Large Language Models Reveal the Neural Tracking of Linguistic Context in Attended and Unattended Multi-Talker Speech

Corentin Puffay[1,2*], **Gavin Mischler**[3], **Vishal Choudhari**[3], **Jonas Vanthornhout**[1], **Stephan Bickel**[4,5], **Ashesh D. Mehta**[4,5], **Catherine Schevon**[6], **Guy M. McKhann**[7], **Hugo Van hamme**[2], **Tom Francart**[1], **Nima Mesgarani**[3*]

***For correspondence:**
corentin.puffay@kuleuven.be (CP);
nima@ee.columbia.edu (NM)

[1]KU Leuven, Dept. Neurosciences, ExpORL, Leuven, Belgium; [2]KU Leuven, Dept. of Electrical engineering (ESAT), PSI, Leuven, Belgium; [3]Columbia University, Dept. of Electrical Engineering, New York, NY, USA; [4]Hofstra Northwell School of Medicine, Uniondale, NY, USA; [5]The Feinstein Institutes for Medical Research, Manhasset, NY, USA; [6]Columbia University, Dept. of Neurology, New York, NY, USA; [7]Columbia University, Dept. of Neurological Surgery, Vagelos College of Physicians and Surgeons , New York, NY, USA

Abstract Large language models (LLMs) have shown strong alignment with contextualized semantic encoding in the brain, offering a new avenue for studying language processing. However, natural speech often occurs in multi-talker environments, which remain underexplored. We investigated how auditory attention modulates context tracking using electrocorticography (ECoG) and stereoelectroencephalography (sEEG) from three epilepsy patients in a two-conversation "cocktail party" paradigm. LLM embeddings were used to predict brain responses to attended and unattended speech streams. Results show that LLMs reliably predict brain activity for the attended stream, and notably, contextual features from the unattended stream also contribute to prediction. Importantly, the brain appears to track shorter-range context in the unattended stream compared to the attended one. These findings reveal that the brain processes multiple speech streams in parallel, modulated by attention, and that LLMs can be valuable tools for probing neural language encoding in complex auditory settings.

Introduction

The emergence of Large Language Models (LLMs) and their high performance on NLP tasks has garnered attention from neuroscience research. In particular, similarities in how LLMs and the human brain process language have been pointed out (***Goldstein et al., 2022***; ***Schrimpf et al., 2021a***; ***Mischler et al., 2024***; ***Caucheteux et al., 2023***), and used to model language processing in the brain. Many studies (***Toneva and Wehbe, 2019***; ***Abnar et al., 2019***; ***Schrimpf et al., 2021b***; ***Hosseini et al., 2024***; ***Anderson et al., 2021***; ***Caucheteux et al., 2021***; ***Caucheteux and King, 2022***; ***Sun et al., 2021***) train linear regression models on word representations extracted from LLMs to predict corresponding neural responses. The correlation between predicted and observed neural responses serves as a metric of how well the model's representations align with brain activity, which can further be

used to investigate the neural basis of language processing.
However, the observed linear relationship between LLM-derived word representations and brain responses can arise from multiple contributing factors, whose influence remains to be fully disentangled. To better characterize the language features encoded by these models, some researchers analyzed activations elicited by naturalistic stimuli while applying controlled manipulations to the corresponding transcripts (e.g., holding syntax invariant while varying semantic content). This enables factorization of word representations into lexical, syntactic, and semantic components (***Caucheteux et al., 2021***). Additionally, word representations generated by LLMs carry both contextual information (related to previous words in the transcript) and information inherent to the word of interest (e.g., name entity, word frequency). To separate contextual from non-contextual lexical information, some studies (***Goldstein et al., 2022***) have used models like GloVe (***Pennington et al., 2014***), which generate word representations solely based on word occurrence statistics without positional or contextual information.

Studies modeling language processing in the human brain involving LLMs were mostly performed in a controlled setup, where participants listened to a single speaker. However, real-life scenarios typically involve multiple talkers, making the speech processing more challenging and altering neural responses through the effects of attention (***Mesgarani and Chang, 2012***; ***O'Sullivan et al., 2019***). Prior studies (***Broderick et al., 2018***; ***Anderson et al., 2024***; ***Brodbeck et al., 2018***) have suggested that the high-level linguistic encoding is likely impaired in challenging listening conditions. This makes such environments an interesting experimental design for studying the encoding of contextual language representations. In this multi-talker paradigm, the listener has to selectively attend to one talker, providing a setup where the encoding of context might be altered by attention. ***Anderson et al.*** (***2024***) investigated the ability of the automatic speech recognition (ASR) model Whisper (***Radford et al., 2022***) to predict EEG responses to continuous speech in a two-speaker scenario. They found that late Whisper's encoder layers (encoding linguistic information) were better predictors than early layers (encoding acoustics) for the attended speaker, but not for the unattended one, suggesting limited or no linguistic processing of the unattended speech. Unfortunately, EEG's signal-to-noise ratio is notoriously low, which raises the possibility that the apparent absence of linguistic processing of unattended speech may stem from limitations in the recording method itself.

Most of the findings with EEG recordings suggest that attention is a necessary condition for contextual linguistic encoding (***Anderson et al., 2024***; ***Broderick et al., 2018***; ***Brodbeck et al., 2018***). However, the EEG prediction scores obtained with linguistic features are typically very low and minimal compared to acoustic features (***Gillis et al., 2021***; ***Puffay et al., 2023***; ***Brodbeck et al., 2018***), possibly due to a low signal-to-noise ratio (SNR) that hinders specific steps of linguistic processing from being captured. While the EEG signal contains responses reflecting various speech and language processing steps, electrocorticography (ECoG) and stereoelectroencephalography (sEEG) enable targeting specific areas, which allows the study of context encoding in a multi-talker scenario with a higher SNR.

In this study, we investigate how attention modulates context tracking. To do so, we use LLM embeddings to predict brain responses measured with ECoG and sEEG in a two-competing-talker attention paradigm. We first compare the prediction scores obtained across the LLM layers for the attended and unattended conversation transcripts. We demonstrate that contextual language features generated by our LLM from both the attended and unattended streams can predict brain responses, with significantly greater predictive power for the attended stream. Secondly, we compare the brain prediction scores obtained when the correct context is provided to our LLM versus an unrelated context, demonstrating the importance of context in improving prediction scores in deeper layers. Finally, by modulating the context length fed to the LLM, we show that the brain

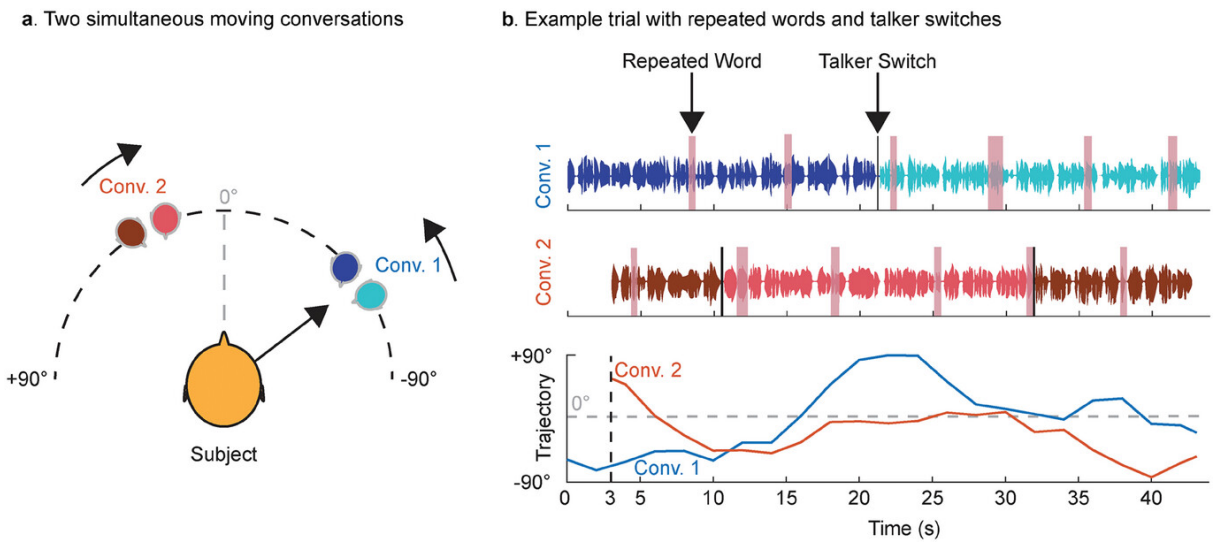


Figure 1. Auditory attention experiment design from ***Choudhari et al.*** (***2024***). a) Every trial consisted of two concurrent conversations moving independently in the front hemifield of the subject. Each conversation had two distinct talkers taking turns. b) Talkers in each conversation repeated a word at random intervals (1-back detection task), as highlighted in pink. The cued (to-be-attended) conversation had a talker switch at 50% trial time mark, whereas the uncued (to-be-unattended) conversation had two talker switches at 25% and 75% trial time marks.

tracks a shorter-range context from the unattended stream than that of the attended stream.

Results

To explore how linguistic context is represented in the human brain during naturalistic listening and selective attention, we analyzed intracranial recordings from three epilepsy patients as they performed a dual-speaker auditory attention task (Figure 1). Participants listened to two overlapping conversations and were instructed to attend to one, while their neural responses were recorded via sEEG and ECoG electrodes implanted over the left hemisphere of their brain. High-gamma (70 - 150 Hz) neural activity was extracted and used as a proxy for local neural engagement (***Edwards et al., 2009***).

To model the contextual encoding across attentional states, we used the Mistral-7B LLM to generate word representations from attended and unattended speech transcripts (Figure 2). Using ridge regression, we quantified how well these embeddings predicted neural responses across different model layers and temporal lags, as well as across varied contextual input given to the LLM. This framework enabled us to examine how attention, contextual input, and context length shape the brain's alignment with LLM-derived representations.

In what follows, we present three key findings: (1) attention enhances brain prediction scores from LLM embeddings, (2) this effect is primarily driven by contextual information, and (3) the encoded context length is shorter for the unattended than the attended speech. Together, these results highlight attentional modulation of linguistic context processing in the brain.

Result 1: Language encoding, measured with LLM brain prediction scores is impacted by attention

In Figure 3, we depict the LLM word representation prediction scores for every layer for both the attended and unattended speech transcripts. The attended and unattended prediction scores for subjects 1, 2, and 3 are depicted in Figures 3a, 3c, and Figure 3e, respectively. We observe a significant difference in the averaged brain score magnitudes across layers between the attended and

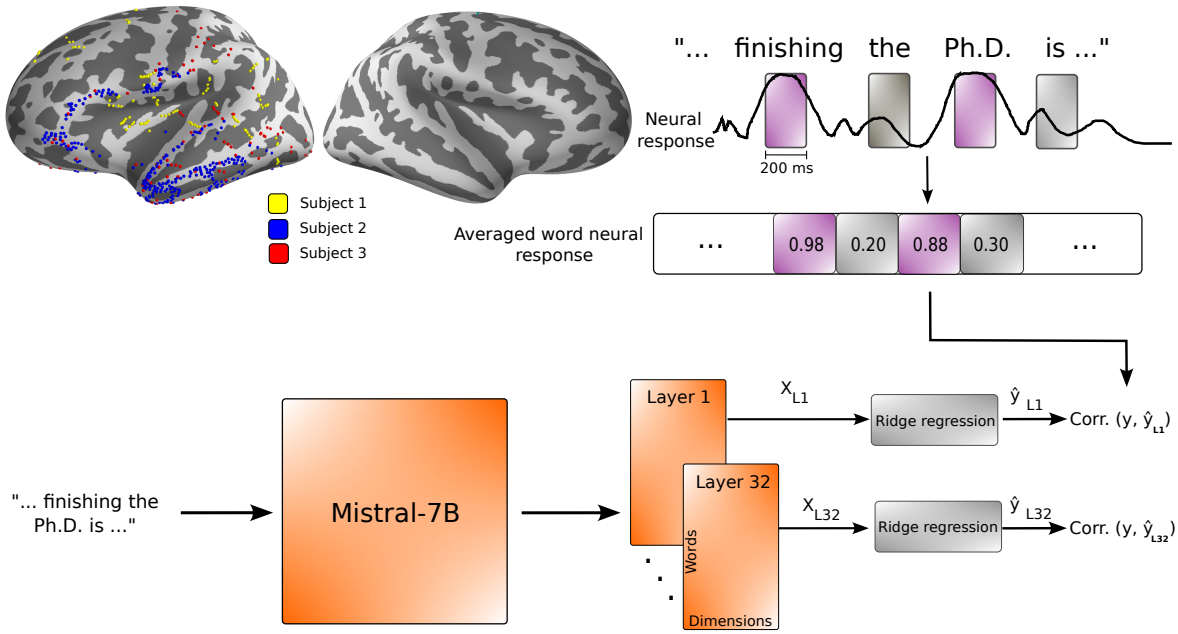


Figure 2. LLM embeddings to predict brain responses. (Inspired from *Mischler et al.* (*2024*)). We averaged the neural responses over 200-ms windows centered at seven lags, from 0 to 500 ms after word onsets. The word representations are extracted using different layers of Mistral-7B to predict the averaged word response further. The correlation between the predicted and actual responses is calculated to compute the brain score.

unattended conditions (two-sided Wilcoxon signed-rank test, $W = 1626{-}1996, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$, for subjects 1, 2 and 3, respectively), suggesting differences in the language information encoded across attention states. The median difference was 0.023, 0.047, and 0.010 for subjects 1, 2 and 3, respectively, consistent with a higher prediction score for the attended condition.

Specifically, in Mistral-7B, the output of layer 0 corresponds to token and positional embeddings only, without any contextual integration. Thus, improvements in brain reconstruction scores observed in subsequent layers can be attributed to the accumulation of contextual information. We observe that the attended condition shows a difference in prediction score, e.g., between layer 0 and layer 6 (two-sided Wilcoxon signed-rank test, $W = 1385, 4253, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$) for subjects 1 and 2. For all subjects, a prediction score difference was observed between layer 0 and 7 (two-sided Wilcoxon signed-rank test, $W = 1380 - 4123, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$) and between layer 0 and 8 (two-sided Wilcoxon signed-rank test, $W = 1336 - 3246, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$). For layer 6, subjects 1 and 2 showed median differences of 0.023 and 0.028, respectively. Additionally, the median difference subject ranges were 0.015-0.026 and 0.012-0.044 for layers 7 and 8, respectively. These values are consistent with a higher prediction score for these layers compared to layer 0. Results were similar for neighboring layers. This indicates that the contextual information contained in word representations from layers 6, 7, and 8 for subjects 1, 2, and 3, respectively, enables a better prediction of the average brain response than layer 0.

In comparison, we also observed a difference in prediction score for the unattended transcript in subject 1 between layer 0 and 6 (two-sided Wilcoxon signed-rank test, $W = 2917, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$). In addition, the median difference was 0.015, consistent with a higher prediction score for layer 6 compared to layer 0, and results were similar for neighboring layers. On the other hand, this effect was not found for subjects 2 and 3. While a difference was observed for the unattended case in subject 2 from layer 0 to 6 (two-sided Wilcoxon signed-rank test, $W = 21951, p = 0.24$), the median difference was -0.01, consistent with a lower prediction score for layer 6 compared to layer 0. Neighboring layers did not show any difference in prediction scores with layer 0. Subject 3 did not show any significant prediction score between layer 0 and layers 6, 7, 8 (two-sided Wilcoxon signed-rank test, W=2606-3176, p=0.04-0.65, not significant after Bonferroni correction, $\alpha = 0.05/6$). This finding suggests that there might be an unattended context encoding for subject 1.

Figures 3b,3d, and 3f depict the attentional modulation of electrodes (i.e., the brain prediction score of the attended condition minus that of the unattended condition).

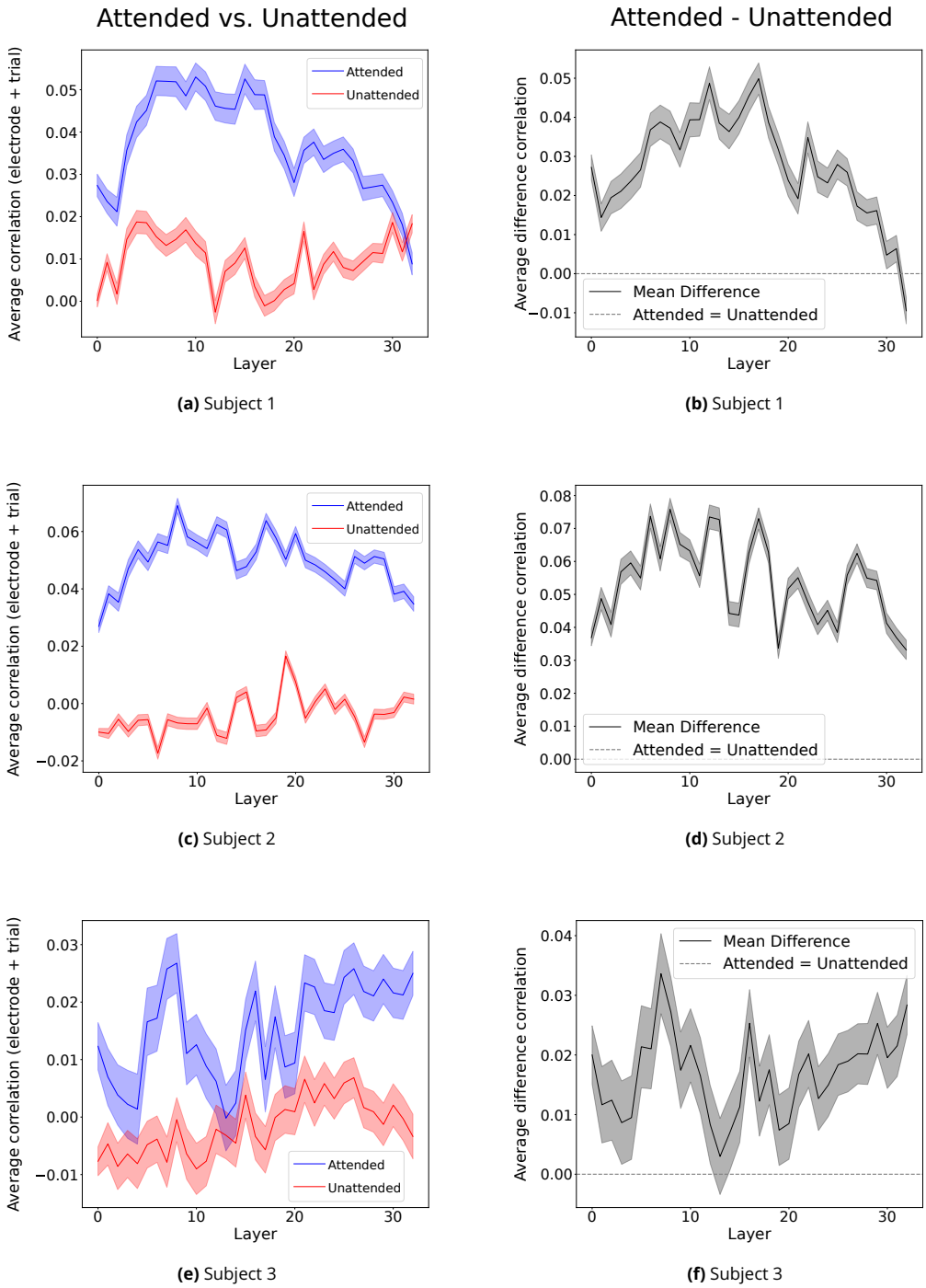


Figure 3. Brain prediction scores of LLM word representations across layers with attention modulation. In (a), (c), and (e) (for subjects 1, 2, and 3, respectively), the two curves depict the average brain score across trials and electrodes for each layer, with SEM across trial-averaged electrodes. Blue: attended speech transcript; red: unattended. (b), (d), and (f) show the attended-unattended score differences.

We did not observe any localized effect of attention when looking at the prediction scores across brain locations. To support this statement, we provide brain mappings of correlations across subjects when looking at the benefit of attention, i.e., the difference in prediction scores between the attended and unattended conditions at layer 10 (see Supplementary Figure 1).

In the Supplementary Material, we run three additional analyses: (1) to evaluate the impact of repeated words on the prediction score (see Text S1, and Supplementary Figure 2), (2) to evaluate any speaker-specific effects that our model could learn and that could impair our conclusions (see Text S2, and Supplementary Figure 3), and (3) to evaluate the impact of attention drops measured behaviorally (see Text S3). None of these potential confounds changed the results of our analyses.

Result 2: Improvement of LLM prediction scores across layers reflects contextual encoding

To verify that the prediction score' increase across Mistral's layers is due to contextual information, we investigate how modifying the context input provided to Mistral influences its ability to predict brain responses. Thus, we repeat the analysis of the previous subsection. However, this time, we either give Mistral the original trial context before the current token (i.e., the one presented to the participants), as done in the previous subsection, or a completely unrelated context extracted from a different story but of the same length, so that it would not be helpful for the LLM to predict the next word (see Context modulation Section for details). We depict the resulting prediction scores for the original and unrelated context conditions in both the attended and unattended conditions in Figures 4a, 4c, and 4e for subjects 1, 2, and 3, respectively.

In the attended case, we observed a difference in the averaged prediction scores across layers between the related and the unrelated conditions for all subjects (two-sided Wilcoxon signed-rank test: $W = (562 - 1999, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$, for subject 1, 2 and 3, respectively). The difference of medians between the related and the unrelated conditions were 0.03, 0.02, and 0.005, consistent with an improvement of the prediction score of the related compared to unrelated conditions. This finding suggests an encoding of the related context in the brain response with attention.

For the unattended condition, conclusions were more ambivalent. Subject 1 showed a significant difference in prediction score (two-sided Wilcoxon signed-rank test: $W = 2021$, $p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$) associated with a positive difference between medians: 0.01. This result is consistent with an improvement of the related condition compared to the unrelated condition. We also observed a significant difference in prediction score for subject 2 (two-sided Wilcoxon signed-rank test: $W = 7750$, $p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$), along with a negative median difference (-0.01), consistent with an improvement of the unrelated condition over the related condition. We did not observe any significant difference between the related and unrelated context conditions for subject 3 (two-sided Wilcoxon signed-rank test, $W = 2463, p = 0.0015$, not significant after Bonferroni correction, $\alpha = 0.05/6$).

In the unattended condition, the results were inconclusive, showing no clear benefit of providing related context to the LLM for generating word representations that improve brain response prediction. One possible explanation is that using too much context may introduce information that is not actually encoded in the brain when the conversation is unattended. This excessive contextual information could confuse the linear encoder, ultimately lowering prediction scores. To investigate this hypothesis, we systematically vary the context length in the next section.

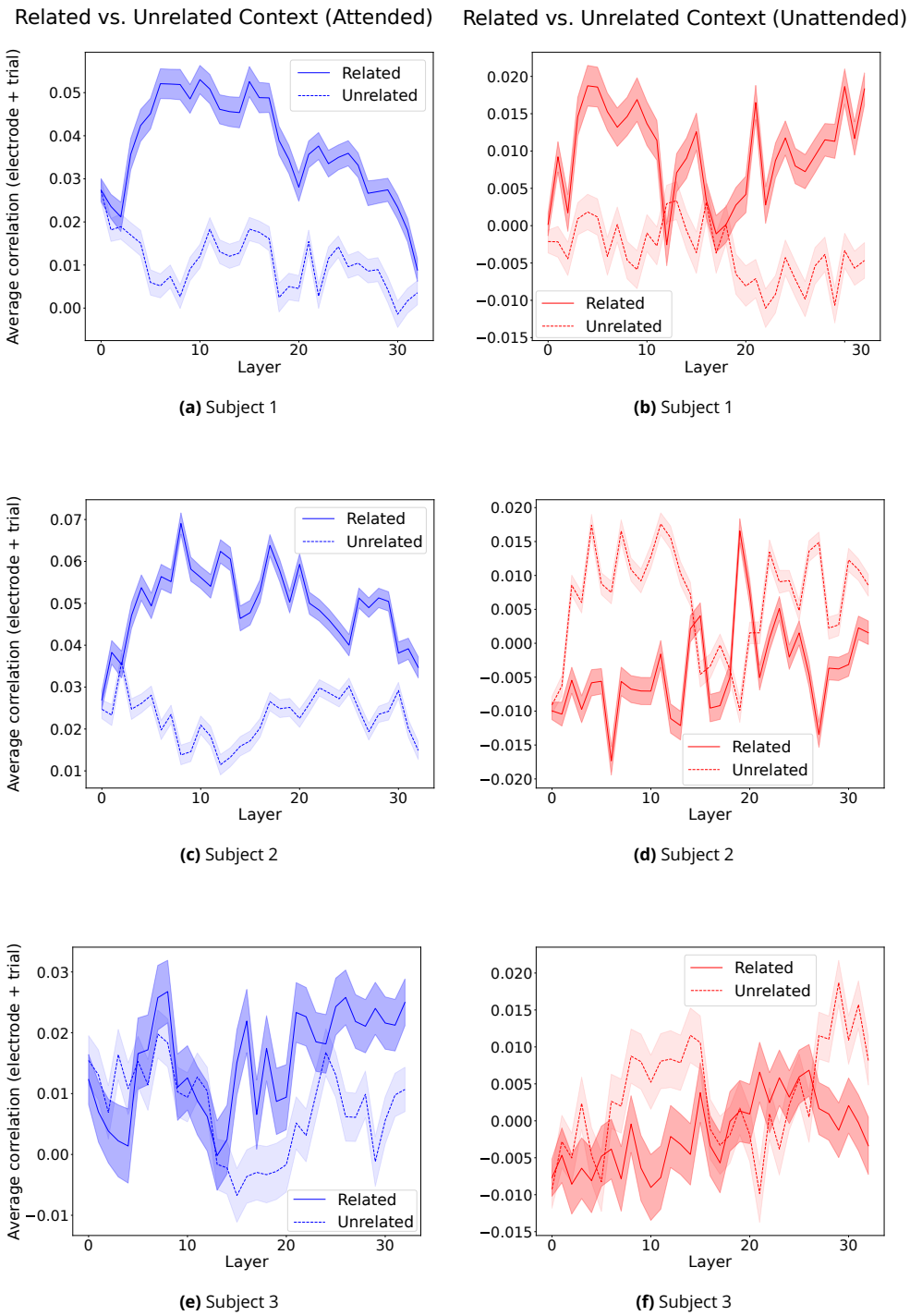


Figure 4. Related vs. unrelated context prediction scores across Mistral layers for both attended and unattended conditions. (a), (c) and (e) depict (for subjects 1, 2, and 3, respectively) the related context (blue curve) and unrelated context (light blue curve) trial-electrode averaged prediction scores across Mistral layers when using the attended word representations. (b), (d) and (f) depict (for subjects 1, 2, and 3, respectively) the related context (blue curve) and unrelated context (light blue curve) trial-electrode averaged prediction scores across Mistral layers when using the unattended word representations.

Result 3: Contextual encoding persists without attention but at shorter context lengths

Here, we characterize the longest context length to which the model benefits in both the attended and unattended conditions. From now on, we depict our results for layer 10, as it is part of the plateau reached for the attended curve in Figure 3a and 3c. Results of neighboring layers from this plateau are reported in Supplementary Figure 5, as their analyses led to the same conclusions. We thus re-ran the analysis with layer 10 for different context lengths (in tokens): 1, 5, 10, 15, 50, and the full context. In Figure 5, we depict for subjects 1, 2, and 3 how the average brain score behaves as a function of context length. To ensure that the observed correlations are related to language processing, rather than being inherent to Mistral-7B's architecture alone, we also fit an encoding model with word representations generated from Mistral-7B initialized with random weights.

We observe that the average prediction score increases steadily with the context length in the attended condition,i.e., a significant difference between 15 tokens and the full context condition (two-sided Wilcoxon signed rank test, $W = 4166, W = 14259, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$, for subject 1 and 2, respectively) with positive median differences: 0.01 and 0.01, for subjects 1 and 2, respectively. While a statistical difference was observed between 5 and 10 tokens in subject 3 (two-sided Wilcoxon signed rank test, $W = 1352, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$), it was a decline as shown by the negative median difference (-0.02) between the two conditions.

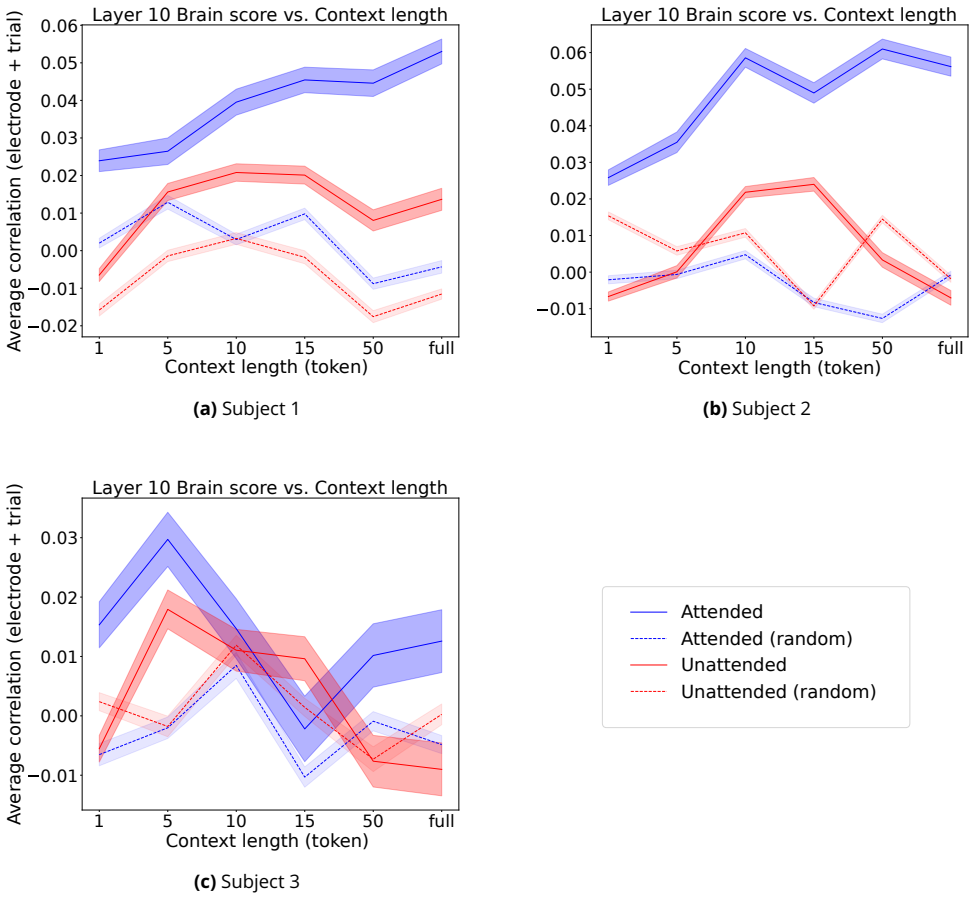
In the unattended condition, the prediction score increase stops at 10 tokens, demonstrated by no significant difference between 10 to 15 tokens conditions (two-sided Wilcoxon signed rank test, $W = 6186, W = 20889, p = 0.23, p = 0.06$ for subjects 1 and 2, respectively). Then it decreased significantly from 15 to 50 tokens (two-sided Wilcoxon signed rank test, $W = 2936, 7593, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$, median differences: 0.01 and 0.03 for subjects 1 and 2, respectively). For subject 3 while a significant difference was found between 5 and 10 tokens (two-sided Wilcoxon signed rank test, $W = 2067, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$), the median difference was negative (-0.005), consistent with a decrease in prediction score between the two conditions. This finding suggests for subjects 1 and 2 that, while some context information about the unattended conversation is encoded in the brain response, its length is shorter than that of the attended conversation.

Across the three subjects, we tested the prediction score's difference between the trained LLM and the untrained LLM condition using two-sided Wilcoxon signed-rank tests across electrodes, averaging across trials at a given context length. This analysis examines whether the encoded speech carries meaningful information beyond what is captured by an untrained model.

In Subject 1, significant differences were observed for almost all context lengths in both attended and unattended conditions (two-sided Wilcoxon signed rank test, attended: $W = 2689 - 227$, $p < 0.001$; unattended: $W = 4705 - 2242$, $p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$), except for the attended condition at context length 5 (two-sided Wilcoxon signed rank test, $W = 6005$, $p = 0.136$). All statistical differences observed were associated with a positive median difference between the trained and untrained conditions, consistent with a higher prediction score when the LLM was trained. In Subject 2, significant differences were observed for all context lengths and both conditions (two-sided Wilcoxon signed rank test, attended: $W = 5700 - 764$, $p < 0.001$; unattended: $W = 6937 - 15275$, $p < 0.001$). While the attended case, all median differences were positive, it was only the case for the 10 and 15 tokens conditions in the unattended case (0.01 and 0.03 for 10 and 15 tokens, respectively), consistent with a higher performance for the trained LLM compared to the untrained LLM. In Subject 3, significant differences were observed for short context lengths (1 and 5) and the full context (two-sided Wilcoxon signed rank test, attended: $W = 1366$, 1049, 2277, $p < 0.003$; unattended: $W = 1943$, 1228, 1829, $p < 0.0001$). On the other hand, intermediate and

long context lengths (10, 15, 50) were not significant in either attended or unattended conditions (two-sided Wilcoxon signed rank test, attended: $W = 2988, 3043, 3214$, $p = 0.33$, $p = 0.41$, $p = 0.73$; unattended: $W = 2759, 2822, 2749$, $p = 0.10$, $p = 0.15$, $p = 0.10$).

Overall, these results indicate that the trained LLM consistently outperforms the untrained LLM, supporting the conclusion that the brain encodes meaningful contextual information from attended and unattended speech embeddings.



(a) Subject 1

(b) Subject 2

(c) Subject 3

Figure 5. Brain prediction scores of Mistral layer 10 embeddings across context lengths. (a), (b), and (c) depict (for subjects 1, 2, and 3, respectively) the attended (blue) and unattended (red) average prediction scores across trials and electrodes using different context lengths. The blue and red dashed curves correspond to the brain prediction scores obtained with an untrained LLM. The corresponding shaded color also depicts the SEM across trial-averaged electrode prediction scores.

Discussion

In this study, we explored to what extent contextual LLM-based linguistic features can predict brain responses to speech in a multi-talker listening environment. Furthermore, we characterized the range of context encoded in the brain when a conversation is attended or unattended.

In Result 1, we showed that the LLM gradually learns contextual information when increasing the layer index, and we found that this improved the brain score until a plateau is reached. The performance then decreases, potentially because Mistral-7B is a generative LLM, which means it will output an answer to a text input. In late layers, the embeddings may thus incorporate more information related to the generated response and unrelated to the input prompt, which could explain the decrease in the brain score in later layers. We found this for the attended condition

but also, in the case of subject 1, for the unattended condition. Although prediction scores reach a lower-magnitude plateau, this suggests contextual tracking without attention. ***Anderson et al. (2024)*** performed a similar analysis combining features extracted from an ASR model (i.e., Whisper), the word surprisal from an LLM (i.e., GPT-2), and the speech envelope with its first derivative. They demonstrated that later Whisper layers were more accurate EEG predictors for the attended stream, providing an added value over the speech envelope and its first derivative. Moreover, they found a unique contribution of the word surprisal, and the speech envelope with its derivative to predict the response measured by certain electrodes. This suggests a non-contextual and contextual tracking of the attended stream. On the other hand, the unique contribution of the word surprisal and of Whisper deep layers disappeared for the unattended stream, suggesting the absence of contextual information encoded in the brain.
Additionally, an earlier EEG study using linear models demonstrated no lexical processing when a speaker was unattended (***Brodbeck et al., 2018***). We explain our different conclusions with the low SNR of EEG recordings. Multiple studies demonstrated the ability of LLM-generated features to be good predictors of the brain response in single-talker paradigms, notably with fMRI (***Caucheteux et al., 2023***; ***Antonello and Huth, 2024***) and ECoG (***Goldstein et al., 2022***; ***Mischler et al., 2024***). Our current finding expands the understanding of contextual encoding in a multi-talker scenario, by suggesting that some contextual linguistic information from the unattended speaker is being processed in the brain, and on a shorter time scale than for the attended speaker. This is the first time it is done with LLM-generated features, and it also questions earlier EEG findings, setting the limits of such recording modality.

In Result 2, we showed that the model relies on the correct contextual information to improve predictions of brain responses. To test this, we compared prediction scores across layers when Mistral-7B was fed the original trial context for each word versus an unrelated context taken from another story ("the black willow"). In the attended case, the unrelated context condition led to lower prediction scores across layers, underscoring the encoding of context information beyond the word's embedding of interest. In the unattended case, word representations may have included information from prior words that were not actually encoded in the brain. This hypothesis was further explored and supported in Result 3.

In Result 3, we explored multiple context lengths to characterize whether the attended and unattended context encoding were in the same length range. We found that for subjects 1 and 2, the unattended context length tracked reached a plateau earlier than that of the attended. Varying the context length provided to models has been done in multiple ECoG studies for single-speaker experiments with LLMs (***Mischler et al., 2024***; ***Goldstein et al., 2022***), also demonstrating a gradual increase of the prediction scores when increasing the context length. A recent EEG study (***Anderson et al., 2024***) investigated how ASR-extracted features relate to the brain response in a multi-talker environment. In line with previous EEG findings (***Brodbeck et al., 2018***), the authors did not find any linguistic contribution for the unattended stream, which contradicts our current findings for subjects 1 and 2. We explain this divergence with the low SNR of EEG, leading to minimal contributions of linguistic features to the EEG response, even in single-speaker paradigms compared to acoustics or lexical segmentation features (***Gillis et al., 2021***; ***Puffay et al., 2023***). For subject 3, we did not find the same context length tracking for both the attended and unattended conditions. Two potential explanations are: (1) the number of electrodes is minimal in the frontal lobe (see Supplementary Figure 4a), which contains areas encoding context (***Caucheteux et al., 2023***), and (2) the electrodes are less responsive, leading to poorer prediction scores than in subjects 1 and 2 (see Supplementary Figures 4c-e).

First and foremost, although ECoG and sEEG provide highly localized brain responses with a high signal-to-noise ratio (SNR), our study includes only three subjects, making it difficult to draw

general conclusions. Additionally, subject 3 did not exhibit the same pattern, potentially due to limited electrode coverage or reduced responsiveness. Nonetheless, the results reveal a compelling pattern regarding contextual tracking across attentional states that warrants further investigation. The limited electrode coverage also limited our ability to identify neuro-anatomical patterns related to the impact of attention on context. To have a homogeneous coverage across subjects, and if fast brain responses are not of interest, fMRI might offer an alternative to investigate neuro-anatomical insights using LLMs (e.g., ***Antonello et al., 2023***; ***Caucheteux et al., 2023***).
Besides discrepancies in electrode coverage between subjects, attention might not be a very localized phenomenon. We can think of attention as operating on different levels, such as acoustic attention, (sub-)lexical attention, and semantic attention. As a result, various brain areas involved in speech processing may be influenced by attention.
However, our text-based features lack acoustic details like tone, accent, and volume. While focusing solely on (sub-)lexical and linguistic information offers valuable insights into how these aspects are encoded, some brain regions might primarily respond to acoustic features of speech (***O'Sullivan et al., 2019***), limiting our ability to capture localized attention effects. A potential solution could be to combine this analysis with Automatic Speech Recognition (ASR) models like Whisper (***Radford et al., 2022***; ***Anderson et al., 2024***), allowing us to incorporate acoustic information while adjusting for context in the word representations.

Bearing that LLM embeddings are possibly restricted in terms of acoustic representation, it is not well-known which information is represented in LLM embeddings. For instance, we observed a difference between the first layer's brain score of the attended and unattended conversations. Although no studies have investigated the information contained in the Mistral-7B layers specifically, the first layer of transformer-based language models is known to produce non-contextualized word representations. Factorizing contextual from non-contextual prediction scores by subtracting the contribution of the first layer is thus a possible approach towards more interpretable LLM analyses. To control for non-contextual information, models such as GloVe can also be used to generate non-contextual word representations (***Goldstein et al., 2022***), and their prediction scores can be used as a baseline to quantify the non-contextual component. Additionally, to better understand what information is contained in each layer of an LLM, one could analyze activations elicited by naturalistic stimuli with controlled manipulations applied to their transcripts (e.g., invariant syntax but different semantic information), and factorize Mistral activations into different representation classes (e.g., lexical, syntactic, or contextual linguistics) (***Caucheteux et al., 2021***).

Lastly, the behavioral data analysis from ***Choudhari et al.*** (***2024***) indicated that the uncued conversation might have been attended occasionally (up to 8% of repeated word detection for subject 1). To alleviate the impact of these attention drops on our conclusions, we excluded the trials containing such events and showed that contextual tracking of the unattended stream remains (See Supplementary Figure 3). Our word detection task occurs every 7 s, thus not involving every word from the trials. Consequently, not every attention drop might be detected in our behavioral measurement. An objective measurement of auditory attention for every segment, such as alpha activity (***Deng et al., 2020***) or the performance of an auditory attention decoding model, could further reinforce our claim. On average across our three subjects, ***Choudhari et al.*** (***2024***) demonstrated a high auditory attention decoding accuracy (e.g., 82% using 2 s segments), suggesting minimal attention drops.

In conclusion, this study presents significant insights into how contextual linguistic features from LLMs, specifically Mistral-7B, can predict ECoG responses to speech in a multi-talker environment. By characterizing the contextual information encoded across attentional states, we gained a deeper understanding of auditory attention, which had previously ignored the contextual component of linguistic tracking. Our modeling approach could be applied to perform auditory attention

decoding, which usually uses acoustic speech features, or to further explore the neural mechanisms of attention with a broader electrode coverage and more participants.
On a broader scope, characterizing context encoding could be used to measure speech understanding objectively. One could, for instance, imagine evaluating the context encoding when listening to a foreign language or degraded speech.

Methods and Materials

Subjects and neural data

We used data from a previous study (***Choudhari et al., 2024***), including 3 subjects, of whom one (subject 1) was from Irving Medical Center at Columbia University, and two (subjects 2 and 3) were from North Shore University Hospital. All participants were undergoing clinical treatment for epilepsy and provided informed consent per the local Institutional Review Board (IRB) regulations. Neural data were collected as subjects performed the task with ECoG. Subject 1 only had sEEG depth electrodes implanted over their left brain hemisphere. Subjects 2 and 3 had sEEG depth electrodes as well as subdural ECoG grid electrodes implanted over the left hemispheres of their brains (see Figure 2).

Subject 1's neural data was recorded using Natus Quantum hardware at a sampling rate of 1024 kHz, and subjects 2 and 3's neural data were recorded using Tucker-Davis Technologies (TDT) hardware at a sampling rate of 1526 Hz. The left and right channels of the audio stimuli played to the participants via earphones were also recorded in sync with neural signals to facilitate the segmenting of neural data into trials for further offline analysis.

Neural data preprocessing

Following ***Choudhari et al.*** (***2024***), neural data were first resampled to 1000 Hz, then a common average reference was performed to reduce recording noise. The neural data were then further downsampled to 400 Hz. Line noise at 60 Hz and its harmonics (up to 180 Hz) were removed using a notch filter. The notch filter was designed using MATLAB's *fir2* function and applied using *filtfilt* with an order of 1000. To extract the envelope of the high gamma band (70 – 150 Hz), the neural data were first filtered with eight Gaussian filters, each with a width of 10 Hz, spaced consecutively between 70 and 150 Hz. The envelopes of the outputs of these filters were obtained by computing the absolute value of their Hilbert transform. The final envelope of the high gamma band was obtained by computing the mean of the individual envelopes yielded by the eight filters and further downsampling to 100 Hz.

Experiment Design

The experiment included 25 multi-talker trials, each with an average duration of 44.2 seconds (SD = 2.0 seconds), and 138 words on average (SD=6.86). As shown in Figure 1 (a), each trial featured two simultaneous and independent conversations, which were spatially separated and moved continuously within the frontal half of the subject's horizontal plane (azimuth range: -90 ° to +90 °). The intensity of both conversations remained equal and constant throughout the experiment; the participants adjusted the overall mixture volume to a comfortable level at the beginning of the task. Both conversations had the same power (RMS); all talkers in the conversations were native American English speakers. Diotic background noise (***Barker et al., 2015***; ***Spahr et al., 2012***) — either non-stationary noise (i.e., street noise, people chatter, honking) or stationary noise (babble)— was mixed in with the conversations at a power 9 or 12 dB lower than the conversation stream (25% of trials with babble -9 dB, 25% with babble -12 dB, 25% with pedestrian -9 dB and 25% with pedestrian -12 dB).

In these conversations, different talkers alternated speaking. As illustrated in Figure 1 (b), in the attended conversation, a talker switch occurred at the 50% mark of the trial duration. For the

unattended conversation, two talker switches happened—one around the 25% mark and the other near the 75% mark. To ensure subjects stay focused on the attended conversation and control for it behaviorally, a 1-back detection task was included in the experiment. Both conversations included occasionally repeated words strategically chosen to avoid overlap in time between the two conversation streams. A repeated word was created by duplicating the talker's voice waveform for that word. On average, the interval between the onset of two repeated words within a conversation was 7.0 seconds (SD = 1.0 seconds). The assignment of male and female talkers to different segments of the conversations was counterbalanced across trials to ensure equal durations of concurrent conversations with either the same or different genders.

Behavioral data

The push-button responses of subjects to repeated words in the conversation they were following were used to determine which conversation the subject was attending. A repeated word was considered correctly detected only if a button press occurred within two seconds of its onset. As reported in ***Choudhari et al.*** (***2024***), 67%, 75%, and 73% of the repeated words in the cued (attended) conversation were correctly detected for subjects 1, 2, and 3, respectively. In contrast, only 8%, 6%, and 1% of the repeated words in the uncued (unattended) conversation were correctly detected, suggesting that there were instances when subjects may have been attending to the unattended conversation as well.

Large language models to generate word representations

In this study, we conducted our analyses using Mistral-7B, a decoder-only transformer model comprising an initial token embedding layer followed by 32 transformer blocks, as described in its architecture overcaptureview (***Jiang et al., 2023***). Mistral-7B yielded the highest brain response prediction scores in a study comparing different open-source LLMs for generating word representations used as ECoG predictors (***Mischler et al., 2024***). Importantly, the initial token embedding layer does not carry positional or contextual information, and we refer to it as layer 0 in this study. We downloaded the model from the Hugging Face platform.

For each trial of both the attended and unattended speech streams, we fed the corresponding transcripts to Mistral-7B and extracted the embeddings of each layer. For each trial, the resulting dimensions were, therefore, $T \times D \times L$, with T the number of tokens, D the dimension of each embedding (i.e., 4096 for Mistral-7B), and L the number of layers (i.e., 32). For multi-token words, we selected the final token embedding to represent the word, as it captures the most complete contextual information. For experiments that varied the context provided to Mistral, we truncated the attention mask to include only the previous M tokens, M being the defined context length.

Ridge regression mapping from embeddings to neural responses.

To evaluate how predictive of different brain responses the LLM word representations are, we implemented ridge regressions to map the different layer embeddings of Mistral to different electrode responses in the brain of our subjects (see Figure 2). We selected 200 ms windows for each electrode to capture word responses, considering seven different lags starting at word onset, with a stride of 50 ms, thus covering a 500 ms span post-word-onset. We temporally averaged neural responses in frames of 200 ms and used ridge regression to predict (averaged) neural responses from the model embeddings as a function of lag. For each trial transcript, layer-specific word representations were computed with varying context lengths specified further.

Thus, for each trial, lag, and electrode, we fitted a ridge regression model with leave-one-trial-out cross-validation to predict the average word responses from the embeddings of each layer of Mistral-7B. The choice to leave out a trial minimizes the possibility of a context leakage that the

model could learn instead of the actual response-language relationship. For each training fold, we swept over a range of regularization parameters (log space between 10^{-2} and 10^{4}). Although there were small variations in brain reconstruction scores (see Supplementary Figure 6), we restricted our analyses to one lag: 200 to 400 ms post-word onset. The selected window has been shown to contain peaks of prediction scores from multiple language-related regions such as anterior superior temporal gyrus (aSTG), temporal pole (TP), Brodmann area 44 (BA44), and Brodmann area 45 (BA45) (***Hong et al., 2024***).

Context modulation

In this study, we manipulated the transcript multiple times before word embedding generation using Mistral-7B.

In Result 2, we compared the brain reconstruction scores using word representations with the original context presented to the participants to the word representations generated using an unrelated context. This experiment aims to show that context contributes to the brain reconstruction score, suggesting an encoding of contextual information in the brain. To generate word representations with an unrelated context, we identified for each word W the number of words K before it within the trial. We then selected K consecutive words from a different transcript ("the black willow", from the MAG-MESC dataset (***Gwilliams et al., 2023***)), adding W at the end of the word's sequence. We further fed this context to Mistral-7B to generate the representation of W with an unrelated context.

In Result 3, we varied the number of tokens fed to Mistral before each token of interest to investigate the impact of the context length used for word representation generation on the brain reconstruction scores. We tested the following lengths (in tokens): 1, 5, 10, 50, for both the attended and unattended conditions.

Acknowledgments

Sukru Samet Dindar for helping me familiarize myself with the dataset. Richard Antonello for the useful discussions and for providing the code base for the ridge regression optimization. Funding was provided by fellowships from Research Foundation - Flanders to Corentin Puffay (1S49823N) and Jonas Vanthornhout (1290821N). Gavin Mischler, Vishal Choudhari, and Nima Mesgarani received funding from the National Institutes of Health, the National Institute on Deafness and Other Communication Disorders, and the Marie-Josee and Henry R. Kravis Foundation.

Author Contributions

Vishal Choudhari (V.C.), Stephan Bickel (S.B.), Ashesh D. Mehta (A.D.M.), Catherine Schevon (C.S.) and Guy M. McKhann (G.M.M) contributed to data collection. Gavin Mischler Moore (G.M.M.), and V.C. contributed to conceptualization, methodology, and revision of the manuscript. Corentin Puffay (C.P.) contributed to the conceptualization, methodology, data analysis, and writing of the manuscript. Jonas Vanthornhout (J.V.) contributed to conceptualization, methodology, and revision of the manuscript. Tom Francart (T.F.) contributed to the revision of the manuscript. Nima Mesgarani (N.M.) supervised the project and contributed to conceptualization and methodology.

Data and Code Availability

The data that support the findings of this study are available on request from the corresponding author (N.M.). The data are not publicly available due to privacy or ethical restrictions. The code base to run our analyses is available on GitHub: https://github.com/corentinpuffay/LLM_ECoG.

Competing interests

The authors declare no competing interests.

References

Abnar S, Beinborn L, Choenni R, Zuidema W. Blackbox Meets Blackbox: Representational Similarity & Stability Analysis of Neural Language Models and Brains. In: Linzen T, Chrupała G, Belinkov Y, Hupkes D, editors. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* Florence, Italy: Association for Computational Linguistics; 2019. p. 191–203. https://aclanthology.org/W19-4820/, doi: 10.18653/v1/W19-4820.

Anderson AJ, Davis C, Lalor EC. Deep-learning models reveal how context and listener attention shape electrophysiological correlates of speech-to-language transformation. PLoS Comput Biol. 2024 Nov; 20(11):e1012537.

Anderson AJ, Kiela D, Binder JR, Fernandino L, Humphries CJ, Conant LL, Raizada RDS, Grimm S, Lalor EC. Deep Artificial Neural Networks Reveal a Distributed Cortical Network Encoding Propositional Sentence-Level Meaning. Journal of Neuroscience. 2021; 41(18):4100–4119. https://www.jneurosci.org/content/41/18/4100, doi: 10.1523/JNEUROSCI.1152-20.2021.

Antonello R, Huth A. Predictive Coding or Just Feature Discovery? An Alternative Account of Why Language Models Fit Brain Data. Neurobiology of Language. 2024 04; 5(1):64–79. https://doi.org/10.1162/nol_a_00087, doi: 10.1162/nol_a_00087.

Antonello R, Vaidya A, Huth A. Scaling laws for language encoding models in fMRI. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. *Advances in Neural Information Processing Systems*, vol. 36 Curran Associates, Inc.; 2023. p. 21895–21907. https://proceedings.neurips.cc/paper_files/paper/2023/file/4533e4a352440a32558c1c227602c323-Paper-Conference.pdf.

Barker J, Marxer R, Vincent E, Watanabe S. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In: *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*; 2015. p. 504–511. doi: 10.1109/ASRU.2015.7404837.

Brodbeck C, Hong LE, Simon JZ. Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. Current Biology. 2018 Dec; 28(24):3976–3983.e5. https://doi.org/10.1016/j.cub.2018.10.042, doi: 10.1016/j.cub.2018.10.042.

Broderick MP, Anderson AJ, Di Liberto GM, Crosse MJ, Lalor EC. Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. Current Biology. 2018; 28(5):803–809.e3. https://www.sciencedirect.com/science/article/pii/S0960982218301465, doi: https://doi.org/10.1016/j.cub.2018.01.080.

Caucheteux C, Gramfort A, King JR. Disentangling syntax and semantics in the brain with deep networks. In: Meila M, Zhang T, editors. *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of Proceedings of Machine Learning Research PMLR; 2021. p. 1336–1348. https://proceedings.mlr.press/v139/caucheteux21a.html.

Caucheteux C, Gramfort A, King JR. Evidence of a predictive coding hierarchy in the human brain listening to speech. Nature Human Behaviour. 2023 Mar; 7(3):430–441. https://doi.org/10.1038/s41562-022-01516-2, doi: 10.1038/s41562-022-01516-2.

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

Choudhari V, Han C, Bickel S, Mehta AD, Schevon C, McKhann GM, Mesgarani N. Brain-controlled augmented hearing for spatially moving conversations in multi-talker environments. Adv Sci (Weinh). 2024 Nov; 11(41):e2401379.

Deng Y, Choi I, Shinn-Cunningham B. Topographic specificity of alpha power during auditory spatial attention. NeuroImage. 2020; 207:116360. https://www.sciencedirect.com/science/article/pii/S1053811919309516, doi: https://doi.org/10.1016/j.neuroimage.2019.116360.

Edwards E, Soltani M, Kim W, Dalal SS, Nagarajan SS, Berger MS, Knight RT. Comparison of Time–Frequency Responses and the Event-Related Potential to Auditory Speech Stimuli in Human Cortex. Journal of Neurophysiology. 2009; 102(1):377–386. https://doi.org/10.1152/jn.90954.2008, doi: 10.1152/jn.90954.2008, pMID: 19439673.

Gillis M, Vanthornhout J, Simon JZ, Francart T, Brodbeck C. Neural markers of speech comprehension: Measuring EEG tracking of linguistic speech representations, controlling the speech acoustics. J Neurosci. 2021 Dec; 41(50):10316–10329.

Goldstein A, Zada Z, Buchnik E, Schain M, Price A, Aubrey B, Nastase SA, Feder A, Emanuel D, Cohen A, Jansen A, Gazula H, Choe G, Rao A, Kim C, Casto C, Fanda L, Doyle W, Friedman D, Dugan P, et al. Shared computational principles for language processing in humans and deep language models. Nat Neurosci. 2022 Mar; 25(3):369–380.

Gwilliams L, Flick G, Marantz A, Pylkkänen L, Poeppel D, King JR. Introducing MEG-MASC a high-quality magneto-encephalography dataset for evaluating natural speech processing. Scientific Data. 2023 Dec; 10(1):862. https://doi.org/10.1038/s41597-023-02752-5, doi: 10.1038/s41597-023-02752-5.

Hong Z, Wang H, Zada Z, Gazula H, Turner D, Aubrey B, Niekerken L, Doyle W, Devore S, Dugan P, Friedman D, Devinsky O, Flinker A, Hasson U, Nastase SA, Goldstein A. Scale matters: Large language models with billions (rather than millions) of parameters better match neural representations of natural language. eLife. 2024 Oct; http://dx.doi.org/10.7554/eLife.101204.1, doi: 10.7554/elife.101204.1.

Hosseini EA, Schrimpf M, Zhang Y, Bowman S, Zaslavsky N, Fedorenko E. Artificial Neural Network Language Models Predict Human Brain Responses to Language Even After a Developmentally Realistic Amount of Training. Neurobiology of Language. 2024 04; 5(1):43–63. https://doi.org/10.1162/nol_a_00137, doi: 10.1162/nol_a_00137.

Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, de las Casas D, Bressand F, Lengyel G, Lample G, Saulnier L, Lavaud LR, Lachaux MA, Stock P, Scao TL, Lavril T, Wang T, Lacroix T, Sayed WE, Mistral 7B; 2023. https://arxiv.org/abs/2310.06825.

Mesgarani N, Chang E. Selective cortical representation of attended speaker in multi-talker speech perception. Nature. 2012 04; 485:233–6. doi: 10.1038/nature11020.

Mischler G, Li YA, Bickel S, Mehta AD, Mesgarani N. Contextual feature extraction hierarchies converge in large language models and the brain. Nature Machine Intelligence. 2024 Dec; 6(12):1467–1477. https://doi.org/10.1038/s42256-024-00925-4, doi: 10.1038/s42256-024-00925-4.

O'Sullivan J, Herrero J, Smith E, Schevon C, McKhann GM, Sheth SA, Mehta AD, Mesgarani N. Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception. Neuron. 2019 Dec; 104(6):1195–1209.e3. https://doi.org/10.1016/j.neuron.2019.09.007, doi: 10.1016/j.neuron.2019.09.007.

Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* Stroudsburg, PA, USA: Association for Computational Linguistics; 2014. .

Puffay C, Vanthornhout J, Gillis M, Accou B, Van Hamme H, Francart T. Robust neural tracking of linguistic speech representations using a convolutional neural network. J Neural Eng. 2023 Aug; 20(4).

Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I, Robust Speech Recognition via Large-Scale Weak Supervision; 2022. https://arxiv.org/abs/2212.04356.

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. Proc Natl Acad Sci U S A. 2021 Nov; 118(45):e2105646118.

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. https://www.pnas.org/doi/abs/10.1073/pnas.2105646118, doi: 10.1073/pnas.2105646118.

Spahr AJ, Dorman MF, Litvak LM, Van Wie S, Gifford RH, Loizou PC, Loiselle LM, Oakes T, Cook S. Development and validation of the AzBio sentence lists. Ear Hear. 2012 Jan; 33(1):112–117.

Sun J, Wang S, Zhang J, Zong C. Neural Encoding and Decoding With Distributed Sentence Representations. IEEE Transactions on Neural Networks and Learning Systems. 2021; 32(2):589–603. doi: 10.1109/TNNLS.2020.3027595.

Toneva M, Wehbe L. In: Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain) Red Hook, NY, USA: Curran Associates Inc.; 2019. .