Imaging Neuroscience
an open access journal

Research Article

Large language models reveal the neural tracking of linguistic context in attended and unattended multi-talker speech

Corentin Puffay[a,b], Gavin Mischler[c], Vishal Choudhari[c], Jonas Vanthornhout[a], Stephan Bickel[d,e], Ashesh D. Mehta[d,e], Catherine Schevon[f], Guy M. McKhann[g], Hugo Van hamme[b], Tom Francart[a], Nima Mesgarani[c]

[a]KU Leuven, Department of Neurosciences, ExpORL, Leuven, Belgium
[b]KU Leuven, Department of Electrical Engineering (ESAT), PSI, Leuven, Belgium
[c]Columbia University, Department of Electrical Engineering, New York, NY, United States
[d]Hofstra Northwell School of Medicine, Uniondale, NY, United States
[e]The Feinstein Institutes for Medical Research, Manhasset, NY, United States
[f]Columbia University, Department of Neurology, New York, NY, United States
[g]Columbia University, Department of Neurological Surgery, Vagelos College of Physicians and Surgeons, New York, NY, United States

Corresponding Authors: Corentin Puffay (corentin.puffay@kuleuven.be), Nima Mesgarani (nima@ee.columbia.edu)

ABSTRACT

Large language models (LLMs) capture long-range contextual structure in natural language and have recently been shown to align with the human brain's contextualized linguistic encoding. This makes them a promising computational probe for studying how context-dependent linguistic information is represented during natural speech perception. Speech perception often occurs in multi-talker environments, where attention must dynamically select among competing streams, yet how contextual information from attended and unattended speech is neurally encoded remains underexplored. Here, we investigate how auditory attention modulates neural tracking of context-dependent linguistic representations using electrocorticography (ECoG) and stereoelectroencephalography (sEEG) recordings from three epilepsy patients engaged in a two-conversation "cocktail party" paradigm. To model neural responses to attended and unattended speech streams, we used contextual word embeddings generated by large language models. We find that LLM-derived features reliably predict brain activity for the attended stream and that contextual information from the unattended stream also contributes to neural prediction. Importantly, these contributions extend beyond low-level acoustic features and shallow syntactic information, and depend on the surrounding linguistic context. Moreover, neural tracking of the unattended stream reflects shorter-range contextual integration than that of the attended stream. Together, these findings indicate that neural responses to speech reflect context-dependent linguistic representations from multiple concurrent speech streams, with attention modulating the depth and timescale of contextual integration. Our results highlight the utility of LLMs for probing higher-level linguistic representations in complex, naturalistic listening environments.

Keywords: auditory attention, large language models, context encoding

1. INTRODUCTION

The emergence of Large Language Models (LLMs) and their high performance on natural language processing tasks has attracted increasing interest from neuroscience. LLMs capture long-range contextual structure in language, and recent work has shown that their representations align with human brain activity during language processing (Caucheteux et al., 2023; Goldstein

Received: 9 September 2025 Revision: 21 February 2026 Accepted: 10 April 2026 Available Online: 15 April 2026

The MIT Press
© 2026 The Authors. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

et al., 2022; Mischler et al., 2024; Schrimpf et al., 2021). This has motivated the use of LLM-derived embeddings as computational probes to study how language is encoded in the brain. A common approach is to extract word representations from LLMs and train linear regression models to predict corresponding neural responses (Abnar et al., 2019; Anderson et al., 2021; Caucheteux & King, 2022; Hosseini et al., 2024; Schrimpf et al., 2021; Sun et al., 2021; Toneva & Wehbe, 2019). The correlation between predicted and observed neural activity serves as a measure of how well the model's representations align with brain responses and allows researchers to investigate which aspects of language (i.e., lexical, syntactic, or semantic) are reflected in the brain response. However, LLM-derived embeddings contain multiple types of information: intrinsic properties of the word itself (e.g., lexical identity, frequency) as well as contextual information from preceding words. To disentangle these contributions, some studies have analyzed LLM activations using controlled manipulations of naturalistic transcripts (Caucheteux et al., 2021), while others have compared contextual embeddings to non-contextual embeddings derived from static models like GloVe (Goldstein et al., 2022; Pennington et al., 2014).

Most prior studies have examined neural encoding of language in controlled, single-speaker paradigms. Real-world listening, however, frequently involves multiple talkers, requiring attention to dynamically select one speech stream over others. Multi-talker environments introduce additional challenges for neural encoding: while low-level acoustic features are robustly tracked even under noisy conditions (Ding & Simon, 2012; Lalor & Foxe, 2010), higher-level linguistic representations rely on abstract, spatially localized, and temporally integrated processes (Hickok & Poeppel, 2007; Poeppel et al., 2009). Reduced acoustic signal-to-noise ratio can disproportionately disrupt these higher-level representations (Peelle et al., 2010). Noninvasive EEG studies suggest that attention is necessary for encoding contextual language features, with unattended speech typically showing minimal predictive power (Anderson et al., 2024; Brodbeck et al., 2018; Broderick et al., 2018). However, the low signal-to-noise ratio of EEG limits our ability to determine whether unattended speech is truly not encoded or simply not detectable. Invasive recordings such as electrocorticography (ECoG) and stereoelectroencephalography (sEEG) offer higher fidelity and spatial precision, enabling direct measurement of local high-frequency activity and more accurate characterization of context-dependent linguistic processing in complex listening scenarios (Chang et al., 2010; Mesgarani et al., 2014). These techniques allow us to investigate how attention shapes the encoding of contextual information in both attended and unattended speech streams.

Here, we use LLM embeddings to model neural responses measured with ECoG and sEEG while participants listen to two concurrent speech streams. First, we compare prediction scores across LLM layers for attended and unattended speech, showing that both streams predict the brain response, with higher power for the attended stream. Second, we manipulate the contextual input provided to the LLM, demonstrating that deeper layers improve prediction only when the preceding context matches the actual speech. Third, by varying the length of the context fed to the LLM, we show that neural responses track a shorter-range context for unattended speech than for attended speech. Finally, we show that LLM embeddings predict neural activity beyond low-level acoustic features. Together, these analyses reveal how attention shapes the depth and timescale of contextual linguistic encoding in the human brain.

2. METHODS AND MATERIALS

2.1. 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 (Fig. 1).

Subject 1's neural data were 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 audio channels played to participants were also recorded in sync with neural signals to facilitate segmenting neural data into trials for offline analysis. The trials were spatialized using head-related transfer functions (HRTFs) and delivered to the subjects via earphones.

2.2. 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 neu-

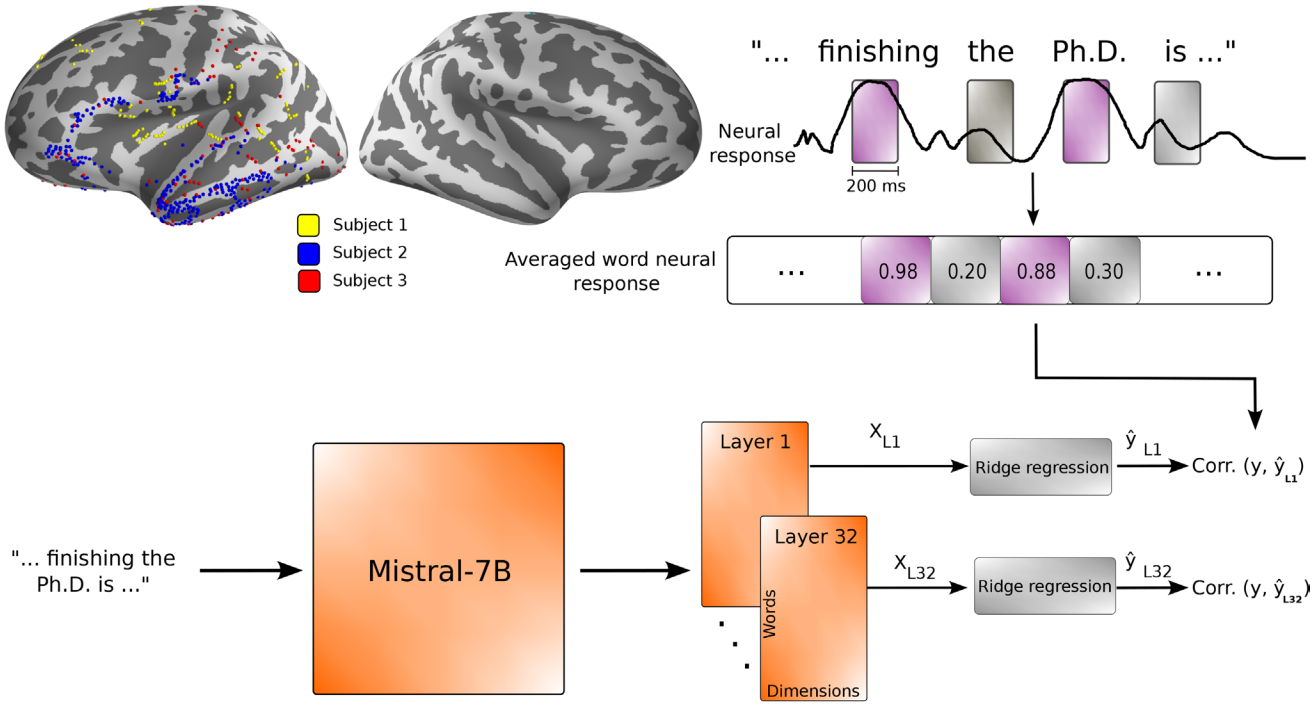

Fig. 1. LLM embeddings to predict brain responses. (Inspired from Mischler et al., 2024). We averaged the neural responses over 200-ms windows centered at seven lags, from 0 to 500 ms after word onsets. The word representations are extracted using different layers of Mistral-7B to further predict the averaged word response. The correlation between the predicted and actual responses is calculated to compute the brain score.

ral 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. A 1000th-order finite impulse response (FIR) notch filter was designed using MATLAB's *fir2* function, and zero-phase filtering was achieved with *filtfilt*. 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.

2.3. 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 2 (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 °), occasionally crossing each other. 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 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, a total of 8 different talkers (4 male, 4 female) alternated speaking. The assignment of genders to conversations was balanced to avoid bias (i.e., 25% male-male, 25% female-female, 25% female-male, and 25% male-female). As illustrated in Figure 2 (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. Participants were instructed at the start of each trial to attend to the conversation that began first and to maintain attention on that conversation throughout the trial, regardless of talker switches. They were told to press a button whenever they detected a repeated word in the attended conversation. Talker switches were not explicitly cued to participants; they were expected to follow the continuity

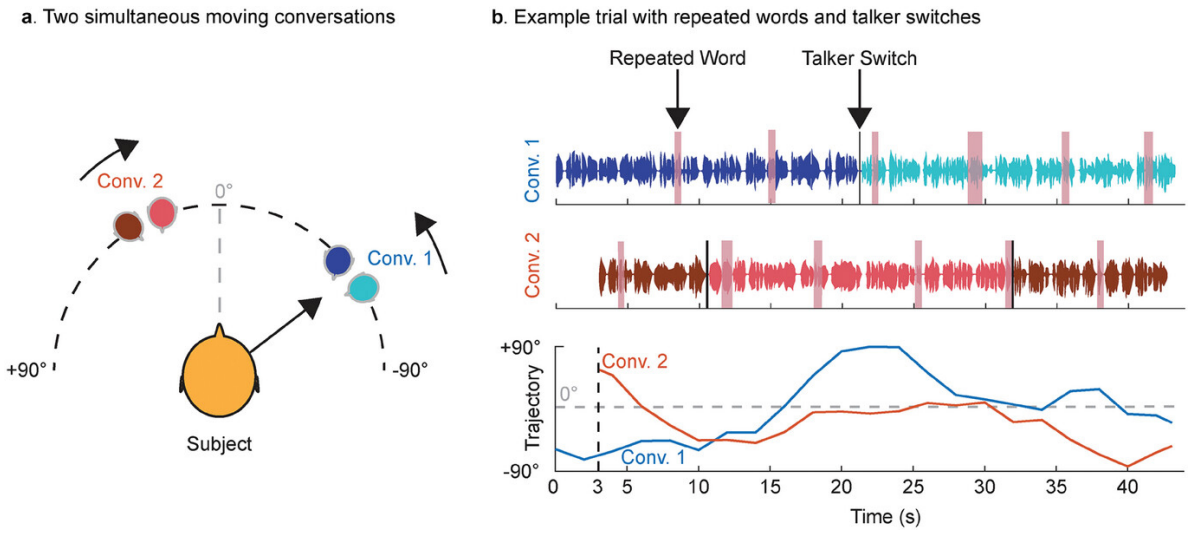

Fig. 2. Auditory attention experiment design from Choudhari et al., 2024. (a) Each trial consisted of two concurrent conversations presented simultaneously in the frontal hemifield. The two conversations (Conv. 1 and Conv. 2) moved independently and each involved two distinct talkers who alternated speaking. (b) In each conversation, talkers occasionally repeated a word at random intervals during a 1-back detection task (highlighted in pink). In the cued (attended) conversation, a single talker switch occurred at approximately 50% of the trial duration, whereas the uncued (unattended) conversation contained two talker switches, occurring at approximately 25% and 75% of the trial duration.

of the conversation rather than a specific talker's voice (Choudhari et al., 2024).

To ensure that subjects remained focused on the attended conversation and control for it behaviorally, a 1-back detection task was included in the experiment. Both conversations included occasional repetition of words, strategically chosen to avoid temporal overlap 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.

2.4. 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 2 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.

2.5. Large language models to generate word representations

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 the architecture and pre-training overview (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 response predictors (Mischler et al., 2024). Importantly, the initial token-embedding layer does not encode positional or contextual information; we refer to it as layer 0 in this study. We downloaded the model from the Hugging Face platform.

Embeddings extracted from Mistral-7B are interpreted as model-internal representations of word meaning that progressively integrate linguistic context across layers. While the initial token embedding layer (layer 0) reflects purely lexical information, successive transformer layers incorporate increasing amounts of contextual information from preceding tokens via self-attention. We deliberately used a pretrained, task-agnostic language model without task-specific fine-tuning, such that the extracted embeddings reflect general-purpose statistical regularities of natural language rather than representations optimized for the experimental paradigm.

This design allows us to investigate how neural activity tracks contextual and non-contextual language representations during naturalistic speech perception, without introducing task-dependent biases.

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.

2.6. Word envelopes as controls for acoustic encoding

To control for acoustic information that can be derived from LLM embeddings, we extracted the speech envelope for each word. To do so, we first generated the auditory spectrogram of the raw audio using the naplib-python package. The model used to generate the auditory spectrogram includes a cochlear filter bank of 128 logarithmically-spaced constant-Q filters, a hair cell model with a low-pass filter and a nonlinear compression function, and a lateral inhibitory network along the spectral axis. The envelope of each frequency band gives the time-frequency representation. To obtain the acoustic envelope, we averaged the auditory spectrogram over the frequency axis to obtain a one-dimensional speech envelope. For each word, we then extracted a lagged-envelope matrix consisting of 7 windows. The first window starts at the word onset and ends at the word length median (i.e., 250 ms); six lags are then extracted, with a stride of 50 ms. We selected the median word length as a reasonable fixed window length for most words, since our regression models do not accept variable-length inputs.

2.7. Ridge regression mapping from word 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 (Fig. 1). 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 over 200 ms frames and used ridge regression to predict (averaged) neural responses from 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 and electrode, we fitted a ridge regression model with leave-one-trial-out cross-validation to predict the average lagged word responses from the embeddings of each layer of Mistral-7B. The decision to exclude a trial reduces the risk of context leakage, which 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}). To allow the model to associate the word embeddings with different brain response delays, we combined the seven lags in our regression models. This choice was supported by an analysis across different time lags, which showed a stable encoding performance (Supplementary Fig. S7).

Additionally, we controlled for the encoding of acoustic information by combining the word embeddings with the corresponding acoustic envelope. For each word, we concatenated the LLM embedding and the lagged envelopes to obtain an input per word of dimension $D + nT$, with $D = 4096$ being the dimensions of Mistral-7B embeddings, $n = 7$ being the number of lags, and $T = 250$ being the number of time samples for each envelope window.

2.8. Context modulation

We manipulated the textual context provided to Mistral-7B prior to computing word representations, either by altering the semantic content of the preceding context or by varying its length. To generate word representations with an unrelated context, we identified for each word W the number of words K before it within a given 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. To examine the effect of context length, 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.

3. 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 performing a dual-speaker auditory attention task and examined how neural responses relate to contextualized language representations derived from attended and unattended speech streams.

In what follows, we present four key findings: (1) attention enhances brain prediction scores from LLM embeddings, (2) this effect is primarily driven by contextual information, (3) the encoded context length is shorter for the unattended than the attended speech, and (4) the information contained in LLM embeddings and predictive of the brain response is not merely acoustic but rather reflects contextual semantic information. Together, these results highlight attentional modulation of linguistic context processing in the brain.

3.1. 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 Figure 3a, 3c, and 3e, respectively. We observe

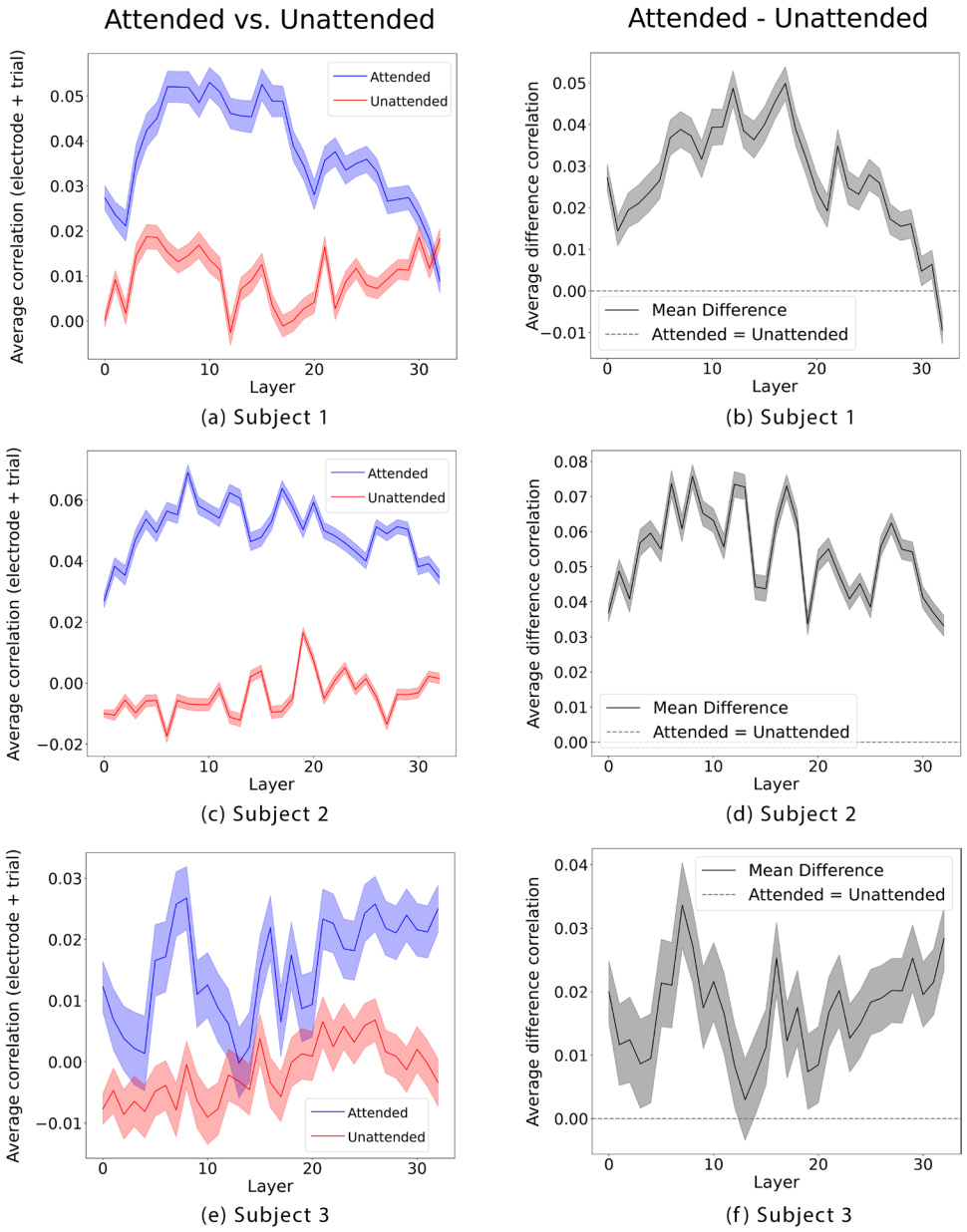

Fig. 3. Brain prediction scores of LLM word representations across layers with attention modulation. In (a), (c), and (e) (for subjects 1, 2, and 3, respectively), the two curves depict the average brain score across trials and electrodes for each layer, with SEM across trial-averaged electrodes. Blue: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

a significant difference in the averaged brain score magnitudes across layers between the attended and 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.

By definition, the output of Mistral-7B's layer 0 corresponds to token and positional embeddings only, without any contextual integration, while higher layers progressively integrate greater contextual information (Voita et al., 2019). Thus, as suggested in Caucheteux et al. (2021), improvements in brain reconstruction scores observed in subsequent layers likely reflect the accumulation of contextual information. We observe that the attended condition shows a difference in prediction score, for example, 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.

Figure 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).

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, that is, the difference in prediction scores between the attended and unattended conditions at layer 10 (Supplementary Fig. S1).

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

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

To verify that the prediction score's 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 Section 2.8 for details). We depict the resulting prediction scores for the original and unrelated context conditions in both the attended and unattended conditions in Figure 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

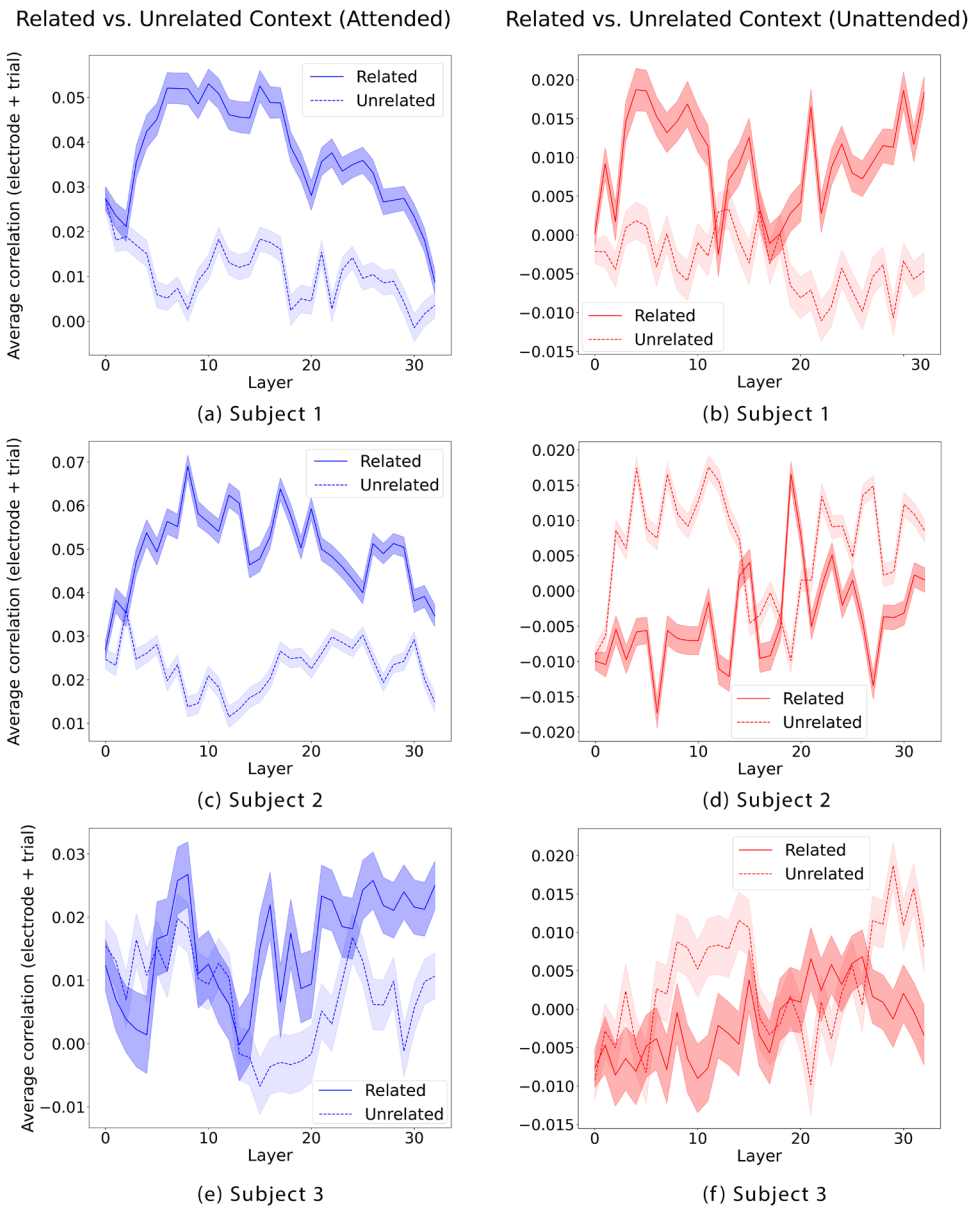

Fig. 4. Related vs. unrelated context prediction scores across Mistral layers for both attended and unattended conditions. (a), (c), and (e) depict (for subjects 1, 2, and 3, respectively) the related context (blue curve) and unrelated context (light blue curve) trial-electrode averaged prediction scores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

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 con-

text 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.

3.3. 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 select layer 10 for visualization because it lies well beyond the non-contextual layers while preceding the saturation or decline observed in later layers, thereby providing a representative snapshot of contextual encoding effects that are consistently observed across multiple higher layers. Results of neighboring layers from this plateau are reported in Supplementary Figure S6, 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, that is, 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-

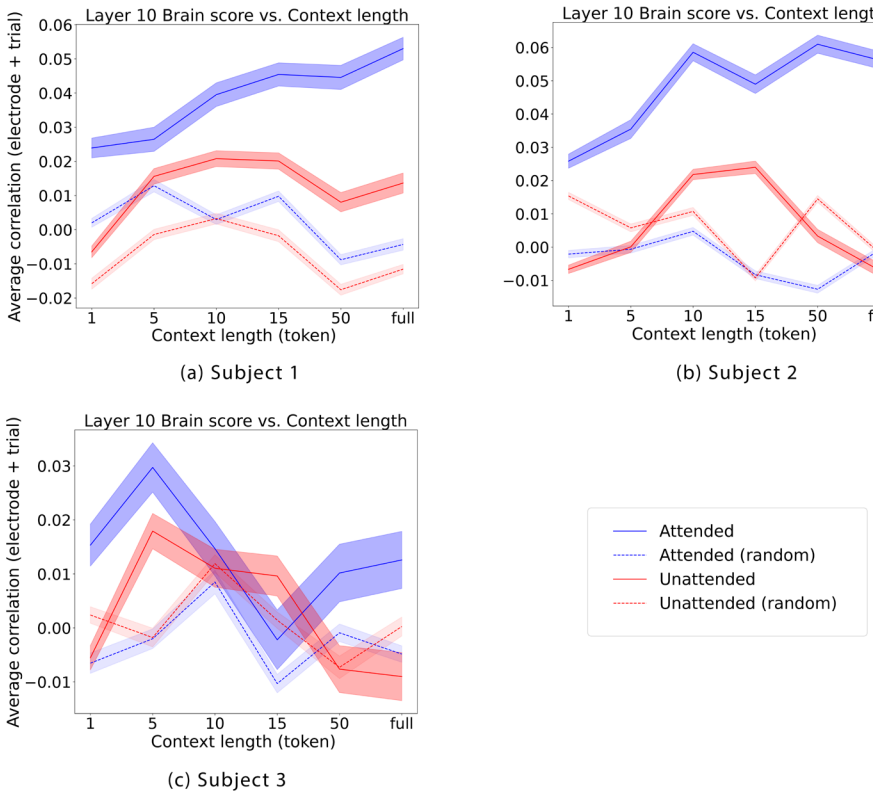

Fig. 5. Brain prediction scores of Mistral layer 10 embeddings across context lengths. (a), (b), and (c) depict (for subjects 1, 2, and 3, respectively) the attended (blue) and unattended (red) average prediction scores across trials and electrodes using different context lengths. The blue and red dashed 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.

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 in 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.

3.4. Result 4: Quantifying the acoustic confounds of the LLM prediction scores

Although LLM word embeddings are derived from text rather than raw audio, they may correlate with acoustic features indirectly via the linguistic structure (e.g., lexical choice, syntactic organization). To determine whether these correlations should be interpreted as evidence of shared higher-level structure rather than acoustic encoding per se, we combine the acoustic envelope of each word with its corresponding LLM embedding in a single regression task and compare the resulting prediction score with that obtained from the acoustic envelope alone. Our hypothesis is that any substantial improvement of the LLM embedding over the speech envelope can be attributed to semantic encoding unrelated to acoustic encoding. We only show results for the first 15 layers of Mistral-7B, as subsequent layers did not improve prediction scores in previous experiments (for more details, please refer to Section 4). Methodological details of the integration of the acoustic envelope into the regression task are reported in Section 2.

We depict the results for subjects 1, 2, and 3 in Figure 6.

For subjects 1 and 2, we see a significant benefit to the prediction score of the LLM embeddings beyond the acoustic envelope for the attended stream (two-sided Wilcoxon signed-rank test, subject 1: $W = 0$, $p < 0.001$; subject 2: $W = 7.0$, $p = 0.0011$, Bonferroni-corrected $\alpha = 0.05 / 6$). For subject 3, we found the opposite effect (two-sided Wilcoxon signed-rank test, $W = 9$, $p < 0.001$, Bonferroni-corrected $\alpha = 0.05 / 6$). This aligns with previous results showing no improvement in prediction scores across layers, possibly due to poor electrode coverage and responsiveness. We thus ignore subject 3 for the rest of this subsection.

For the unattended stream, the conclusions were more mixed when considering the maximal context for each word. For subject 1, a benefit of the LLM embeddings beyond acoustics was observed (two-sided Wilcoxon signed-rank test, $W = 9$, $p < 0.01$, Bonferroni-corrected $\alpha = 0.05 / 6$), while not for subject 2 (two-sided Wilcoxon signed-rank test, $W = 31$, $p = 0.11$).

Overall, we note that prediction scores obtained using the acoustic envelope alone are generally low, reflecting the relatively small number of electrodes with high scores compared with the model integrating LLM embeddings

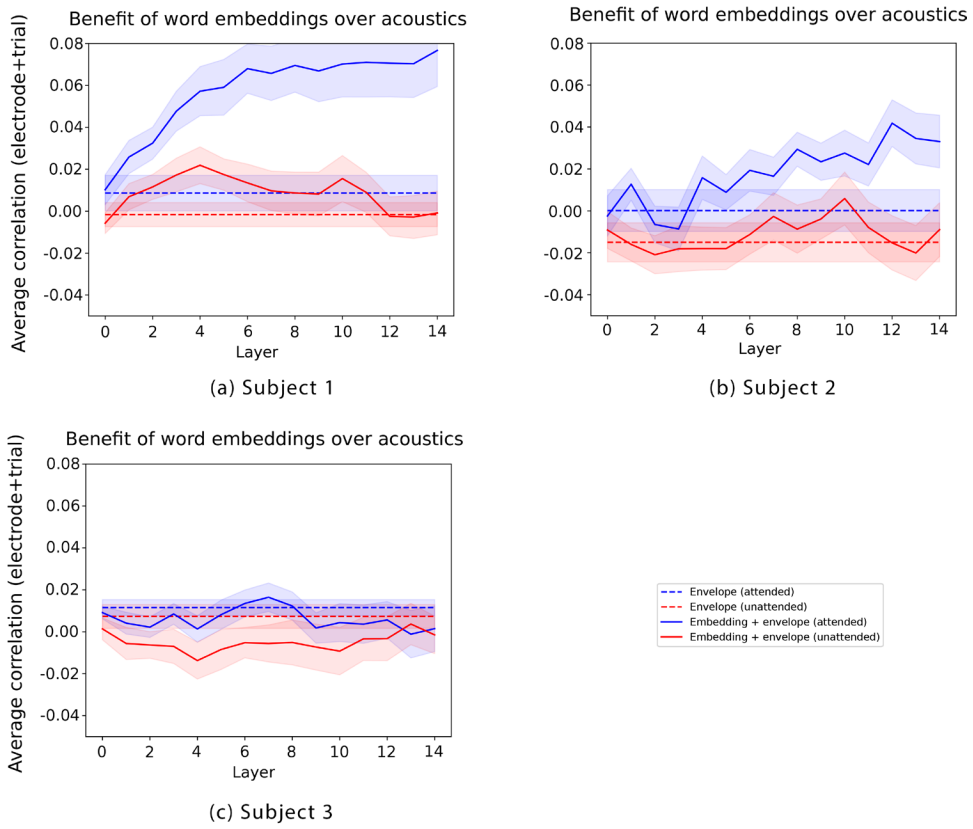

Fig. 6. Prediction scores obtained with the speech envelope only vs. the speech envelope combined with LLM embeddings generated with the full context for each word. (a), (b), and (c) depict (for subjects 1, 2, and 3, respectively) the attended (continuous blue) and unattended (continuous red) average prediction scores across trials and electrodes for the combination of LLM embeddings generated with the maximal context for each word and the corresponding acoustic envelope. The blue and red dashed lines correspond to the prediction scores obtained only with the attended and unattended acoustic envelopes, respectively. The corresponding shaded color also depicts the SEM across trial-averaged electrode prediction scores.

(Supplementary Fig. S8). This is in line with the minimal speech-responsive electrode selection that was made previously with this dataset (Choudhari et al., 2024), and gives compelling evidence that LLM embeddings not only carry semantic information beyond acoustics on average, but also lead to higher prediction scores for more electrodes. These results confirm that the prediction scores estimated in the previous experiments for subjects 1 and 2 stem from higher-level speech features than mere acoustic encoding when considering the attended stream. As hypothesized previously, the context length encoded for the unattended stream may be shorter (Fig. 5).

We thus depict, for the unattended stream, the benefit of LLM embeddings beyond acoustics, using a shorter context length of 10 tokens for subjects 1 and 2. We chose 10 tokens because it is the length at which the prediction score saturates in Figure 5 for subjects 1 and 2. We depict the results in Figure 7.

We observe a significant increase in the prediction score when adding LLM embeddings to the acoustic envelope at a context length of 10 tokens for subjects 1 and 2 (two-sided Wilcoxon signed-rank test, subject 1: $W = 2.0$, $p < 0.001$; subject 2: $W = 3.0$, $p < 0.001$, Bonferroni-corrected $\alpha = 0.05 / 6$). These results confirm that, at shorter context lengths, the prediction scores estimated in the previous experiments for subjects 1 and 2 are driven by higher-level speech features rather than by mere acoustic encoding when considering the unattended stream.

4. DISCUSSION

In this study, we explored the extent to which contextual LLM-based linguistic features can predict brain responses to speech in a multi-talker listening environment beyond acoustics. 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

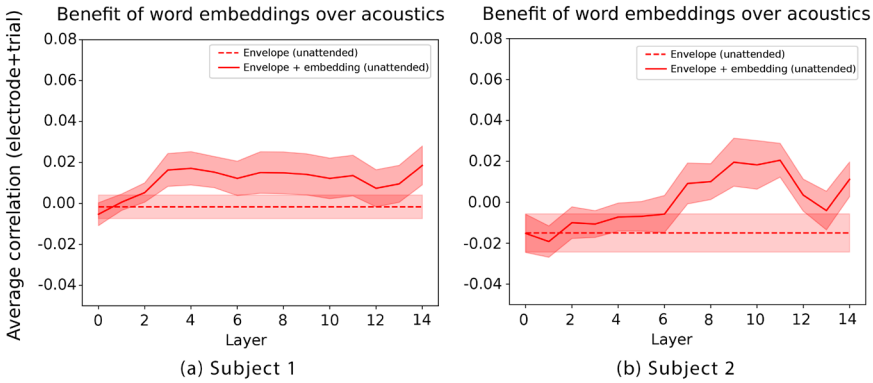

Fig. 7. Unattended stream prediction scores obtained with the speech envelope only vs. the speech envelope combined with LLM embeddings generated with a shorter context length (10 tokens) for each word. (a) and (b) depict (for subjects 1 and 2, respectively) the unattended (continuous red) average prediction scores across trials and electrodes for the combination of LLM embeddings generated with a 10-tokens context length for each word and the corresponding acoustic envelope. The red dashed line corresponds to the prediction scores obtained with the unattended acoustic envelope only. The corresponding shaded color also depicts the SEM across trial-averaged electrode prediction scores.

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 (Antonello et al., 2023; Caucheteux et al., 2023) 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 to estimate the processing of linguistic features.

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 embedding of the word 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.

Taken together, the layer-wise analysis (Result 1) and the context-manipulation experiment (Result 2) suggest that the predictive power of LLM embeddings does not arise from lexical or part-of-speech information alone, but from representations that depend on preceding linguistic context. Because the identity and syntactic category of the target word were preserved while the surrounding context was altered, the observed drop in prediction scores indicates sensitivity to context-dependent representations. This contextual sensitivity plausibly reflects semantic integration processes, extending beyond shallow syntactic features or low-level acoustic tracking.

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 tracking 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 (Goldstein et al., 2022; Mischler et al., 2024), 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 (Supplementary Fig. S5a), 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 (Supplementary Fig. S5b-d).

In Result 4, we quantified the added value of LLM-generated word embeddings over their acoustic envelope, to ensure that the prediction scores obtained with embeddings were not only reflecting acoustic processing, but also semantic. We found a significant benefit of LLM embeddings beyond acoustics for both the attended and unattended streams, with a shorter context length encoding for the unattended stream, which confirmed our findings that non-acoustic linguistic processing happened with and without attention. Interestingly, the prediction score values were almost identical between the embeddings from layer 0 and the acoustic envelope. Considering that the speech envelope and Mistral-7B's layer 0 carry information about the word of interest only, this reinforces the claim that the prediction score increase observed across layers may reflect contextual processing. This result aligns with a previous study (Caucheteux et al., 2021), using embeddings from layer 0 as non-contextual lexical controls.

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 (Anderson et al., 2024; Radford et al., 2022), allowing us to incorporate acoustic information while adjusting for context in the word representations.

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 (Supplementary Fig. S4). Our word detection task occurs every 7 seconds, 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 second 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 (Bhaya-Grossman et al., 2025) or degraded speech.

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.

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.

FUNDING

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.

DECLARATION OF COMPETING INTEREST

The authors declare no competing interests.

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.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available with the online version here: https://doi.org/10.1162/IMAG.a.1227#supplementary-data

REFERENCES

Abnar, S., Beinborn, L., Choenni, R., & Zuidema, W. (2019). Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In T. Linzen, G. Chrupała, Y. Belinkov, & D. Hupkes (Eds.), *Proceedings of the 2019 ACL workshop blackboxnlp: Analyzing and interpreting neural networks for nlp* (pp. 191–203). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4820

Anderson, A. J., Davis, C., & Lalor, E. C. (2024). Deep-learning models reveal how context and listener attention shape electrophysiological correlates of speech-to-language transformation. *PLoS Computational Biology*, *20*(11), e1012537. https://doi.org/10.1371/journal.pcbi.1012537

Anderson, A. J., Kiela, D., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Raizada, R. D. S., Grimm, S., & Lalor, E. C. (2021). Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *Journal of Neuroscience*, *41*(18), 4100–4119. https://doi.org/10.1523/JNEUROSCI.1152-20.2021

Antonello, R., Vaidya, A., & Huth, A. (2023). Scaling laws for language encoding models in fMRI. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (pp. 21895–21907, Vol. *36*). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/4533e4a352440a32558c1c227602c323-Paper-Conference.pdf

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

Bhaya-Grossman, I., Leonard, M. K., Zhang, Y., Gwilliams, L., Johnson, K., Lu, J., & Chang, E. F. (2025). Shared and language-specific phonological processing in the human temporal lobe. *Nature*, *649*, 140–151. https://doi.org/10.1038/s41586-025-09748-8

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, *28*(24), 3976–3983.e5. https://doi.org/10.1016/j.cub.2018.10.042

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803–809.e3. https://doi.org/10.1016/j.cub.2018.01.080

Caucheteux, C., Gramfort, A., & King, J.-R. (2021). Disentangling syntax and semantics in the brain with deep networks. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning* (pp. 1336–1348). PMLR. https://proceedings.mlr.press/v139/caucheteux21a.html

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

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

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, *13*(11), 1428–1432. https://doi.org/10.1038/nn.2641

Choudhari, V., Han, C., Bickel, S., Mehta, A. D., Schevon, C., McKhann, G. M., & Mesgarani, N. (2024). Brain-controlled augmented hearing for spatially moving conversations in multi-talker environments. *Advanced Science (Weinheim)*, *11*(41), e2401379. https://doi.org/10.1002/advs.202401379

Deng, Y., Choi, I., & Shinn-Cunningham, B. (2020). Topographic specificity of alpha power during auditory spatial attention. *NeuroImage*, *207*, 116360. https://doi.org/10.1016/j.neuroimage.2019.116360

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859. https://doi.org/10.1073/PNAS.1205381109

Gillis, M., Vanthornhout, J., Simon, J. Z., Francart, T., & Brodbeck, C. (2021). Neural markers of speech comprehension: Measuring EEG tracking of linguistic speech representations, controlling the speech acoustics. *Journal of Neuroscience*, *41*(50), 10316–10329. https://doi.org/10.1523/jneurosci.0812-21.2021

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

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

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2024). Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, *5*(1), 43–63. https://doi.org/10.1162/nol_a_00137

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7b. *arXiv*. https://arxiv.org/abs/2310.06825

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience*, *31*(1), 189–193. https://doi.org/10.1111/j.1460-9568.2009.07055.x

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1010. https://doi.org/10.1126/science.1245994

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

O'Sullivan, J., Herrero, J., Smith, E., Schevon, C., McKhann, G. M., Sheth, S. A., Mehta, A. D., & Mesgarani, N. (2019). Hierarchical encoding of attended auditory objects in multi-talker speech perception. *Neuron*, *104*(6), 1195–1209.e3. https://doi.org/10.1016/j.neuron.2019.09.007

Peelle, J. E., Johnsrude, I., & Davis, M. H. (2010). Hierarchical processing for speech in human auditory cortex and beyond. *Frontiers in Human Neuroscience*, *4*, 51. https://doi.org/10.3389/fnhum.2010.00051

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

Poeppel, D., Idsardi, W. J., & van Wassenhove, V. (2009). Speech perception at the interface of neurobiology and linguistics. In B.C.J. Moore (Ed.), *The perception of speech: From sound to meaning* (pp. 249–274). Oxford University Press. https://doi.org/10.1093/oso/9780199561315.003.0011

Puffay, C., Vanthornhout, J., Gillis, M., Accou, B., Van Hamme, H., & Francart, T. (2023). Robust neural tracking of linguistic speech representations using a convolutional neural network. *Journal of Neural Engineering*, *20*(4). https://doi.org/10.1088/1741-2552/acf1ce

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv*. https://arxiv.org/abs/2212.04356

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences United States of America*, *118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Spahr, A. J., Dorman, M. F., Litvak, L. M., Van Wie, S., Gifford, R. H., Loizou, P. C., Loiselle, L. M., Oakes, T., & Cook, S. (2012). Development and validation of the AzBio sentence lists. *Ear and Hearing*, *33*(1), 112–117. https://doi.org/10.1097/aud.0b013e31822c2549

Sun, J., Wang, S., Zhang, J., & Zong, C. (2021). Neural encoding and decoding with distributed sentence representations. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 589–603. https://doi.org/10.1109/TNNLS.2020.3027595

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Proceedings of the 33rd international conference on neural information*

processing systems. Curran Associates Inc. https://doi.org/10.5121/csit.2019.91200

Voita, E., Sennrich, R., & Titov, I. (2019). The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 4396–4406). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1448