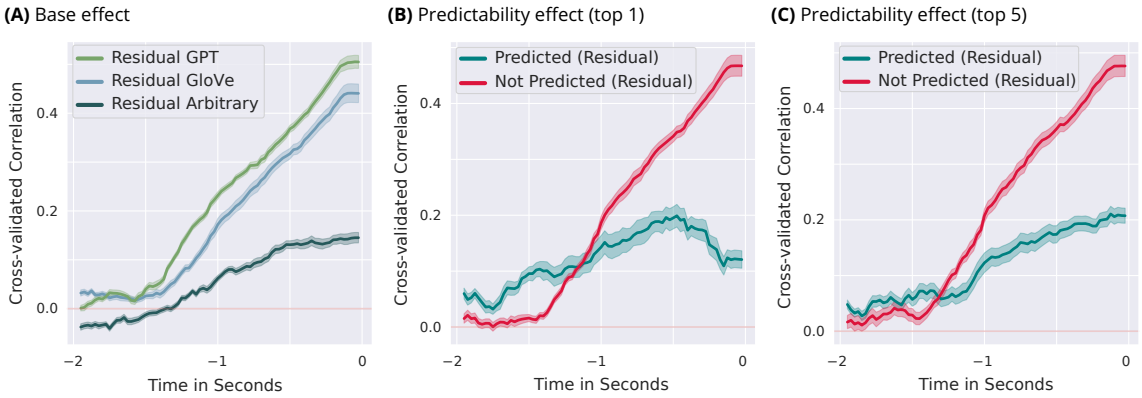

Figure S4. Predicting Acoustics prior to word onset from residual word embeddings in the multi-subject dataset. First panel from the left shows the overall encoding performance of residualised GPT-2, GloVe and arbitrary vectors. Results mirror brain encoding results both in encoding time courses and differences between models (see Figure S3 C). Middle panel shows the sensitivity to the predictability of the word for GPT-2-XL's top-1 prediction and the right panel shows the the sensitivity to GPT-2-XL's top-5 prediction when using residualised GloVe embeddings. Shaded areas show 95-% confidence intervals computed over the 9 dimensions (8 mels + envelope). Results mirror qualitative differences observed in the brain encoding study.

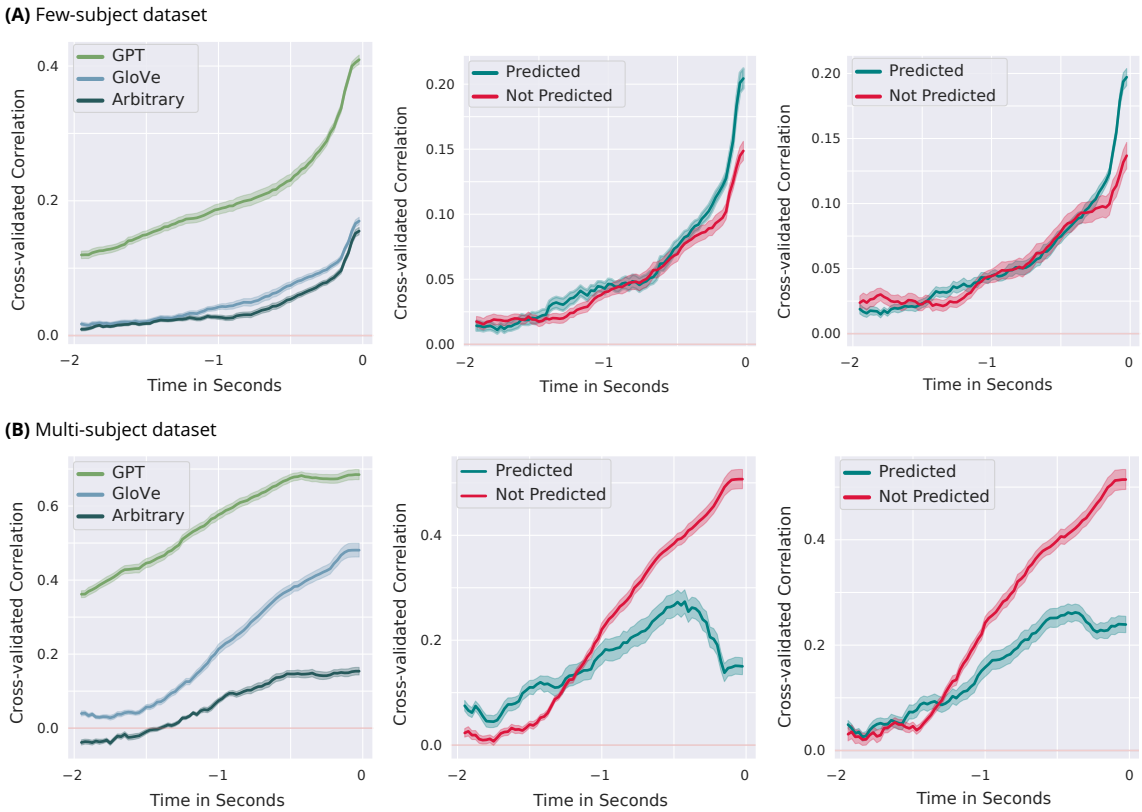

Figure S5. Predicting Acoustics prior to word onset from original embedding vectors for both datasets. First panel from the left shows the overall encoding performance of residualised GPT-2, GloVe and arbitrary vectors. Results closely mirror brain encoding results both in encoding time course and differences between models (see Figure 1B and C and S1). Middle panel shows the sensitivity to the predictability of the word for GPT-2-XL's top-1 prediction and the right panel shows the the sensitivity to GPT-2-XL's top-5 prediction when using original GloVe embeddings. Shaded areas show 95-% confidence intervals computed over the 9 dimensions (8 mels + envelope). Results mirror qualitative differences observed in the brain encoding studies.