bioRxiv preprint doi: https://doi.org/10.1101/2025.04.24.648897; this version posted April 24, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 Megarani**[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***) typically train linear regression using word representations generated by the LLMs to predict the corresponding neural responses. The correlation between the predicted and the actual neural response is then used as a metric for language modeling performance. Such metric can also be used

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.24.648897; this version posted April 24, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

to investigate how the brain response relates to language processing.

However, the linear relationship observed between word representations and the brain response can be explained by various factors whose influence is not well established. Using a more controlled dataset (e.g., with invariant syntax but varying words), researchers factorized the word representations into lexical, compositional, syntactic, or semantic representations (***Caucheteux et al., 2021***) to understand the language features used by models. To separate the contextual from non-contextual information contained in word representations, (***Goldstein et al., 2022***) used GloVe (***Pennington et al., 2014***), a model generating these features purely based on words' frequency derived from a corpus, not relying on the position of words in a text.

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 ASR model Whisper (***Radford et al., 2022***) to predict EEG responses to continuous speech in a two-speaker scenario. They found that late Whisper 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 context is provided to our LLM versus not, and show that context is necessary to improve the prediction scores in deeper layers. Finally, by modulating the context fed to the LLM, we show that the brain 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

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.24.648897; this version posted April 24, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

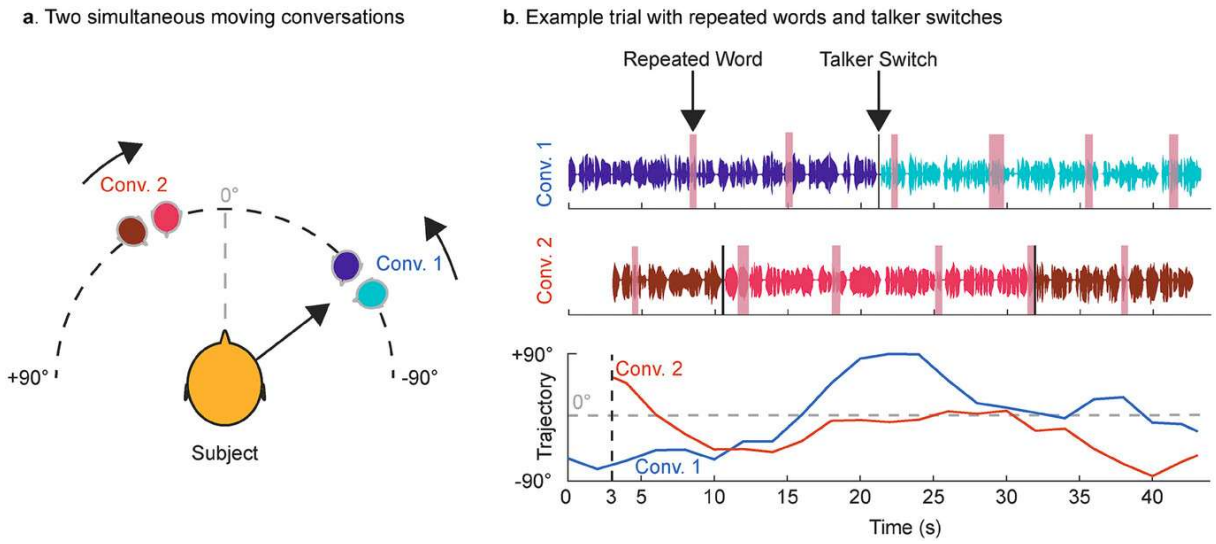


Figure 1. AAD experiment design from ***Choudhari et al.*** (***2024***). a) Every trial consisted of two concurrent conversations moving independently in the front hemifield of the subject. Each conversation had two distinct talkers taking turns. b) Talkers in each conversation repeated a word at random 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.

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 embeddings 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: LLM brain prediction scores are impacted by attention

In Figure 3, we depict the LLM word embedding 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 that the averaged brain score magnitudes for the full context condition across layers are higher in the attended condition than in the unattended one (Wilcoxon signed-rank test: $W = 3, p < 0.001; W = 0, p < 0.001; W = 0, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$, for subject 1, 2 and 3, respectively), suggesting differences in the contextual information encoded across attention states.

Specifically, we observe that the attended condition shows a prediction score increase, e.g., from layer 0 to layer 6 (Wilcoxon signed-rank test, $W = 1385, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$), from layer 0 to 7 (Wilcoxon signed-rank test, $W = 4123, p < 0.001$, Bonferroni-corrected

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.24.648897; this version posted April 24, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

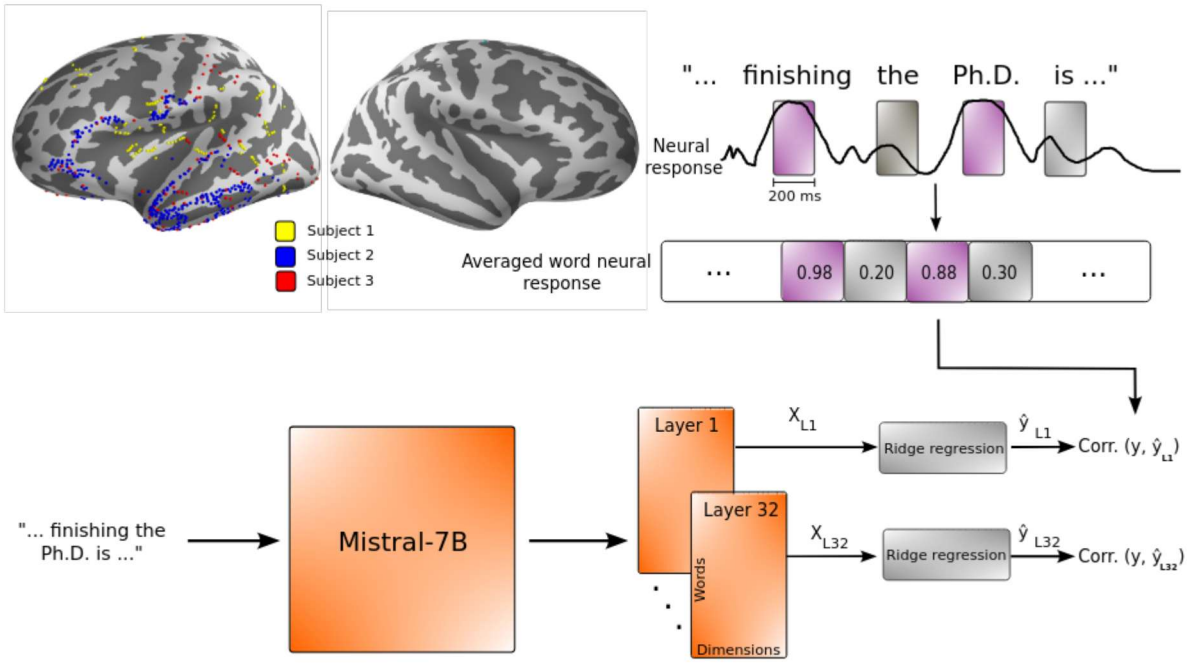
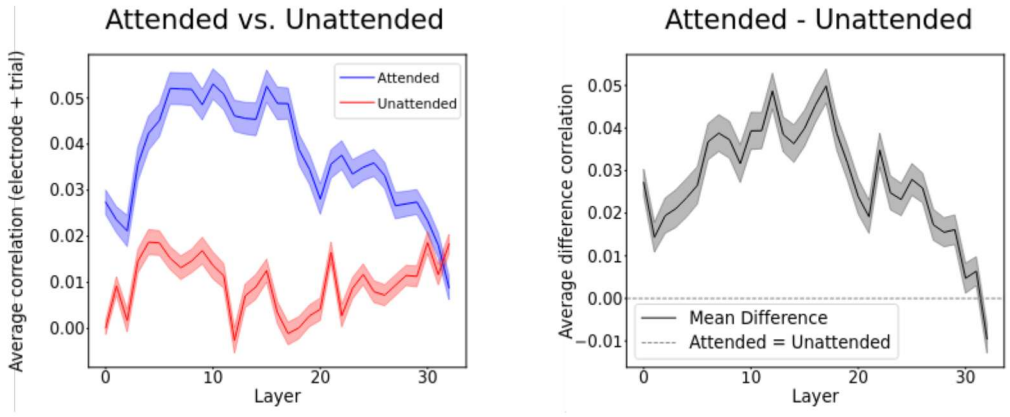


Figure 2. LLM embeddings to predict brain responses. (Inspired from *Mischler et al.* (*2024*)). We averaged the neural responses over 200-ms windows centered at seven lags, from 0 to 500 ms after word onsets. The word embeddings are extracted using different layers of Mistral-7B to predict the averaged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

$\alpha = 0.05/6$) and from layer 0 to 8 (Wilcoxon signed-rank test, $W = 1526, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$) for subjects 1, 2, and 3, respectively. Results are similar for neighboring layers and are reported in Supplementary Table 1. This informs us that the contextual information contained in layers 6, 7, and 8 for subjects 1, 2, and 3, respectively, enables a better prediction of the average brain response than in earlier layers.

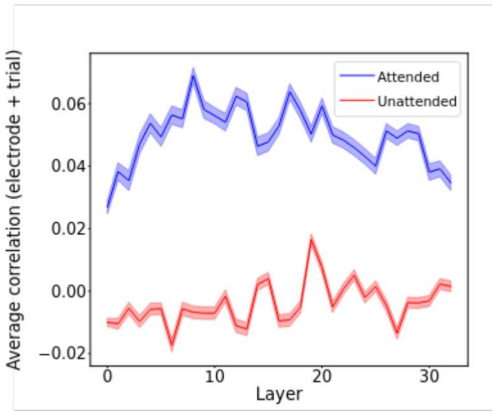
In comparison, we observe a (lower) magnitude increase for the unattended transcript in subject 1 from layer 0 to 6 (Wilcoxon signed-rank test, $W = 2917, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$), while no such increase is observed for the unattended case in subject 2 from layer 0 to 7 (Wilcoxon signed-rank test, $W = 21951, p = 0.24$) and subject 3 from layer 0 to 8 (Wilcoxon signed-rank test, $W = 21951, p = 0.042$, not significant after Bonferroni correction, $\alpha = 0.05/6$). Results are similar for neighboring layers and are reported in Supplementary Table 2. This finding suggests a better contextual encoding of the attended stream, while there might be an unattended speech 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).

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.24.648897; this version posted April 24, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

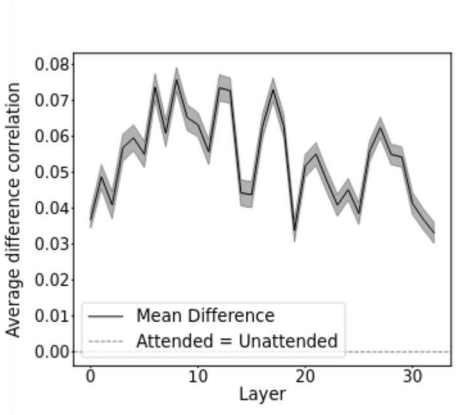


(a) Subject 1

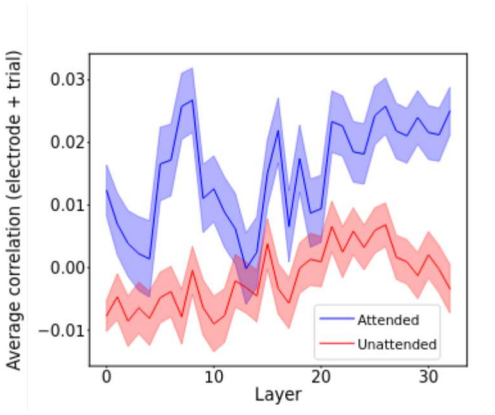
(b) Subject 1



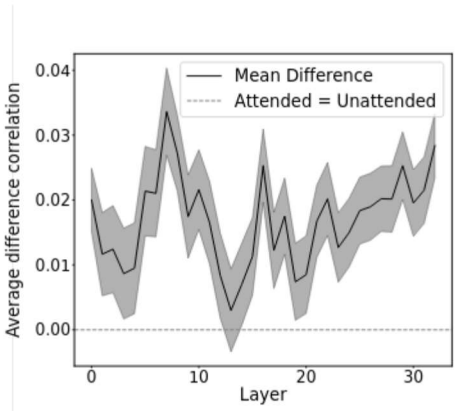
(c) Subject 2



(d) Subject 2



(e) Subject 3



(f) Subject 3

Figure 3. Brain prediction scores of LLM word embeddings 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.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.24.648897; this version posted April 24, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

In the Supplementary Material, we run two additional analyses: (1) to evaluate the impact of repeated words, (2) to evaluate any speaker-specific effects that our model could learn and that could impair our conclusions.

Result 2: Context tracking is present with and without attention

To verify that the prediction score's increase across Mistral's layers is due to contextual information, we investigate how varying 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 full trial context before the current token, as done in the previous subsection, or solely the current token. We depict the resulting brain scores for all three subjects in both the attended and unattended conditions in Figures 4a, 4c, and 4e for subjects 1, 2, and 3, respectively.

For all subjects, we observe that for the attended transcripts with the full context, the brain score increases across layers (for statistics, see the previous subsection). At the same time, the condition in which no context is provided shows a decrease or no change in brain score across layers. For example, this occurs for subject 1 between layers 0 and 6 (Wilcoxon signed-rank test, $W = 4673, p < 0.001$, Bonferroni-corrected, $\alpha = 0.05/6$). The prediction scores were not significantly different for subject 2 between layer 0 and 7 (Wilcoxon signed-rank test, $W = 20201, p = 0.022$, not significant after Bonferroni correction, $\alpha = 0.05/6$), and for subject 3 between 0 and 8 (Wilcoxon signed-rank test, $W = 2659, p = 0.06$, not significant after Bonferroni correction, $\alpha = 0.05/6$). Results are similar for neighboring layers and are reported in Supplementary Table 3. This finding suggests that the increased prediction score observed in the previous section for the attended case is likely due to contextual information.

In the unattended condition, as seen in the previous section, we observe a score increase across layers in the condition with full context for subject 1. In contrast, no such behavior is observed without context (e.g., between layers 0 and 6 or 7 or 8: Wilcoxon signed-rank test, $W = 5905, W = 21638, W = 2622$, not significant after Bonferroni correction, $\alpha = 0.05/6$ for subjects 1, 2, and 3, respectively). Results are similar for neighboring layers and are reported in Supplementary Table 4. Although the superiority of the full context condition is, in the unattended case, quite erratic and does not happen in all layers, it suggests that the origin of the prediction score increase across layers in subject 1 is likely due to contextual information encoding.

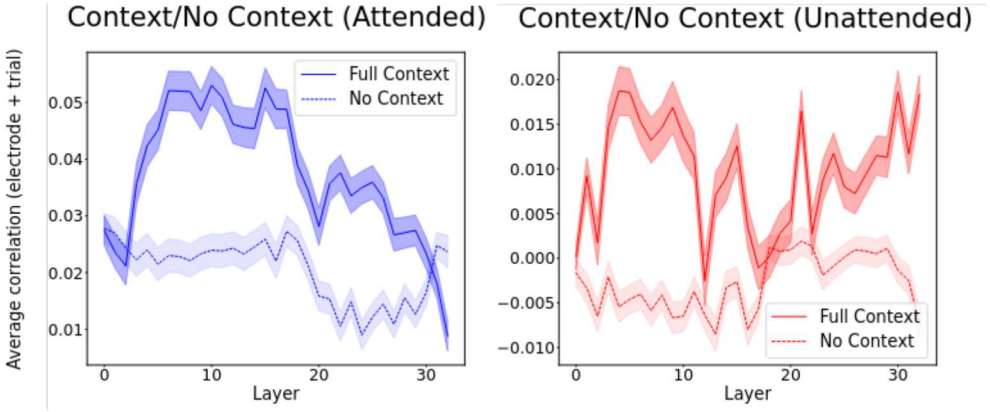
Result 3: Length of context tracking varies with attention

In Subsection , we performed a binary analysis, comparing word embeddings when all tokens before the token of interest were fed to Mistral, as opposed to word embeddings with only the token of interest fed to Mistral, excluding context.

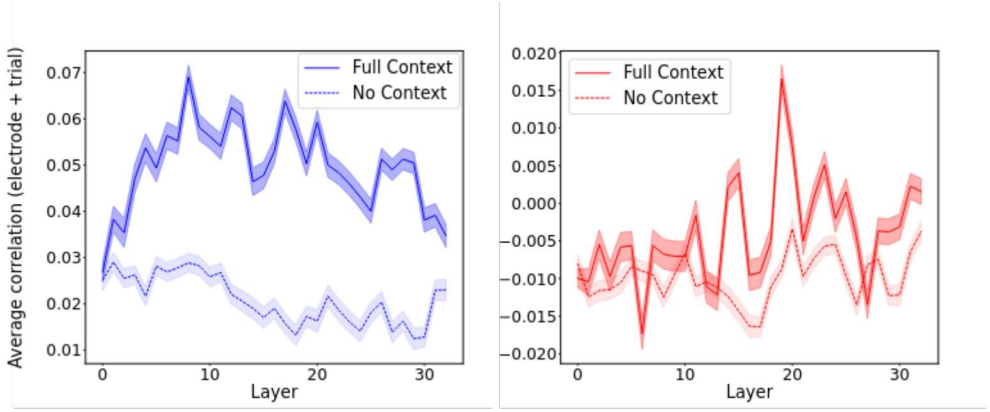
Here, we characterize the longest context length to which the model benefits in both the attended and unattended conditions. From now on, we depict our results for layer 10, as it is part of the plateau reached for the attended curve in Figure 3a and 3c. Neighboring layers from this plateau are not shown in this study (see Supplementary Figures 5 and 6) 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 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.

We observe that the average brain score increases steadily with the context length in the attended condition (i.e., a significant increase from 15 tokens to the full context condition, Wilcoxon signed rank test, $W = 4166, W = 14259, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$, for subject 1 and 2, respectively). This steady increase did not happen for subject 3 beyond five tokens. In the

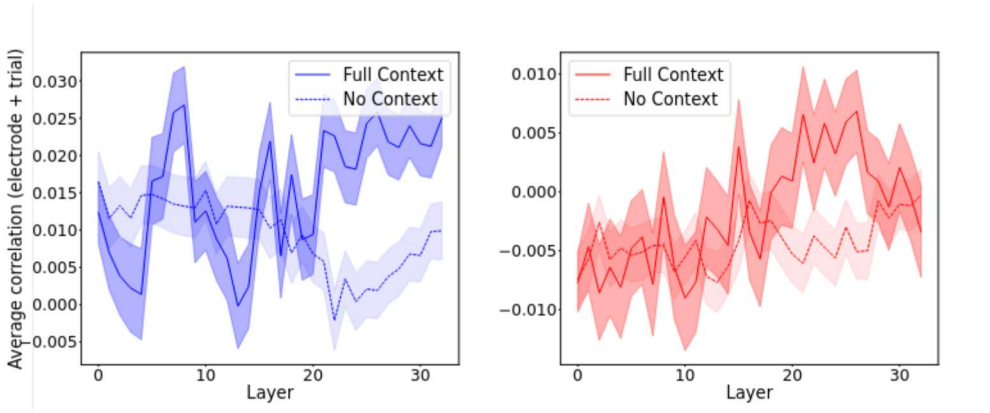
bioRxiv preprint doi: https://doi.org/10.1101/2025.04.24.648897; this version posted April 24, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



(a) Subject 1 **(b)** Subject 1



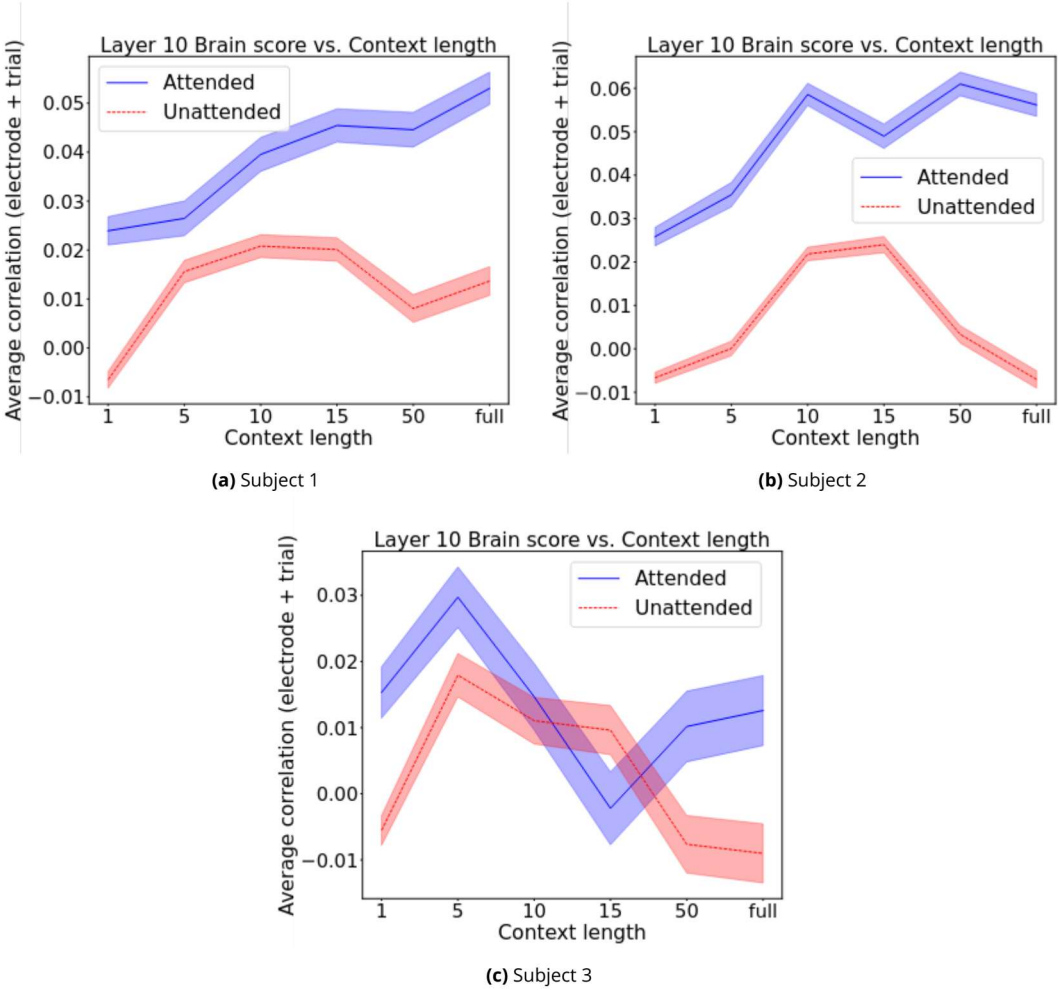
(c) Subject 2 **(d)** Subject 2



(e) Subject 3 **(f)** Subject 3

Figure 4. Full vs. no context brain scores across Mistral layers for both attended and unattended conditions. (a), (c) and (e) depict (for subjects 1, 2, and 3, respectively) the full context (blue curve) and no context (light blue curve) trial-electrode averaged brain scores across Mistral layers when using the attended word embeddings. (b), (d) and (f) depict (for subjects 1, 2, and 3, respectively) the full context (blue curve) and no context (light blue curve) trial-electrode averaged brain scores across Mistral layers when using the unattended word embeddings.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.24.648897; this version posted April 24, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



(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 brain scores across trials and electrodes using different context lengths. The corresponding shaded color also depicts the SEM across trial-averaged electrode brain scores.

unattended condition, the brain score increase stops at 10 tokens (no significant increase from 10 to 15 tokens, Wilcoxon signed rank test, $W = 4603, W = 10842, p = 0.23, p = 0.06$ for subjects 1 and 2, respectively). Then it decreases (significant decrease from 15 to 50 tokens, Wilcoxon signed rank test, $W = 2936, W = 7593, p < 0.001, p < 0.001$, Bonferroni-corrected $\alpha = 0.05/6$, for subjects 1 and 2, respectively). This steady increase did not happen for subject 3 beyond five tokens. 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.

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 infor-

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.24.648897; this version posted April 24, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

mation 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 brain 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 highlighted that the model uses contextual information to better predict the brain response. We compared the brain score across layers with either the full trial context fed to Mistral or only the token of interest. We found no brain score increase over layers in the latter case. Although Mistral-7B has also been shown to have predictive power for ECoG responses to single-talker speech, and to benefit from contextual information (***Mischler et al., 2024***), our approach for this second result can be criticized. Mistral is trained to learn a representation of words within a given context, and there is uncertainty about what information remains in the word-embedded representation if only one token is provided. An alternative would be to use models that are not optimized on contextual information. For example, ***Goldstein et al.*** (***2022***) implemented a similar regression task, using non-contextual (static) word embeddings generated with the GloVe model (***Pennington et al., 2014***), and showed significant brain prediction scores without contextual information. As previously mentioned, ***Anderson et al.*** (***2024***) used the speech envelope and its first derivative, which could also be used as a non-contextual acoustic control. An obvious disadvantage would be the acoustic dominance of these features, and a larger uncertainty about what information is being used by the model to predict the brain response, while static word embeddings provided by the layer 0 of Mistral-7B, or the GloVe model only contain the word onset timing and the word frequency.

Despite this caveat, the present result reveals two important points: (1) a prediction score increase past layer 0 is due to context, and (2) the difference in layer 0 between attended and unattended without context shows that there is a difference in non-contextual information encoding across attentional states.

In Result 3, we explored multiple context lengths to characterize if 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

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.24.648897; this version posted April 24, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

increase of the brain 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. Two potential explanations are: (1) the number of electrodes is minimal in the frontal lobe (see Supplementary Figure 3) which contains areas encoding context (***Caucheteux et al., 2023***), and (2) the electrodes are less responsive, leading to poorer brain scores than in subjects 1 and 2 (see Supplementary Figure 4).

Attention and context are complex phenomena that, when brought together, pose a lot of questions. 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 embeddings.

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 embeddings. 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 embeddings (***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, a more controlled dataset could be used (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***).

In conclusion, this study presents significant insights into how contextual linguistic features

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.24.648897; this version posted April 24, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 was first resampled to 1000 Hz, then a common average reference was performed to reduce recording noise. The neural data was 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 28 multi-talker trials, each with an average duration of 44.2 seconds (SD = 2.0 seconds). 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%

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.24.648897; this version posted April 24, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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. Both conversations included occasionally repeated words as part of a 1-back detection task, with these 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 did our analyses using Mistral-7B (***Jiang et al., 2023***), which led to the highest brain response prediction scores in a study comparing different open source LLMs to generate word embeddings used as ECoG predictors (***Mischler et al., 2024***). Mistral-7B is an LLM based on a transformer architecture (***Vaswani et al., 2017***). It possesses seven billion parameters and 32 layers. 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). We picked the last token for multi-token words to align our token embeddings with words, since it contains the full word's context. 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 embeddings 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 embeddings 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

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.24.648897; this version posted April 24, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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}). We restricted our analyses to the one lag leading to the highest average prediction scores across layers for every subject: 200 to 400 ms post-word onset.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.24.648897; this version posted April 24, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.24.648897; this version posted April 24, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

499 **Vaswani A**, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you
500 need. . 2017; .