

# Natural language processing models reveal neural dynamics of human conversation

Corresponding Author: Dr Jing Cai

**This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.**

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Comments:

This study reports that natural language processing models revealed neural dynamics of human conversation. While the experimental procedures of the study appear to be well thought out and executed, some flaws exist in the processing and interpretation of the data which need to be addressed. These concerns and others are explained below:

Major points:

1. The authors did not correct for multiple comparisons across multiple brain regions throughout the manuscript (e.g., Fig. 2A). They should address this point.
2. The authors did not describe the limitations of this study. They should describe at least two limitations to increase the reliability of the study.
3. The authors should examine whether the neural response is correlated not only to the IQ score (Fig. S4) but also to age and sex.

Minor points:

1. In the legend of Fig. 1, the authors describe as follows: "A total of 1910 bipolar-referenced channel channels that demonstrated a low rate of epileptiform discharges were primarily distributed in the frontal, temporal and mesial areas in both hemispheres." The word "channel" seems duplicated.
2. The authors should add the "age" column in Table S1.
3. In Fig. S5, the number on top of each bar is not located completely above the bar. The authors should correct the figure for readability.
4. Please read the "How to submit" page of Nature Communications website. The following sentence is described there: To cite a preprint, please follow this style: Babichev, S. A., Ries, J. & Lvovsky, A. I. Quantum scissors: teleportation of single-mode optical states by means of a nonlocal single photon. Preprint at <http://arXiv.org/quant-ph/0208066> (2002).

Reviewer #2

(Remarks to the Author)

NCOMMS-24-27413

In this manuscript, the authors use natural language processing models to distinguish between the neural mechanisms involved in speech production versus comprehension, as well as transitions between the two. The manuscript is well-written and includes multiple levels of analysis. However, additional clarification and justification of methods is needed. Ultimately, the manuscript could benefit from several additional analyses to increase its impact to that typical of papers published in Nature Communications.

Major Comments

Introduction:

- Conversation and turn-taking are not synonymous, as extended periods of single speaker speech are also very common in

natural conversation. Including periods of both longer speech production and comprehension could help account for potential edge effects at transition points. It would be nice to analyze whether there are differences in neural activity between longer and shorter periods of production/comprehension, especially since the task-based language you use involve rapid alternation, so it is important to control for this potential confounding factor.

#### Methods:

- The goal of the paper is stated as investigating “how linguistic information is represented in the brain during conversation,” but it is not clear what linguistic information the NLP model outputs represent and therefore what it is that you are correlating neural activity to. Do these NLP embeddings represent a specific element of language or specific type of content, for example?
- Relatedly, the authors mention that a possible reason for the differences found between GPT-2 and BERT models is that BERT contained future word information from the same sentence. If the embeddings mean or represent different things for the two models, it is important to better explain what they each represent (based on how they are derived by the model, for example) and this means that they cannot be used to test generalizability of neural-NLP correlations across models.
- The authors explain the estimation of number of channels, which they reference as their sample size for analysis, but they later also refer to number of words as sample size that could have affected results. Please clarify and include power analysis to justify minimum number of words needed per participant.
- For the control task with passive listening and speaking, it is unclear how the sentence length (number of words per sentence) compares to the average sentence length in natural conversations. If these exchanges are shorter/more abrupt (have more transitions per unit time), that difference could influence results. Please include average sentence length (and SD) for natural conversations.

#### Results:

- For the main analyses of natural conversations, there is a notable difference in average total number of words during production versus comprehension, with almost 30% more words during listening than speaking (“These included  $2728 \pm 1804$  (mean  $\pm$  s.t.d) words during production and  $3460 \pm 2581$  words during comprehension”) which may impact results. Since the main objective is to compare speech production to speech comprehension, it is important for the sample sizes to be balanced between conditions.
- They state that “This model also importantly provided access to the hidden embeddings at scales that are tenable for neuronal analysis.” It wasn’t very clear what this meant. How and why are these used for neural analysis? Please expand on this.
- It was not clear why they assigned white matter locations to their nearest gray matter rather than just excluding white matter contacts since they are less likely to contribute to EEG. This is particularly questionable since they state that they do not significantly change the correlations, but the largest number of channels per brain area is still in cerebral white matter (576 channels in cerebral white matter, compared to the second largest number of channels at 177 in the middle temporal region). I would recommend excluding white matter—otherwise, please clarify why it would not be more appropriate to do so.
- Correlation coefficients of less than 0.2 are typically considered not meaningful even though the correlation may still reach significance due to large sample size. Authors need to point out that the correlation is low and therefore needs to be validated in other ways.
- In paragraph 8 of Results, they mention low-level sound features, but it is not clear what these features are or how they were derived in your case since the NLP models used were only trained on and only ingest text, not audio waveforms. Please clarify.
- In paragraph 10 of Results, they suggest that their experimental design distinguishes between pseudo-interactions and natural conversations, but their results do not provide convincing evidence of how these are different. What is the intrinsic difference between them-- is it the spontaneity and unpredictability? Is it related to a sense of agency? Is it due to invoking of memory in natural conversations but not in forced vocalization/listening? Is it that natural conversations involve more salient personal information and therefore more strongly activate attention? There may be several additional analyses that could begin to tease apart differences in conversation, such as: comparing neural activity during production/comprehension of word vs non-word sounds, or otherwise controlling for various potential confounds such as specific conversation content (autobiographical memories, for example). Being able to more precisely disambiguate the intrinsic differences between natural conversations and pseudo-conversations would be more impactful for the field and would support future experiments using natural language by informing experimenters exactly they are studying with natural speech vs task-based language.
- In paragraph 11 of Results, they mention that the results reflect involvement of a distributed rather than localized network, but their analyses are limited to distinct brain areas. Since language is known to be particularly distributed, functional connectivity analyses would be more suited for answering questions related to differences in the neural process that underlie such broad language processes such as speech production versus comprehension.

#### Discussion:

- The term “linguistic information” is used very loosely/broadly throughout the manuscript, but language is an exceptionally broad construct, so the results are less meaningful without further investigation into which elements of spoken/heard language are at play. Linguistic information, for example, may be related to parts of speech or the specific content of language or features such as linguistic distance, so the manuscript would be strengthened by further investigation into which elements of linguistic information are being related to neural activity here.
- The interpretation and discussion of how this work fills existing gaps in knowledge is weak. Overall, it is not very clear what the value add is or how this work contributes meaningfully to the field and what types of future analyses are now enabled by these findings.

#### Minor comments:

- In the first paragraph of Materials and Methods, “enrolment” should be “enrollment”

- The third sentence in the first paragraph of the discussion is missing a verb. Should be "when the participants were not only engaged in language comprehension..." or something similar.
- Abstract: two occurrences of "through" in the first sentence.

### Reviewer #3

#### (Remarks to the Author)

Large Language Models have made gigantic waves in science, technology, culture, and society, yet they are still largely blackbox devices. The fact that artificial neural networks are extremely simple models of animal neural processing, there is a hypothesis that at least some aspects of how LLMs work reflects neural brain processing, and therefore could be used as a model of brain dynamics, which is far more of a blackbox than LLMs. Therefore, a few labs have been reporting correlation between hidden representations of LLMs and brainwave activity in humans. This paper advances these results in a novel way, and the results are, in my opinion, important enough for publication.

There are several noteworthy results: 1) Correlation between LLM and brain activity; 2) Interesting patterns of brain activity on p.5, some of which is consistent with previous results, but some is novel.

I find the methodology sound, statistical methods appropriate, and the writing clear.

I have a few recommendations:

- 1) After a very general introduction, in paragraphs 1 and 2, the authors hint at what is to be done in the paper. I think the introduction could be focused much more in asking the question of correlation between LLM and brain activity, if I am correct that this is the key result. It would be especially useful to articulate why the paradigm of conversation is innovative by comparison to other similar work in this vain.
- 2) Once the introduction is made much more question-directed, I think the first part of the discussion can similarly be directed at answering the main question.
- 3) More comparison of brain results to other recent reports is necessary.
- 4) GPT-3 and GPT-4 systems have been argued to be far more conversationally able than GPT-2. If it is easy to reevaluate the results on these, the paper will be more current, and the correlations maybe higher.
- 5) I would appreciate more details about the layers than just saying lower and higher layers.
- 6) "Multiple Comparisson" in Figure 1F has an extra "s"
- 7) Manning et al.'s 2020 paper in PNAS would be important to cite, not just the Belinkov papers

Version 1:

Reviewer comments:

### Reviewer #1

#### (Remarks to the Author)

I appreciate the authors' revision. However, I believe that not all of my comments have been fully or appropriately addressed.

The authors mentioned in their rebuttal letter that their multiple tests were independent. However, these tests were NOT independent. For example, as described below, after conducting 17 tests, the authors used the results of these tests to extract representative regions of interest for their subsequent analysis.

-----  
Further, across 17 brain areas with a high number of channels implanted ( $n \geq 25$ , Method), there were 9 areas that showed significantly high percentage of channels ( $p < 0.01$ ) that correlated to NLP embeddings, including areas in the temporal and frontal cortex, thalamus and limbic system structure (Fig. 2a). Thus, in subsequent analyses, we used these 9 areas as representative regions of interest.  
-----

In general, there are cases where multiple comparison correction is unnecessary. However, this is true only when the tests are entirely independent. The methods the authors employed may risk yielding false positives.

Additionally, while the authors created Figure S12, there is no reference to it in the main text.

(Remarks on code availability)

### Reviewer #2

(Remarks to the Author)  
My comments have been addressed.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)  
I am happy with the revisions to my comments as well as the comments by other reviewers. The changes are extensive, and they significantly elevated the manuscript.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)  
I appreciate the authors' revision and believe that my comments have been appropriately addressed.

In the text, the supplementary figures appear in the following order: Fig. S1, Fig. S2, Fig. S12, Fig. S13, ...  
Thus, the supplementary figures should be re-numbered to ensure sequential order.

(Remarks on code availability)

**Open Access** This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

## Summary

We greatly appreciate the reviewer's comments and constructive feedback on our manuscript entitled "**Natural language processing models reveal neural dynamics of human conversation**" and kindly thank them for their generous time and support. Overall, we have taken their comments very seriously and have spent extensive time addressing their suggestions, adding new controls, providing new neuronal recordings and analyses, expanding on our discussion and including many new figures and subplots. Overall, we believe that these changes have helped significantly improve the quality and rigor of the paper.

We also thank the reviewers for their kind comments and suggestions. We were encouraged to hear that "*The manuscript is well-written and includes multiple levels of analysis.*" and "*This paper advances these results in a novel way, and the results are, in my opinion, important enough for publication... I find the methodology sound, statistical methods appropriate, and the writing clear.*".

Finally, with specific respect to the updated manuscript and following the reviewers' recommendations, we have made a number of substantive changes and additions. These include the following:

1. Addition of control experiments in which we included listening and repeating 'jabberwocky sentences' that held no meaning but that still involved sensory perception and motor articulation with four participants and 465 recording channels.
2. To confirm the generalizability of our findings, we added four additional families of state-of-the-art large language models, each of which was independently trained on different datasets with various parameters and number of layers.
3. We included non-parametric rank correlations to relate neural activity with NLP embeddings, as well as additional analysis and updates.
4. We controlled for the length of periods of speech perception and production.
5. We conducted a power analysis of correlation to estimate the number of words given the estimated correlation between neural activity and language embeddings.
6. We included controls showing that the neural correlation is not dependent on age or sex.
7. We broadly extended the supplementary materials to provide further description of our analyses as well as expanded our introduction and discussion in line with suggestions made by the reviewers.

Taken together, we believe the additional data, analysis and reworking of the language of the manuscript significantly increases the robustness of our work and strengthens the conclusions of the research. Below are the point-by-point responses to the reviewers' comments with reviewer comments in *grey italics* and our responses in black typeface. All changes in the manuscript itself and supplementary materials are **highlighted in blue**.

Reviewer #1 (Remarks to the Author):

Comments:

This study reports that natural language processing models revealed neural dynamics of human conversation. While the experimental procedures of the study appear to be well thought out and executed, some flaws exist in the processing and interpretation of the data which need to be addressed. These concerns and others are explained below:

Major points:

1. The authors did not correct for multiple comparisons across multiple brain regions throughout the manuscript (e.g., Fig. 2A). They should address this point.

We appreciate your concern regarding multiple comparisons across brain regions. While it is true that we examined correlations between each individual channel and NLP embeddings, we emphasize that these analyses were conducted *independently* for each channel. The subsequent grouping by region was performed to examine how the significant channels are distributed across various brain areas. Therefore, a traditional multiple comparison correction, typically applied when asking whether there exist any brain areas among all areas that significantly respond to a task, is not applicable in this context. Instead, our analysis focuses on the individual relationship between each brain channel and the NLP embeddings, shown in Figure 2b and Figure 3b.

Nevertheless, we appreciate your comments and acknowledge the importance of addressing potential concerns about multiple comparisons. To further clarify our approach and strengthen the interpretation of our findings, we performed a new analysis which included selecting channels that are significantly correlated to NLP embeddings with Bonferroni correction of the NLP nodes, layers and frequency bands. Then we grouped these channels by brain area (17 areas). Using a chi-square proportion test, again with Bonferroni correction for the areas, we examined whether there exist any areas that show significantly higher percentage of correlated channels. Again, there were – with similar ratios in different brain areas as found using our other methods. This reworked analysis provides additional support for the robustness of our results and addresses potential concerns regarding multiple comparisons. These results are shown below in figure R1 which is added to the supplementary materials as S12.

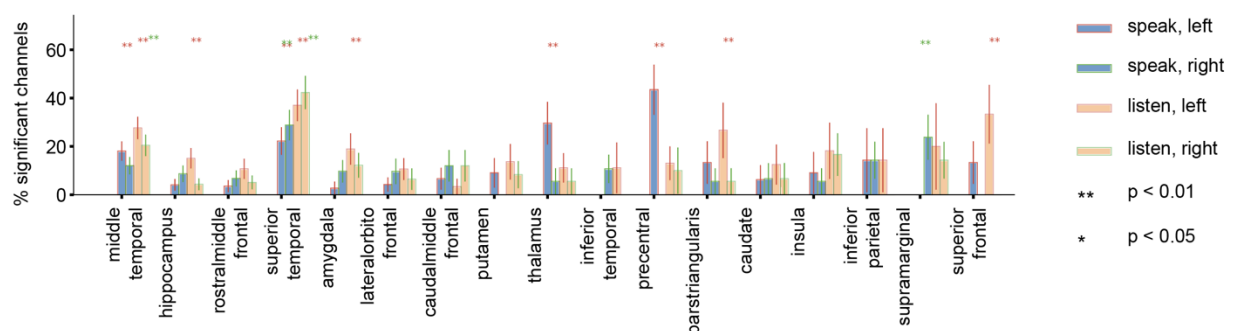


**Figure R1(S12). Percentage of significant correlated channels with Bonferroni correction to brain areas.** To ensure the rigorousness of our findings, we performed a control analysis with Bonferroni correction to brain areas. Specifically, we firstly examined whether an individual channel significantly correlated to NLP embeddings, and performed Bonferroni correction of the NLP nodes, layers and frequency bands. Then we grouped these channels by brain areas (17 areas), and used chi-square proportion test with Bonferroni correction to study whether there exist

any areas that show significantly higher percentage of correlated channels. These results were compared to **Figure 2a**, where we performed Bonferroni correction for the NLP nodes and layers only, and averaged channels over frequency bands.

2. The authors did not describe the limitations of this study. They should describe at least two limitations to increase the reliability of the study.

Thank you for your comment. We now further expand on two main limitations in the revised manuscript:

First, as pointed out by Reviewer 2, our current work does not aim to identify the specific linguistic, audio, or cognitive factors that contribute to the neural response under the natural conversation. Rather, our study investigates the compositional representations of language information conveyed during natural conversation, and demonstrates the differences of neural supports between natural conversation and passive language activities. Natural conversations are complicated, involving spontaneity, unpredictability, and the interplay of various cognitive processes. To rigorously disentangle the contributions of these factors, it is necessary to design experiments with extensive control conditions and analysis to eliminate irrelevant factors, which is the work we are already planning.

Second, the variability in electrode locations across participants is a recognized challenge in intracranial EEG studies. This heterogeneity makes direct comparisons between individuals difficult and limits our ability to precisely identify the variance across individuals. This limitation is compounded by the necessarily sparse sampling inherent in intracranial investigations. To address this issue, we employed area-level analyses and focused on identifying patterns of neural activity with combination of participants. Future studies involving other recording methods could address this limitation as will the gradual buildup of larger datasets.

Last, we focused on text-based NLP models in this work. These models used transcribed words as input. Therefore, the embeddings lack the information of the auditory and phonetic features of the conversation. Further studies will include NLP models that use audio files as inputs to capture the full richness of speeches.

We edited the **Discussion** accordingly by adding the following sentences:

However, there are limitations of this study: Our primary focus is on demonstrating the presence of differences between natural conversation compared to passive listening and repeating rather than delving into their underlying mechanisms. Due to the inherent complexity of natural conversation, future studies will be needed to pinpoint the precise linguistic or cognitive factors differentiating it from passive listening and repeating. Further, variations in electrode placement across participants hinder direct comparisons of neural activity, making it challenging to precisely quantify individual differences in brain responses. Lastly, the reliance on text-based NLP models limits its ability to capture the full richness of spoken language, as it excludes acoustic information.

3. The authors should examine whether the neural response is correlated not only to the IQ score (Fig. S4) but also to age and sex.

Thank you for the suggestion. In Figure R2, we showed the percentage of responding channels as a function of participants' age and sex. The percentage of channels were not correlated to the age of the participants (Kendall Tau test, statistics = 0.26,  $p = 0.21$  for speak; statistics = 0.011,  $p = 0.96$  for listen). Similarly, we examined whether the sex of the participants plays a role in determine the percentage of correlated channels. Again, we found that the percentage was not significantly modulated by the sex of the participants (Oneway ANOVA, statistics = 0.23,  $p = 0.64$  for speak; statistics = 0.003,  $p = 0.96$  for listen). Therefore, neural responses appeared to be largely unrelated to the age and sex of the participants. These two subplots were added to Supplementary Figure S4.

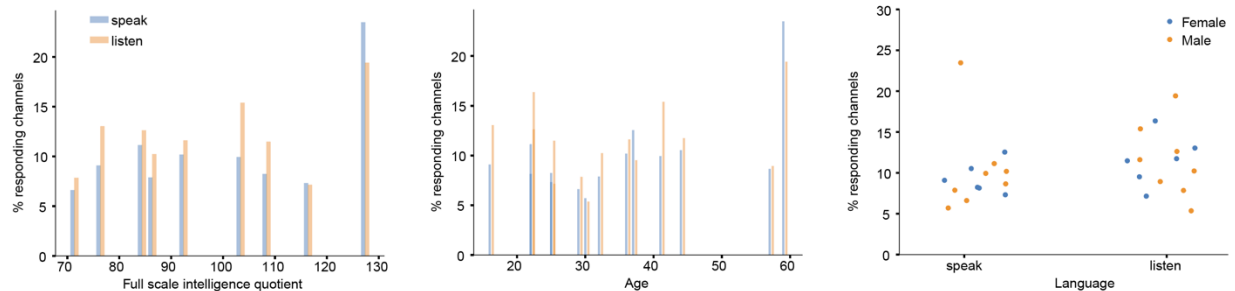


**Figure R2(S4). Responding channel by participant's intelligence quotient (IQ), age and sex.** To examine whether the neural response is correlated to the IQ score for the participants, we performed full scale intelligence quotient test for 9 out of 14 participants, and calculated the percentage of responding channels as a function of their IQ (*left*). Each pair of bars represent a participant, and the location of these bars are plotted based on their IQ values. There is no significant correlation between the percentage of channels and participants IQ (Kendall Tau test, statistics = 0.167,  $p = 0.61$  for speak; statistics = 0.11,  $p = 0.76$  for listen). *Similar, the percentage of channels is not correlated to the age of the participants (middle, Kendall Tau test, statistics = 0.26,  $p = 0.21$  for speak; statistics = 0.01,  $p = 0.96$  for listen).* Last, the percentage channels are not dependent on the sex of the participants (*right, One-way ANOVA, statistics = 0.23,  $p = 0.64$  for speak; statistics = 0.003,  $p = 0.96$  for listen).*

Minor points:

1. In the legend of Fig. 1, the authors describe as follows: "A total of 1910 bipolar-referenced channel channels that demonstrated a low rate of epileptiform discharges were primarily distributed in the frontal, temporal and mesial areas in both hemispheres." The word "channel" seems duplicated.
2. The authors should add the "age" column in Table S1.
3. In Fig. S5, the number on top of each bar is not located completely above the bar. The authors should correct the figure for readability.
4. Please read the "How to submit" page of Nature Communications website. The following sentence is described there: To cite a preprint, please follow this style: Babichev, S. A., Ries, J. & Lvovsky, A. I. Quantum scissors: teleportation of single-mode optical states by means of a nonlocal single photon. Preprint at <http://arXiv.org/quant-ph/0208066> (2002).

Thank you so much for highlighting these points. We have changed the text in the manuscript to correct them all.

Reviewer #2 (Remarks to the Author):

NCOMMS-24-27413

In this manuscript, the authors use natural language processing models to distinguish between the neural mechanisms involved in speech production versus comprehension, as well as transitions between the two. The manuscript is well-written and includes multiple levels of analysis. However, additional clarification and justification of methods is needed. Ultimately, the manuscript could benefit from several additional analyses to increase its impact to that typical of papers published in Nature Communications.

## Major Comments

### Introduction:

- Conversation and turn-taking are not synonymous, as extended periods of single speaker speech are also very common in natural conversation. Including periods of both longer speech production and comprehension could help account for potential edge effects at transition points. It would be nice to analyze whether there are differences in neural activity between longer and shorter periods of production/comprehension, especially since the task-based language you use involve rapid alternation, so it is important to control for this potential confounding factor.

You raise an important point regarding the distinction between conversation and turn-taking. While our study primarily focused on the rapid alternation of speech production and comprehension, we are aware that natural conversation often involves extended periods of single-speaker speech, and the length of the speech may impact the channel-NLP correlation.

To address this, we performed additional analysis by grouping the periods of single speaker speech into short (word number < 50 percentile) and long (word number > 50 percentile) for each participant, and then calculated the percentage of channels that significantly correlated to NLP embeddings respectively in each group, following the same procedure as described in the manuscript.

Using this approach, we found that the short period resulted in slightly higher percentage of correlated channels for both language production and perception (8.7% for short period versus 6.9% for long period during speaking; 12.4% versus 7.3% for listen). However, all of them are significantly higher than chance (Chi-square proportion test:  $p < 2 \times 10^{-4}$  for all the short and long periods during speaking and listening). In addition, the channels that respond during short and long periods are significantly overlapping (T-test,  $p < 2 \times 10^{-14}$  for both speech production and perception). Therefore, while there were minor differences in the overall proportion of responsive channels, the underlying neural mechanisms appear to be largely shared across speech segment durations.

We added this control in the **Methods**:

In addition, to examine whether the neural-NLP correlation is modulated by only short or long durations of speech before transitions, we categorized word sequences into short or long durations using 50 percentile threshold. Even after this separation, we found a significantly higher percentage of responding channels for both groups than chance (T-test,  $p < 2 \times 10^{-14}$  for

both speaking and listening), suggesting that the correlation is not caused by either short or long durations.

We also added this control in the **Results**:

In addition, we showed that the neural-NLP correlation is robust regardless the duration of the speech between transitions, or using non-parametric rank correlations (Materials and Methods).

Methods:

- The goal of the paper is stated as investigating “how linguistic information is represented in the brain during conversation,” but it is not clear what linguistic information the NLP model outputs represent and therefore what it is that you are correlating neural activity to. Do these NLP embeddings represent a specific element of language or specific type of content, for example?

Thank you for raising this question which we have now clarified in the updated manuscript. In NLP models, the processing of language occurs in a hierarchical manner. However, unlike rule-based systems, these models do not have explicit, human-readable rules for each linguistic level. Instead, they learn statistical patterns from vast amounts of text data, which allows them to approximate these linguistic processes. As described in more detail before (Vig, J. and Belinkov, Y., 2019 arXiv:1906.04284; Liu, N.F., *et al.*, 2019. arXiv:1903.08855), NLP model embeddings do not necessarily represent a single, specific element of language or content, and the layers do not correspond directly to known linguistic categories. Instead, they typically encode a complex, multidimensional representation of linguistic information. The embeddings of these NLP models often use distributed representations, where linguistic information is spread across many dimensions rather than localized to specific units. Through this, these NLP embeddings therefore provide a comprehensive representation of linguistic information rather than specific elements.

That being said, previous analyses have used probe analysis to identify key linguistic features processed in these models. They found that different layers of LLMs tend to specialize in distinct aspects of language processing. Lower (input) layers, for example, often capture more surface-level features such as word morphology and the word-level features, while middle and higher layers show a strong affinity for semantic information, phrase and sentence level, contextual representations, as well as abstract and higher-level linguistic concepts.

In our work, we correlate neural activities to embeddings from different layers. The percentage of correlated channels will indicate the language processing for that layer. Then we specifically examined the correlations with the layer categories of low and middle-high layers. We found that the neural dynamics of most areas preferably correlated with representations in the middle and higher layers of the NLP model, suggesting that these activity patterns reflected contextual or sentence-level information integration over the course of conversation rather than simply individual words which would be represented by lower layers. Notably, these layer distributions were significantly altered when the participants were passively listening and repeating sentences, confirming that these neural signals reflected processes that were specific to natural conversation rather than simply the processing of any linguistic signal.

Taken together, this approach therefore allowed us to provide a comprehensive representation of linguistic information communicated during conversation. More importantly, by performing control comparisons (e.g., natural conversation vs. passive listening or sentence repetition), we could quantify the degree to which certain areas selectively processed information related to conversation. Such analyses would not be possible by selecting specific elements of language.

These explanations have now been updated in **Results** under the section ‘Neural-NLP relations across frequency bands and NLP layers’:

In NLP, the processing of language occurs in hierarchical layers, which extract the statistical patterns from vast amounts of text data. Although layers do not correspond directly to known linguistic categories, lower (input) network layers preferentially reflect information about individual words independent of their context, whereas middle layers capture the more complex linguistic structures, and higher (output) layers reflect integrated compositional sentence-level information.

We also added the description of NLP models in the **Methods**:

NLP models process language hierarchically, but without explicit, human-readable rules for each linguistic level. Instead, they learn patterns from large text datasets, approximating linguistic processes statistically. Their embeddings do not necessarily represent single language elements or correspond directly to linguistic categories. Rather, they encode complex, multidimensional linguistic information distributed across many dimensions, providing a comprehensive representation of language.

- Relatedly, the authors mention that a possible reason for the differences found between GPT-2 and BERT models is that BERT contained future word information from the same sentence. If the embeddings mean or represent different things for the two models, it is important to better explain what they each represent (based on how they are derived by the model, for example) and this means that they cannot be used to test generalizability of neural-NLP correlations across models.

Thank you for your comment. This is an interesting point about the potential differences between GPT-2 and BERT embeddings. While it is true that these models have distinct architectures and training objectives, it is helpful to note that both models have demonstrated a capacity to capture meaningful semantic and syntactic information from language.

GPT-2 is a generative model trained on a massive text dataset to predict the next word in a sequence. This unidirectional approach allows it to capture contextual information from preceding words. In contrast, BERT is a bidirectional model trained on a masked language modeling task, enabling it to consider both preceding and succeeding context. This difference in architecture can lead to variations in how these models represent language, particularly in terms of how they encode future word information.

Despite these architectural differences, our findings indicate that both GPT-2 and BERT embeddings exhibit significant correlations with neural activity. This suggests that these models, while capturing language in distinct ways, are able to extract features that are relevant to human

brain processing. Indeed, the presence of such correlations across fundamentally different model architectures strengthens our findings for the *generalizability* of neural-NLP correlations.

In addition to these two NLP models, we added more recent models that contain various number of parameters and trained independently. They majorly composed 4 families of pretrained models, including Bloom (with 560 million and 1 billion parameters), Falcon (1b and 40b), dolly (3b and 12b), and llama (3b, 7b, 13b, wizard-13b, 30b, chronos-33b). These models contain different numbers of layers and number of hidden embeddings per layer. To account for this variation, we took the first 100 principal components and then correlated them with the neural activity. We found that there were significant higher percentage of channels correlated to all models examined (Chi square proportion test,  $p < 1 \times 10^{-23}$ ) during both speaking and listening. Further, the percentage of correlated channels showed an increase trend with the size of NLP models ( $R > 0.80$  and  $p < 2 \times 10^{-3}$ ).

This new analysis further extends NLP models with diverse sizes and model families, with all of them show significant higher percentage of correlated channels than chance, indicating the generalizability across various of models. In other words, these findings appear to be robust and generalizable property of neural activity.

We added the Figure below into the Supplementary Figure S13:

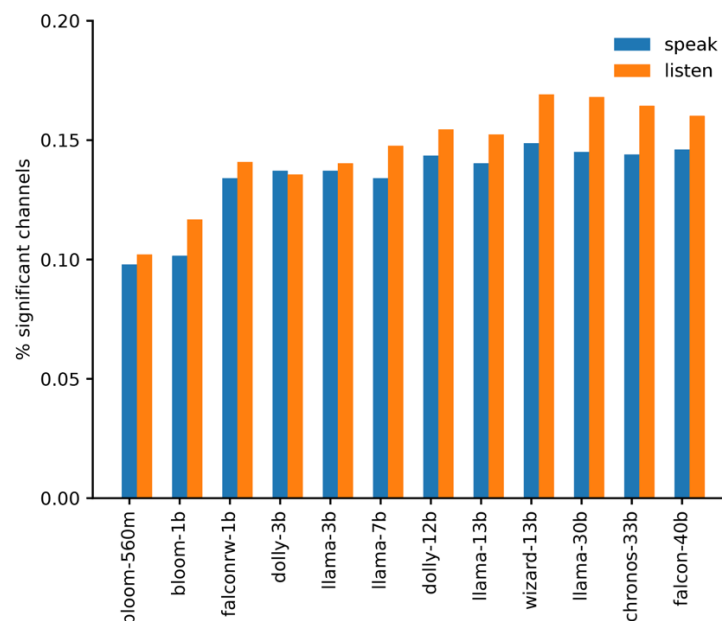


**Figure R3 (S13). Percentage of significant channels examined by various NLP models.** By applying NLP models with various size, number of layers, and families, we found that there were significant percentage of channels correlating to all models for both speech production and comprehension (Chi square proportion test,  $p < 1 \times 10^{-23}$ ), with larger models showing higher percentage of correlated channels ( $R > 0.80$  and  $p < 2 \times 10^{-3}$ ).

We edited the **Results** section:

Thirdly, to confirm that the above findings were generalizable, we compared the neural activity patterns to those of another NLP model trained on a different dataset. Here, we used a BERT (base) model with a bidirectional network architecture and trained on different linguistic materials (46). Again, we found that the proportion of correlated channels was significantly higher than chance (Chi-square test, statistic = 15278,  $p < 10^{-100}$ ), and was greater than or equal to that observed with the GPT-2 model (Fig. S7b). Further, similar findings were observed across a series of NLP models with different layers and architectures (Fig. S13, Chi-square proportion test,  $p < 1 \times 10^{-23}$ ), with an increase of the percentage of correlated channels to the size of NLP models ( $R > 0.80$  and  $p < 2 \times 10^{-3}$ ). These suggest that the relationships between neural activity patterns in the brain and those of the language models are generalizable properties of neuronal responses.

We added the figure in the Supplementary Material, and a paragraph in **Methods** to describe these models:

To confirm the generalizability of our results, in addition to these two NLP models, we examined other models that contain various number of parameters and are trained independently. They majorly composed 4 families of pretrained models, including Bloom (with 560 million and 1 billion parameters), Falcon (1b and 40b), dolly (3b and 12b), and llama (3b, 7b, 13b, wizard-13b, 30b, chronos-33b). These models contain different number of layers and number of hidden embeddings per layer. To account this variation, we took the first 100 principal components and then correlated them with the neural activity. We found that there were significant higher percentage of channels correlated to all models examined (Figure S13, Chi square proportion test,  $p < 1 \times 10^{-23}$ ) during both speaking and listening. Further, the percentage of correlated channels showed an increase trend with the size of NLP models ( $R > 0.80$  and  $p < 2 \times 10^{-3}$ ). By extending NLP models with much diverse sizes and model families, we showed that all of them were significant higher percentage of correlated channels than chance, indicating the generalizability across various of models. In other words, these findings appear to be robust and generalizable property of neural activity.

- The authors explain the estimation of number of channels, which they reference as their sample size for analysis, but they later also refer to number of words as sample size that could have affected results. Please clarify and include power analysis to justify minimum number of words needed per participant.

For these natural conversations, there was only so much time that the patient could spend in the conversation before either getting tired or being interrupted for any number of reasons (clinical activities, visitors, food, etc). Consequently, we did not design the analysis for a specified amount of data but sought statistically significant findings in the data we obtained.

Further, we recorded 1000s of words for speaking and listening (for example, one participant engaged in a conversation for 19.4 minutes, resulting in 1697 words perceived and 1787 words produced). This is quite similar to the length of time used in many block design experiments. But in our natural conversations, there were no artificial intermissions including pauses for breaks, instructions, etc.

Nonetheless, we recognize the potential concern and have therefore conducted a power analysis of correlation to estimate the minimum number of words required per participant to detect a correlation between neural activity and language embeddings. Used Fisher's z-transformation, and assuming a moderate sample correlation ( $R = 0.1$ ), an alpha level of 0.05, and a power of 0.8, our calculations indicate that approximately 783 words spoken or perceived per participant would be sufficient. As our dataset contains larger number of words per participant, we are confident that our sample size is adequate to detect meaningful effects.

Finally, we showed that the number of words that a participant spoke and listened to was not correlated with the percentage of responding channels in **Figure S5**. There is no correlation with the percentage of channels to the rank of words by participants ( $p = 0.95$ ). This suggests that the sample size is not a factor contributing to our results, hence having a longer conversation is not expected to have an impact on the percentage of the responding channels.

We added the following paragraph in **Methods**:

Further, we conducted a power analysis of correlation to estimate the minimum number of words required per participant to detect a significant correlation between neural activity and language embeddings. We used Fisher's z-transformation for the correlation coefficient test:

$$N = \left( \frac{z_\alpha + z_\beta}{C(r)} \right)^2 + 3$$

where

$$C(r) = \frac{1}{2} \log \left( \frac{1+r}{1-r} \right)$$

Assuming a moderate sample correlation ( $R = 0.1$ ), an alpha level of 0.05, and a power of 0.8, our calculations indicate that approximately 783 words spoken or perceived per participant would be sufficient. As our dataset contains larger number of words per participant ( $2728 \pm 1804$  (mean  $\pm$  s.t.d) words during production and  $3460 \pm 2581$  words during comprehension), our sample size was adequately powered to detect meaningful effects.

- For the control task with passive listening and speaking, it is unclear how the sentence length (number of words per sentence) compares to the average sentence length in natural conversations. If these exchanges are shorter/more abrupt (have more transitions per unit time), that difference could influence results. Please include average sentence length (and SD) for natural conversations.

You raise a valid point about the potential influence of sentence length on our findings. That said, unlike structured language tasks, natural conversations may not have clear-cut sentence boundaries. There is occasionally overlapping speech, interruptions, co-constructions, and frequent turn-taking, making it challenging to define precise sentence lengths. The following table provides a representative example of conversation between the experimenter and the participant, showing that the alternation between speakers may make it unclear to define a sentence.

| Speaker      | Transcription                                                   |
|--------------|-----------------------------------------------------------------|
| Participant  | <i>A cat name Alex.</i>                                         |
| Experimenter | <i>And he's just a cat. He's just a cat. He's not the best.</i> |
| Participant  | <i>He is</i>                                                    |
| Experimenter | <i>cat</i>                                                      |
| Participant  | <i>Fantastic</i>                                                |
| Participant  | <i>He's like an indoor outdoor cat.</i>                         |

On the other side of the spectrum, some participants might connect language units with words like ‘and’, creating exceptionally long utterances. For example, one participant said:

| Speaker     | Transcription                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|-------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Participant | <i>So I had to make a whole other VLAN in our network and then try to figure out what the best way to have it be to get the traffic from the call center, uh, desktops over to that new VLAN and then tie that new VLAN into our, um, MPLS shared network between us and our hosting provider, so we have this little semi-private network, and so it's basically figuring out how to get the data from here to there and back again in the quickest time possible.</i> |

Therefore, a precise sentence boundary in spontaneous conversation can be challenging due to its dynamic nature.

Nonetheless, given your comments, we addressed this by analyzing speech segments separated by the speaker transitions. In other words, instead of manually labeling sentences from the conversation material, we identify the transitions between speakers and use the words in between the transitions, and examine whether channels responding to only short or long durations. Across all participants, we identified an average of  $12.3 \pm 3.9$  (mean  $\pm$  s.t.d) words between transitions (ranging from 6.2 to 18.9). Then, as described in our response to the first comment of Reviewer 2, we showed that short durations between transitions resulted in slightly higher percentage of correlated channels for both language production and perception, but all of them are significantly higher than chance and there is significant overlap between short and long durations, indicating a largely shared neural mechanisms across speech segment durations.

Together, our conclusions are generalizable between conversation materials from short and long durations, even if the specific length of sentences may not be well-defined.

We added the results into the Supplementary Materials:

In addition, to examine whether the neural-NLP correlation is modulated by only short or long durations of speech before transitions, we categorized word sequences into short or long durations using 50 percentile threshold. Even after this separation, we found a significantly higher percentage of responding channels for both groups than chance (T-test,  $p < 2 \times 10^{-14}$  for both speaking and listening), suggesting that the correlation is not caused by either short or long durations.

We also added this control in the **Results**:

In addition, we showed that the neural-NLP correlation is robust regardless the duration of the speech between transitions, or using non-parametric rank correlations (Materials and Methods).

Results:

- For the main analyses of natural conversations, there is a notable difference in average total number of words during production versus comprehension, with almost 30% more words during listening than speaking (“These included  $2728 \pm 1804$  (mean  $\pm$  s.t.d) words during production and  $3460 \pm 2581$  words during comprehension”) which may impact results. Since the main objective is to compare speech production to speech comprehension, it is important for the sample sizes to be balanced between conditions.

Unequal word counts between speech production and comprehension could have an impact on our findings. To address this, we conducted a new control analysis by matching the number of words between the two conditions for each participant. Specifically, we set a cutoff word number, which was chosen to be the minimum of the number of words during speak and listen. This was performed for each participant respectively.

We found that there is a slight numerical difference in the percentage of significant channels between speech production (10%) and comprehension (13%) when word counts are matched, with significantly more channels responding during comprehension (chi-square proportion test, statistics = 14.1,  $p = 1.7 \times 10^{-4}$ ). Further, both conditions exhibit a substantially higher percentage of significant channels compared to chance (chi-square proportion test, statistics  $> 100$ ,  $p < 2 \times 10^{-23}$  for both conditions). These findings are largely consistent with our main analysis without word number matching (compared to 11% for production and 15% comprehension), indicating that the variation in word number does not drive the apparent difference between channels responding to speaking and listening. Further, these findings also suggest that the presence of neural activity correlating to NLP embedding is conserved in both speech production and comprehension.

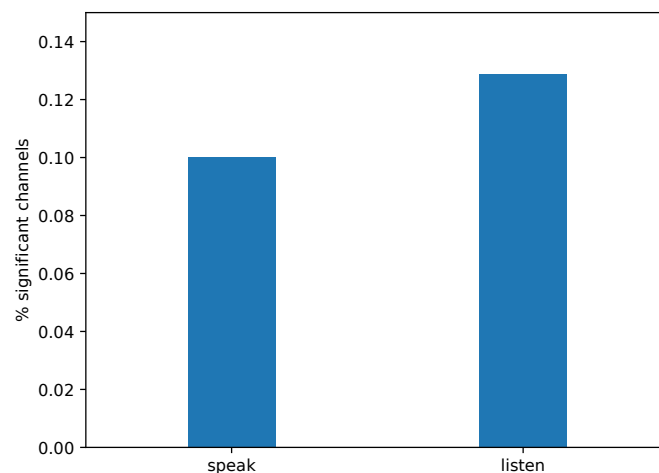


**Figure R4. Control analysis with matching word number for speech production and comprehension.** With same number of words for speech production and comprehension, we still found more channels responding to speech comprehension (chi-square proportion test, statistics = 14.1,  $p = 1.7 \times 10^{-4}$ ). However, both production and comprehension involved significant channels higher than chance (chi-square proportion test, statistics > 100,  $p < 2 \times 10^{-23}$  for both conditions).

We included this analysis in **Methods**:

To address potential biases due to unequal word counts between speech production and comprehension, we matched the number of words in each condition for each participant by setting a cutoff equal to the minimum word count for speaking and listening. We found that there was a slight numerical difference in the percentage of significant channels between speech production (10%) and comprehension (13%) when word counts were matched, with conditions a substantially higher percentage of significant channels compared to chance (chi-square proportion test, statistics > 100,  $p < 2 \times 10^{-23}$  for both conditions). These findings are largely consistent with our main analysis without word number matching (compared to 11% for production and 15% comprehension), indicating that the variation in word number does not drive the apparent difference between channels responding to speaking and listening.

- They state that “This model also importantly provided access to the hidden embeddings at scales that are tenable for neuronal analysis.” It wasn’t very clear what this meant. How and why are these used for neural analysis? Please expand on this.

Here is what we meant by this sentence: This NLP model provided access to hidden embeddings in a vectorized format. This vectorized version can then be directly compared to the neural data. That is, we have two sets of time series data that can be directly and efficiently compared. In addition, given the substantial volume of recorded words (thousands), the GPT-2 small model is sufficient for our analysis, offering a tenable computational efficiency without compromising performance.

We updated the sentence in our manuscript in **Results**:

Importantly, this model provided access to hidden embeddings in a vectorized format. Given the substantial volume of recorded words (thousands), the GPT-2 small model is used in our analysis.

In addition, we also edited the **Introduction** to make it clear how we applied the NLP models to study neural activity underlying natural conversation:

Here, we utilize these models as artificial, hierarchically structured vectorized representations of language during natural dialogue. This approach allows us to investigate the neural basis by which the brain processes entire word sequences within the context of conversation as one process, rather than breaking it into small pieces of components. Further, by examining the correlation between neural channel activity and NLP embeddings, we aim to identify brain regions specifically involved in encoding language-related information. This method enabled us to explore how specific sequences of words, together with their compositional semantic and contextual features, are represented in the brain during both speaking and listening, despite differences in language content. Finally, this approach allowed us to compare neural patterns that

respond selectively during speaker transitions with those that process word sequences. Together, our approach offers a comprehensive investigation into the neural mechanisms underlying natural conversation by directly examining the integrated processes of speaking, listening, and speaker transitions. This method provides a holistic view of the neural substrates involved in these interconnected aspects of language communication.

- It was not clear why they assigned white matter locations to their nearest gray matter rather than just excluding white matter contacts since they are less likely to contribute to EEG. This is particularly questionable since they state that they do not significantly change the correlations, but the largest number of channels per brain area is still in cerebral white matter (576 channels in cerebral white matter, compared to the second largest number of channels at 177 in the middle temporal region). I would recommend excluding white matter—otherwise, please clarify why it would not be more appropriate to do so.

We appreciate your consideration and have now also performed your suggested analysis. Firstly, to clarify, white matter contacts were not assigned the nearest gray matter in our main analysis. Instead, they were grouped in the “other” categories (Figure 2a, 2b, 3a), and analyzed separately from the major gray area channels. We have now also clarified this point in the revised the legend of Figure S6: [In the main analysis, brain areas identified as white matter were categorized as ‘Other’ and included in all the analysis.](#)

Furthermore, we performed a control analysis in Figure S6 to show that the percentage of channels responding to NLP embeddings remains largely the same whether we categorize white matter as ‘other’ category or combine them with gray matter. In other words, the control to assign white matter channels to the nearest gray matter is to further confirm that our conclusions were not impacted by the specific treatment of white matters.

Thirdly, given your recommendation, we have performed an additional analysis by excluding white matter contacts altogether. We found that on average, there were 11.5% channels responding to NLP embeddings during language production and 15.2% channels during comprehension, as compared to 11.0% and 15.0% if white matter channels were included. Both the language production and comprehension without white matter were significantly higher than random fluctuations (Chi-square proportion test: statistics = 37.7,  $p = 8.2 \times 10^{-10}$  for speaking; statistics = 127.4,  $p = 1.5 \times 10^{-29}$  for listening). Therefore, the neural-NLP correlations were present regardless of whether white matter channels were included.

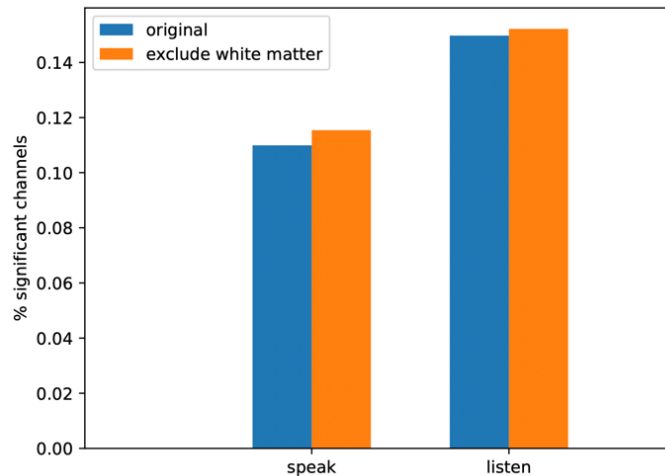


**Figure R5. Control analysis to exclude channels from white matter areas.** By excluding white matter channels, we found that on average, there were 11.5% channels responding to NLP embeddings during language production and 15.2% channels during comprehension. Both were significantly higher than random fluctuations (Chi-square proportion test: statistics = 37.7,  $p = 8.2 \times 10^{-10}$  for speaking; statistics = 127.4,  $p = 1.5 \times 10^{-29}$  for listening). These percentages were very similar to those when white matters were included (11.0% and 15.0% for production and comprehension). Therefore, the neural-NLP correlations were present regardless of whether white matter channels were included.

We added this analysis in **Results**:

The results were also robust if assigned white matter locations to their nearest gray matter (**Fig. S6**), or if these channels were excluded entirely (Chi-square proportion test: statistics = 37.7,  $p = 8.2 \times 10^{-10}$  for speaking; statistics = 127.4,  $p = 1.5 \times 10^{-29}$  for listening).

- Correlation coefficients of less than 0.2 are typically considered not meaningful even though the correlation may still reach significance due to large sample size. Authors need to point out that the correlation is low and therefore needs to be validated in other ways.

While it is true that correlation coefficients below 0.2 are often considered weak, they are robustly observed across various brain areas and are not driven by only one participant, suggesting a meaningful relationship between the neural activity and NLP models.

Firstly, to address your comments more directly, we conducted additional control analysis by performing non-parametric hypothesis tests for rank correlations. Specifically, we used Kendall Tau test and Spearman R to examine the correlations between channels and NLP embeddings. Using Kendall Tau test, we found there were 11.3% channels responding to NLP during language production, whereas there were 12.6% to comprehension (Chi-square proportion test, statistics = 150.0,  $p = 1.1 \times 10^{-36}$  for speaking; statistics = 230.1,  $p = 5.5 \times 10^{-52}$  for listen). Similar percentage of channels were selected using the Spearman R test (speaking: 11.8%, statistics = 187.7,  $p = 1.0 \times 10^{-42}$ ; listening: 12.9%, statistics = 253.0,  $p = 5.8 \times 10^{-57}$ ). These together, indicate that the channel-NLP correlations were not limited to specific statistical methods, and were generalizable to various methods for correlation tests.

Secondly, there are multiple factors that contribute to the neural-NLP correlations. These are majorly due to the complexity of the human brain and the nature of natural language processing. In fact, 0.2 is the estimated ceiling level in one of the datasets in the fMRI work for language comprehension: Schrimpf, M., *et al.*, 2021. PNAS, 118(45), p.e2105646118. Further, there are inherent differences between human brain networks and artificial language models, including the structure, the processing mechanism, the inputs, etc. Lastly, the location of electrodes was not determined based on any of our tasks, and was not optimized for areas that respond to conversations. All these factors contribute to the overall correlation coefficients.

Therefore, it is true that the absolute correlation coefficients indicate they are weak correlations, potentially due to various factors explained above, but more importantly, they are statistically robust and generalizable. Crucially, we have used additional analysis with different measures of the relationship between the model and the neural activity and get essentially the same results. This indicates that it is neither the specific statistical test nor method of correlation analysis which drives the findings.

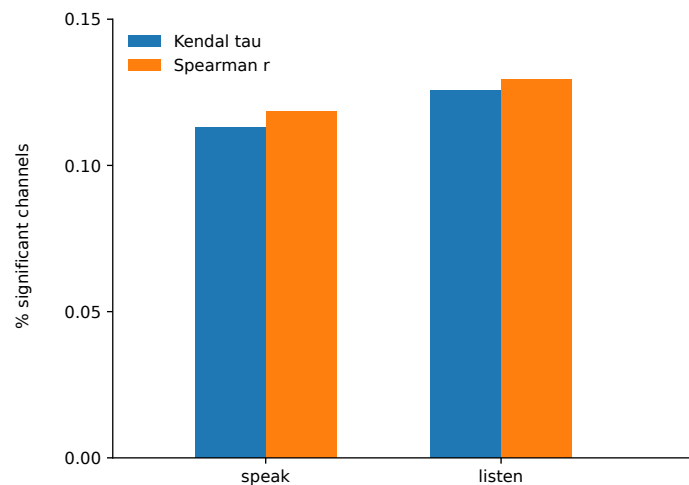


**Figure R6. Control analysis with non-parametric rank correlations.** We used Kendall Tau test and Spearman R to examine the correlations between channels and NLP embeddings. Using Kendall Tau test, we found there were 11.3% channels responding to NLP during language production, whereas there were 12.6% to comprehension (Chi-square proportion test, statistics = 150.0,  $p = 1.1 \times 10^{-36}$  for speaking; statistics = 230.1,  $p = 5.5 \times 10^{-52}$  for listen). Similar percentage of channels were selected using the Spearman R test (speaking: 11.8%, statistics = 187.7,  $p = 1.0 \times 10^{-42}$ ; listening: 12.9%, statistics = 253.0,  $p = 5.8 \times 10^{-57}$ ). These together, indicate that the channel-NLP correlations were not limited to specific statistical methods, and were generalizable to various methods for correlation tests.

We added the results in **Methods**:

Finally, to ensure the observed percentage of correlated channels were not limited only to Pearson correlations, we conducted additional control analysis by performing non-parametric hypothesis tests for rank correlations. Specifically, we used Kendall Tau test and Spearman R to examine the correlations between channels and NLP embeddings. Using Kendall Tau test, we found there

were 11.3% channels responding to NLP during language production, whereas there were 12.6% to comprehension (Chi-square proportion test, statistics = 150.0,  $p = 1.1 \times 10^{-36}$  for speaking; statistics = 230.1,  $p = 5.5 \times 10^{-52}$  for listen). Similar percentage of channels were selected using the Spearman R test (speaking: 11.8%, statistics = 187.7,  $p = 1.0 \times 10^{-42}$ ; listening: 12.9%, statistics = 253.0,  $p = 5.8 \times 10^{-57}$ ). These together, indicate that the channel-NLP correlations were not limited to specific statistical methods, and were generalizable to various methods for correlation tests.

We also added this result in **Results**:

In addition, we showed that the neural-NLP correlation is robust regardless the duration of the speech between transitions, or using non-parametric rank correlations (Materials and Methods). These together, indicate the robustness of neural-NLP relations during language processes of natural conversation.

- In paragraph 8 of Results, they mention low-level sound features, but it is not clear what these features are or how they were derived in your case since the NLP models used were only trained on and only ingest text, not audio waveforms. Please clarify.

You are right that our NLP models were exclusively trained on text data and do not incorporate audio features. However, this does not exclude the possibility that low-level acoustic properties could indirectly influence our findings. For example, if neural responses were primarily driven by word pitch, and there was a correlation between word pitch and word meaning, this could inadvertently lead to a correlation between NLP embeddings and neural activity. Therefore, to address this, we conducted analyses examining the relationship between low-level acoustic features, such as the amplitude and pitch, and neural responses. We showed the NLP-correlated channels responded similarly to amplitude and pitch as compared to all recorded channels (2-side permutation test combined, amplitude:  $p = 0.41$  for speak,  $p = 0.21$  for listen; pitch:  $p = 0.79$  for speak,  $p = 0.60$  for listen;  $n = 10,000$ ). In this way, we were able to demonstrate that the observed correlations primarily driven by the semantic and syntactic content of the speech, rather than low-level acoustic properties.

- In paragraph 10 of Results, they suggest that their experimental design distinguishes between pseudo-interactions and natural conversations, but their results do not provide convincing evidence of how these are different. What is the intrinsic difference between them-- is it the spontaneity and unpredictability? Is it related to a sense of agency? Is it due to invoking of memory in natural conversations but not in forced vocalization/listening? Is it that natural conversations involve more salient personal information and therefore more strongly activate attention? There may be several additional analyses that could begin to tease apart differences in conversation, such as: comparing neural activity during production / comprehension of word vs non-word sounds, or otherwise controlling for various potential confounds such as specific conversation content (autobiographical memories, for example). Being able to more precisely disambiguate the intrinsic differences between natural conversations and pseudo-conversations would be more impactful for the field and would support future experiments using natural

language by informing experimenters exactly they are studying with natural speech vs task-based language.

We appreciate your insightful comments regarding the distinction between natural conversations and pseudo-interactions and have now performed your suggested added analysis. While our study demonstrated significant differences in neural activity between these conditions, we acknowledge that it is challenging to identify the detailed linguistic and cognitive features that impact the neural responses between natural conversation and pseudo-interactions.

Instead, in this work, we focused on using NLP models as a compositional representation of neural features, and demonstrated the difference of neural computation between natural conversation and passive language production and comprehension. This work is among the first studies to reveal the neural processes that support elements of conversation.

Natural conversations involve much more complicated interplay of factors including spontaneity, unpredictability, agency, memory, and attention, in addition to the uncontrolled semantic and syntactical features. To comprehensively identify the factors contributing to the neural difference between natural conversation and passive speaking and listening, it requires new experimental designs from beginning. These experimental paradigms will need to be specifically structured to pinpoint the exact underlying causes with careful controls to eliminate contribution from other factors. It also needs preliminary experiments for power estimation to design the number of channels / participants required for each specific factor. These themselves, will require massive recordings and analytical work, which will not be possible to be addressed within one manuscript, and they are also beyond the scope of this study.

Nevertheless, we followed the suggestion and performed control analysis by instructing a new set of four participants with 465 channels to passively repeat fake sentences that sounds like English, but do not convey any actual meanings (i.e. jabberwocky, for example: “Dorty of the pikap incaged a flup of mool.”). These tasks were performed with the same four participants who performed the passive repeating and listening to sentences control, and each participants spoke and heard the same number of sentences and fake sentences. In this way, we can directly compare the percentage of channels under these two control tasks to identify the contribution of meaningful language materials encoded in channels. We found that there were significantly higher percentage of channels responding to real sentences compared to jabberwocky for both comprehension and production (ANOVA Repeated measurement,  $F = 34.1$ ,  $p = 0.0043$  for speaking;  $F = 18.5$ ,  $p = 0.012$  for listening, **Figure R7**). This showed that coherent, meaningful sentences are processed with a higher percentage of channels compared to meaningless fake sentences, indicating the contribution of semantic feature in neural processing.

Together, aside for these added results, our findings also provide a foundational framework of natural conversation for future research. The observed differences between natural and pseudo-conversations indicate the importance of considering these factors in future studies. Building on our work, further investigations will explore deeper into the specific neural correlates of these factors, ultimately leading to a more refined understanding of human communication.

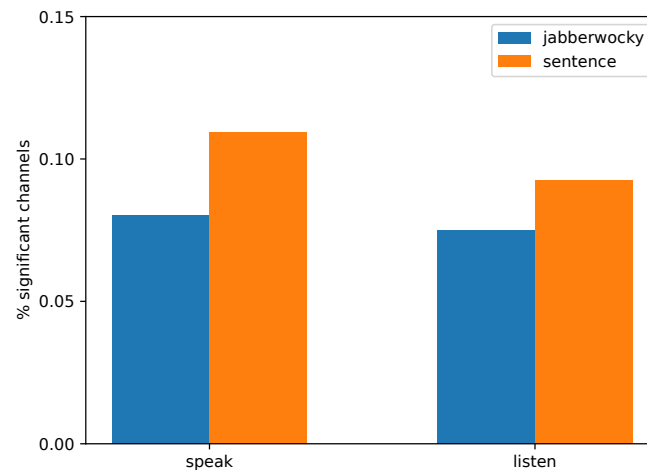


**Figure R7. Channels responding to passively repeating fake or real sentences.** We performed control analysis by instructing the participants to passively repeat fake sentences that sounds like English, but do not convey any actual meanings (i.e. jabberwocky, for example: “Dorty of the pikap incaged a flup of mool.”). These tasks were performed with the same four participants who performed the passive repeating and listening to sentences control, and each participants spoke and heard the same number of sentences and fake sentences. In this way, we can directly compare the percentage of channels under these two control tasks to identify the contribution of meaningful language materials encoded in channels. We found that there were significantly higher percentage of channels responding to real sentences compared to jabberwocky for both comprehension and production (ANOVA Repeated measurement,  $F = 34.1$ ,  $p = 0.0043$  for speaking;  $F = 18.5$ ,  $p = 0.012$  for listening). This showed that coherent, meaningful sentences are processed with a higher percentage of channels compared to meaningless fake sentences, indicating the contribution of semantic feature in neural processing.

We added the jabberwocky control in the **Methods**:

#### Control task with passive listening and speaking Jabberwocky

To further dissociate the processes related to motoric and perceptual aspects of speeches in a conversation from those that reflected meaningful communication relayed through dialogue, we performed a Jabberwocky control when the participants were instructed to listen and repeat lists of pseudowords that sound like English sentences. These lists of pseudowords, while structurally similar to real sentences, lacked actual meaning (e.g. “Dorty of the pikap incaged a flup of mool.”). In this way, we isolate the contribution of the sensory perception and motor articulation from the meaningful speech processing.

We added the jabberwocky control results in **Results**:

Finally, we performed the Jabberwocky control on the same 4 participants, which was composed of strings of pseudowords that sounds like words, but did not convey actual meanings (e.g. “Dorty of the pikap incaged a flup of mool.”). Therefore, even though the participants were producing and perceiving similar sound signatures, they were not engaged in meaningful

conversation, allowing us to dissociate neural signals simply involved in the motoric and acoustic-phonetic components of speech from those involved in meaningful speech of a conversation. We found that there were significantly higher percentage of channels responding to real sentences compared to jabberwocky for both comprehension and production (ANOVA Repeated measurement,  $F = 34.1$ ,  $p = 0.0043$  for speaking;  $F = 18.5$ ,  $p = 0.012$  for listening). These suggest that linguistic meanings conveyed through conversation contribute significantly to the neural processing, indicating the complexity of language comprehension and production.

We edited the **Discussion** accordingly to address this limitation of our study:

However, there are limitations of this study: Our primary focus is on demonstrating the presence of differences between natural conversation compared to passive listening and repeating rather than delving into their underlying mechanisms. Due to the inherent complexity of natural conversation, future studies will be needed to pinpoint the precise linguistic or cognitive factors differentiating it from passive listening and repeating.

- In paragraph 11 of Results, they mention that the results reflect involvement of a distributed rather than localized network, but their analyses are limited to distinct brain areas. Since language is known to be particularly distributed, functional connectivity analyses would be more suited for answering questions related to differences in the neural process that underlie such broad language processes such as speech production versus comprehension.

We agree with you on the importance of considering functional connectivity for understanding distributed language networks. Indeed, these analyses can provide valuable insights into the interactions between brain regions during language processing.

Therefore, we started with functional connectivity analyses based on calculating pairwise correlations between all channels for each participant as done previously (Lee, Y.Y. and Hsieh, S., 2014. PloS one, 9(4), p.e95415.; Srinivasan, R., Winter, W.R., Ding, J. and Nunez, P.L., 2007. Journal of neuroscience methods, 166(1), pp.41-52.). To eliminate the influence of neighboring channels (we used bipolar reference), we excluded adjacent electrode pairs. We then calculate these correlation matrices during different task conditions (e.g., natural speech production vs. comprehension) respectively. Next, to estimate whether these correlations are significantly higher than random fluctuation of the voltage recording, we compare them with the correlation between the same channels, but from different time periods, with one channel during speaking and the other during listening. In this way, we estimate the correlation of randomness between the channels, as the control condition. We then calculate the z-score of the actual correlation compared to the control condition, as shown in **Figure R8**. Similarly, we repeated the same analysis using the channel activity when the participants were passive repeating or listening to meaningless English sentences (jabberwocky).

While our initial findings suggest potential differences in significant connectivity patterns, we also observed substantial overlap in connectivity matrices across the jabberwocky condition. Correlation values between these matrices ranged from 0.26 to 0.93 for both speaking and listening across all four participants, with the mean of 0.48 and 0.67 for speaking and listening, indicating a strong influence of shared neural processes, possibly related to auditory processing

or motor planning. Therefore, while we observe similar functional connectivity patterns between speaking and listening, our current experimental paradigm cannot distinguish whether they originate from low-level auditory / motor processing, or high-level linguistic and social encoding.

Given the complex interplay of sensory, motor, and cognitive processes involved in language, we believe that a more carefully designed approach is necessary to illustrate the specific contributions of these factors to functional connectivity patterns. This would require careful experimental design and advanced analysis techniques beyond the scope of the current study. Therefore, while functional connectivity is a promising avenue for future research, our current findings suggest that caution is needed in interpreting differences in connectivity patterns as direct reflections of language-specific processes.

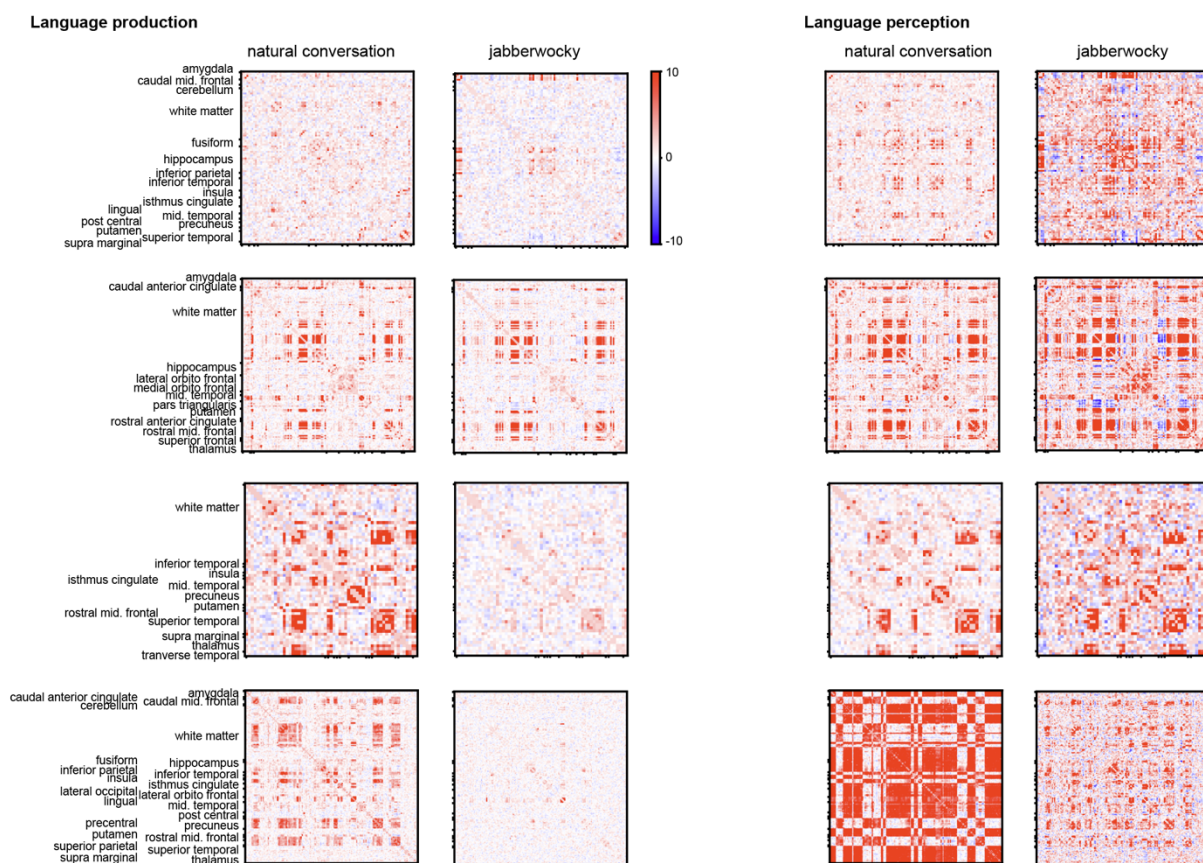


**Figure R8. Functional connectivity between language perception and production.** While our initial findings suggest potential differences in significant connectivity patterns, we also observed substantial overlap in connectivity matrices across the jabberwocky condition. Therefore, a careful experimental paradigm design is required to tell apart neural patterns from low-level auditory / motor processing from high-level linguistic and social encoding.

Discussion:

- The term “linguistic information” is used very loosely/broadly throughout the manuscript, but language is an exceptionally broad construct, so the results are less meaningful without further investigation into which elements of spoken/heard language are at play. Linguistic information, for example, may be related to parts of speech or the specific content of language or features such as linguistic distance, so the manuscript would be strengthened by further investigation into which elements of linguistic information are being related to neural activity here.

We agree with your observation regarding the breadth of the term ‘linguistic information’. Following your suggestion, we have now edited all occurrence of ‘linguistic information’ in the manuscript, and changed them to more specific terms. In addition, we point out in the updated manuscript that language includes various levels of analysis, from phonology to semantics. A comprehensive examination of all potential linguistic features within the context of a single study is both impractical and beyond the scope of this work.

Rather than focusing on isolated linguistic elements, our approach leverages the power of large language models to capture the holistic nature of language processing. By employing a pre-trained GPT-2 model, we aimed to capture a broad spectrum of linguistic information without being restricted to specific features like parts of speech or syntactic structure. This approach allows us to investigate the neural correlates of language in a more ecologically valid manner, reflecting the complexities of natural conversation.

While it is certainly possible to conduct more granular analyses of specific linguistic features, such an approach would require the design of experimental paradigms and control conditions for each feature. Given the potential features in linguistic landscape, it is not feasible to exhaustively explore all potential features within a single study. Instead, our methodology provides a novel perspective on understanding the neural basis of language by examining its relationship to high-dimensional language embeddings, offering a foundation for future research to delve deeper into specific linguistic components.

We have now included these considerations and limitations in the **Discussion**:

However, there are limitations of this study: Our primary focus is on demonstrating the presence of differences between natural conversation compared to passive listening and repeating rather than delving into their underlying mechanisms. Due to the inherent complexity of natural conversation, future studies will be needed to pinpoint the precise linguistic or cognitive factors differentiating it from passive listening and repeating.

- The interpretation and discussion of how this work fills existing gaps in knowledge is weak. Overall, it is not very clear what the value add is or how this work contributes meaningfully to the field and what types of future analyses are now enabled by these findings.

Thank you very much for pointing out your confusion. Natural conversation involves complex computations. We frequently transition between comprehension and production, extracting meaning from speech and planning our own responses. Understanding these neural mechanisms is challenging due to the dynamic, high-dimensional, and contextual nature of conversation. Previous research has often used reductionist approaches, breaking down conversations into smaller components and employing blocked design tasks with predetermined materials. While these studies have provided insights into specific aspects of language processing, they may not

fully capture the complexity of free-flowing speech. As a result, we lack a clear understanding of how sequences of words are represented in the brain during natural conversation, whether common neural processes underlie speaking and listening, and how brain activity relates to information conveyed during transition, especially considering the temporal dynamics of natural speech.

Recent NLP advancements offer a platform for studying natural linguistic interactions. These models capture word sequences and their composition within phrases and sentences, providing a structured representation of language. By correlating neural activity with NLP embeddings, we identified brain regions involved in language encoding and comprehensively investigated the neural mechanisms underlying natural conversation, encompassing speaking, listening, and speaker transitions.

We edited the **Introduction** significantly to address your concern:

The dynamic nature of dialogue, the diversity of information exchanged, and its contextual nature, however, have made the neural mechanisms that underlie conversation in humans a challenge to understand (12). To address these challenges, previous research has largely adopted a reductionist approach, breaking down conversations into smaller, more manageable components. In particular, most studies have used block design tasks that involve predetermined language materials and scripted turn-taking in conversations, and focus on limited aspects of language or transition processes (13-15). Given the dynamics nature of free-flowing speech (16), little work has directly studied neural representation of natural conversation as a continuous process (17), and how language is represented in the brain during natural conversation has remained a challenge to understand.

Specifically, whereas certain brain areas have been shown to distinguish well-formed meaningful sentences from degraded or non-meaningful stimuli, suggesting their involvement in language processing (18-21), the detailed neural process by which sequences of words may be represented in the brain remains largely unknown. Further, while there is often broad overlap between areas involved in language production and comprehension (22-26), whether common neural processes are involved in representing speaking and listening during conversation remains poorly understood. Finally, although certain areas of the brain have been implicated in conversation or turn-taking (13, 27-29), little is known about whether or how brain activity relates to information conveyed during dialogue, especially when considered with the temporal dynamics of natural speech since the methods for analyzing natural communication are still limited.

The recent advancement of natural language processing (NLP) models based on artificial deep learning neural networks has provided a new prospective platform by which to study continuous, natural linguistic interactions. These models have been shown to display high-level performative interactions with human subjects in conversations (30), and can achieve state-of-the-art benchmarks in comprehension-based tasks and question-answering (31-33). These models are capable of capturing specific word sequences and their composition within phrases and sentences through hierarchical layers using vectors (34-36). By providing a structured representation of language, these models may offer a crucial link between the linguistic content and recorded neural activity. Indeed, NLP models have also demonstrated high performance in explaining brain activity during passive listening (37-39), suggesting their capability in representing neurobiological activity and mechanisms. For example, recent studies indicated a shared

contextual space and similar geometric patterns with human brain that facilitates communication (40), where middle and higher layers of the models provide best explanatory power for neural activity (41). In this way, this approach presents a quantifiable method for studying both language production and comprehension, regardless of the specific words and sentences participants use.

Here, we utilize these models as artificial, hierarchically structured vectorized representations of language during natural dialogue. This approach allows us to investigate the neural basis by which the brain processes entire word sequences within the context of conversation as one process, rather than breaking it into small pieces of components. Further, by examining the correlation between neural channel activity and NLP embeddings, we aim to identify brain regions specifically involved in encoding language-related information. This method enabled us to explore how specific sequences of words, together with their compositional semantic and contextual features, are represented in the brain during both speaking and listening, despite differences in language content. Finally, this approach allowed us to compare neural patterns that respond selectively during speaker transitions with those that process word sequences. Together, our approach offers a comprehensive investigation into the neural mechanisms underlying natural conversation by directly examining the integrated processes of speaking, listening, and speaker transitions. This method provides a holistic view of the neural substrates involved in these interconnected aspects of language communication.

Minor comments:

- In the first paragraph of Materials and Methods, “enrolment” should be “enrollment”
- The third sentence in the first paragraph of the discussion is missing a verb. Should be “when the participants were not only engaged in language comprehension...” or something similar.
- Abstract: two occurrences of “through” in the first sentence.

Thank you for these comments. All of them have been edited in the manuscript.

Reviewer #3 (Remarks to the Author):

Large Language Models have made gigantic waves in science, technology, culture, and society, yet they are still largely blackbox devices. The fact that artificial neural networks are extremely simple models of animal neural processing, there is a hypothesis that at least some aspects of how LLMs work reflects neural brain processing, and therefore could be used as a model of brain dynamics, which is far more of a blackbox than LLMs. Therefore, a few labs have been reporting correlation between hidden representations of LLMs and brainwave activity in humans. This paper advances these results in a novel way, and the results are, in my opinion, important enough for publication.

Thank you so much for your kind comments and support.

There are several noteworthy results: 1) Correlation between LLM and brain activity; 2) Interesting patterns of brain activity on p.5, some of which is consistent with previous results, but some is novel.

I find the methodology sound, statistical methods appropriate, and the writing clear.

I have a few recommendations:

1) After a very general introduction, in paragraphs 1 and 2, the authors hint at what is to be done in the paper. I think the introduction could be focused much more in asking the question of correlation between LLM and brain activity, if I am correct that this is the key result. It would be especially useful to articulate why the paradigm of conversation is innovative by comparison to other similar work in this vain.

We appreciate your suggestion for improving the introduction. While we agree that the correlation between neural activity and language models is a central aspect of our study, we believe that framing the research question solely around this correlation would oversimplify the broader goals of our investigation.

Our primary objective is to elucidate the neural mechanisms underlying natural conversation. The correlation between neural activity and language models serves as a crucial tool to achieve this goal by providing a quantitative measure of how well neural representations align with linguistic information conveyed in language. An unconstrained language communication involves unlimited amount of language information with a temporal flow with is not entirely predictable or consistent, making it hard to examine with traditional analytic approaches. To address it, we utilize the development of deep learning models in natural language processing, and creatively applied the embeddings of these models as representation of language. NLP models offer a consistent and elegant way to represent language content using vectors between the modules. Using these models as a bridge, we can ask whether neural activities are representing language features, and we can compare language perception and production in a straightforward way, though the content of each may not overlap. Because our approach does not require a priori knowledge of word timings or fixed transition periods, it is possible to use it in natural conversation.

To make our points more clear, we have rewritten a significant portion of the **Introduction**. We believe this modification enhances the clarity of the introduction without compromising the overall focus of the paper:

Natural conversation is a fundamental mode of human communication, allowing us not only to convey but also derive complex information from language. This core process involves the frequent transition between two distinct but complementary linguistic computations: comprehension and production. Speech comprehension involves a structured succession of processes that extract information from acoustic-phonetic signals allowing us to understand the meaning of sentences and to comprehend the thematic and contextual information being conveyed (1-6). In contrast, speech production planning involves a reverse process whereby higher-order conceptual information is converted to motor planning features for articulation (7-9). These processes are necessary for communicating information during conversations that usually involve rapid alternations between speakers every few seconds (10, 11), including processes for extracting meaning as well as planning and producing our own verbal output.

The dynamic nature of dialogue, the diversity of information exchanged, and its contextual nature, however, have made the neural mechanisms that underlie conversation in humans a challenge to understand (12). To address these challenges, previous research has largely adopted a reductionist approach, breaking down conversations into smaller, more manageable components. In particular, most studies have used block design tasks that involve predetermined language materials and scripted turn-taking in conversations, and focus on limited aspects of language or transition processes (13-15). Given the dynamics nature of free-flowing speech (16), little work has directly studied neural representation of natural conversation as a continuous process (17), and how language is represented in the brain during natural conversation has remained a challenge to understand.

Specifically, whereas certain brain areas have been shown to distinguish well-formed meaningful sentences from degraded or non-meaningful stimuli, suggesting their involvement in language processing (18-21), the detailed neural process by which sequences of words may be represented in the brain remains largely unknown. Further, while there is often broad overlap between areas involved in language production and comprehension (22-26), whether common neural processes are involved in representing speaking and listening during conversation remains poorly understood. Finally, although certain areas of the brain have been implicated in conversation or turn-taking (13, 27-29), little is known about whether or how brain activity relates to information conveyed during dialogue, especially when considered with the temporal dynamics of natural speech since the methods for analyzing natural communication are still limited.

The recent advancement of natural language processing (NLP) models based on artificial deep learning neural networks has provided a new prospective platform by which to study continuous, natural linguistic interactions. These models have been shown to display high-level performative interactions with human subjects in conversations (30), and can achieve state-of-the-art benchmarks in comprehension-based tasks and question-answering (31-33). These models are capable of capturing specific word sequences and their composition within phrases and sentences through hierarchical layers using vectors (34-36). By providing a structured representation of language, these models may offer a crucial link between the linguistic content and recorded neural activity. Indeed, NLP models have also demonstrated high performance in explaining

brain activity during passive listening (37-39), suggesting their capability in representing neurobiological activity and mechanisms. For example, recent studies indicated a shared contextual space and similar geometric patterns with human brain that facilitates communication (40), where middle and higher layers of the models provide best explanatory power for neural activity (41). In this way, this approach presents a quantifiable method for studying both language production and comprehension, regardless of the specific words and sentences participants use.

Here, we utilize these models as artificial, hierarchically structured vectorized representations of language during natural dialogue. This approach allows us to investigate the neural basis by which the brain processes entire word sequences within the context of conversation as one process, rather than breaking it into small pieces of components. Further, by examining the correlation between neural channel activity and NLP embeddings, we aim to identify brain regions specifically involved in encoding language-related information. This method enabled us to explore how specific sequences of words, together with their compositional semantic and contextual features, are represented in the brain during both speaking and listening, despite differences in language content. Finally, this approach allowed us to compare neural patterns that respond selectively during speaker transitions with those that process word sequences. Together, our approach offers a comprehensive investigation into the neural mechanisms underlying natural conversation by directly examining the integrated processes of speaking, listening, and speaker transitions. This method provides a holistic view of the neural substrates involved in these interconnected aspects of language communication.

2) Once the introduction is made much more question-directed, I think the first part of the discussion can similarly be directed at answering the main question.

Thank you for the comment in structuring the introduction and discussion. We edited the text accordingly in the first paragraph of the **Discussion** by specifically talking about the neural response during the three main components of the natural conversation: language production, perception, and transition, as follows:

Major challenges in understanding how our brain supports natural conversation include limitations in precisely recording the relevant neural activity, the difficulty of richly and accurately representing word sequences, and the ability to place both of these problems in the context of natural human interactions (52, 53). By comparing intracranially recorded neural patterns during natural dialogue between individuals to those of NLP models, we have taken a novel route toward overcoming these significant hurdles. This approach importantly allowed us to examine directly the interconnected neural processes of speech production, comprehension, and transition, providing a comprehensive analysis of the crucial components of natural conversation as an integrated system. We found that compositional information about the word and contextual features as captured by the models could similarly explain neural activity patterns found in the brain when the participants were not only engaged in language comprehension, but also during language production. These neural patterns were carried across multiple frequency bands and showed heterogeneity over brain areas, and were largely absent when randomizing the neural data, suggesting that these neural patterns reflected the specific word sequence being

conveyed in natural conversation. Similarly, channels that responded to speaker transitions were not localized in certain brain areas or frequency bands. These together, indicate a well-coordinated and integrated neural pattern underlying the process of natural conversation.

Further, we used 2 paragraphs in Discussion to address the specific questions we asked in the Introduction: (1) whether common neural processes are involved in representing speaking and listening during conversation; (2) whether or how brain activity relates to information conveyed during dialogue. We edited the first sentence of the two paragraphs accordingly:

(1) In addition, comparing neural encodings during speech production and comprehension, we found that information being conveyed in speech production, though, did not simply mirror that of comprehension.

(2) Further, neural patterns partially overlapped between those that encode word sequences conveyed between speakers with those that respond to speaker-role transitions.

3) More comparison of brain results to other recent reports is necessary.

We appreciate your suggestion for additional comparisons with existing literature. Recent advances in neurolinguistics have indeed demonstrated the potential of NLP models to elucidate neural representations during language processing. These deep learning networks have shown neural-NLP correlations, initially observed in functional Magnetic Resonance Imaging (fMRI) studies [Caucheteux 2020, Schrimpf 2020]. They have also revealed that neural activity patterns can be predicted using embeddings derived from various NLP models with different architecture, training and number of parameters. More recently, intracranial electrocorticography (ECoG) recordings, analyzed through hyperscanning techniques, have provided further insights [Zada 2024] into this process.

Together, these studies suggest that NLP models can capture contextual information more effectively than syntactic or articulatory models, indicating a shared contextual space with human brain that facilitates communication. In addition, subsequent research has identified shared geometric patterns between brain activity and NLP embeddings [Goldstein 2024]. Lastly, it has been demonstrated that middle to higher layers of NLP models provide the best explanatory power for neural activity [Kumar 2024]. These findings collectively demonstrate the potential of NLP models to reveal the neural mechanisms underlying natural conversation.

Building on these results from the literature, we have now edited the **Introduction** as follows:

The recent advancement of natural language processing (NLP) models based on artificial deep learning neural networks has provided a new prospective platform by which to study continuous, natural linguistic interactions. These models have been shown to display high-level performative interactions with human subjects in conversations (30), and can achieve state-of-the-art benchmarks in comprehension-based tasks and question-answering (31-33). These models are capable of capturing specific word sequences and their composition within phrases and sentences through hierarchical layers using vectors (34-36). By providing a structured representation of language, these models may offer a crucial link between the linguistic content and recorded neural activity. Indeed, NLP models have also demonstrated high performance in explaining

brain activity during passive listening (37-39), suggesting their capability in representing neurobiological activity and mechanisms. For example, recent studies indicated a shared contextual space and similar geometric patterns with human brain that facilitates communication (40), where middle and higher layers of the models provide best explanatory power for neural activity (41). In this way, this approach presents a quantifiable method for studying both language production and comprehension, regardless of the specific words and sentences participants use.

In addition, we also added a paragraph in the **Discussion** to discuss the neural activity during natural conversation compared to block design language tasks:

To support smooth speech production and comprehension in natural conversation, neural activity is characterized by a distributed and widespread patterns of activation (17). We found that responding channels spanned across various cortical and subcortical areas (e.g. frontal and temporal cortices, hippocampus, and amygdala), and were significantly reduced during pseudo conversation controls. These findings collectively suggest that natural conversations involve an integration of neural signals across various processing pathways. In contrast, previous studies showed localized neural engagement (e.g. STG for comprehension, precentral for production) for language specific activities (9, 37, 54, 55). By examining neural activity in these regions during natural conversation, we found a higher percentage of responding channels from the established language areas, with more than 20% of responding channels located in superior temporal cortices during comprehension and the left precentral area during production. This heterogeneity suggests that, while conversation engages widespread areas of the brain, language areas still maintain a central role in the natural verbal communication (55). Together, the broad and heterogeneous distribution of neural response suggests the complex and interconnected nature of language processing during real-world communication.

Caucheteux, C. and King, J.R., 2020. Language processing in brains and deep neural networks: computational convergence and its limits. *BioRxiv*, pp.2020-07.

Goldstein, A., Grinstein-Dabush, A., Schain, M., Wang, H., Hong, Z., Aubrey, B., Schain, M., Nastase, S.A., Zada, Z., Ham, E. and Feder, A., 2024. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature communications*, 15(1), p.2768.

Kumar, S., Sumers, T.R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K.A., Griffiths, T.L., Hawkins, R.D. and Nastase, S.A., 2024. Shared functional specialization in transformer-based language models and the human brain. *Nature communications*, 15(1), p.5523.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E.A., Kanwisher, N., Tenenbaum, J. and Fedorenko, E., 2020. Artificial neural networks accurately predict language processing in the brain. *BioRxiv*, pp.2020-06.

Zada 2024: Zada, Z., Goldstein, A., Michelmann, S., Simony, E., Price, A., Hasenfratz, L., Barham, E., Zadbood, A., Doyle, W., Friedman, D. and Dugan, P., 2024. A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. *Neuron*.

4) GPT-3 and GPT-4 systems have been argued to be far more conversationally able than GPT-2. If it is easy to reevaluate the results on these, the paper will be more current, and the correlations maybe higher.

We agree on your point about the potential benefits of using more advanced language models like GPT-3 or GPT-4. However, since our work requires comparing channels to the hidden embeddings of NLP models, we are limited to only *open-sourced* models, while GPT-3 and GPT-4 do not make their model weights and architectures publicly available. In addition, we are limited by the computational resources. We currently cannot load models with more than 40 billion parameters.

However, given these limitations, we added more recent models that contain various number of parameters and trained independently (as explained to address Reviewer #2's comment). They are composed of 4 families of pretrained models, including Bloom (with 560 million and 1 billion parameters), Falcon (1b and 40b), dolly (3b and 12b), and llama (3b, 7b, 13b, wizard-13b, 30b, chronos-33b). These models contain different number of layers and number of hidden embeddings per layer. Due to differences in model depth and embedding dimensions, we standardized model outputs by reducing them to 100 principal components before correlating with neural activity. We found that there were significant higher percentage of channels correlated to all models examined compared to chance ( $p < 1 \times 10^{-23}$ ) during both speaking and listening (**Figure R3**). Further, we found a strong positive correlation ( $R > 0.80$ ,  $p < 2 \times 10^{-3}$ ) between model size and the proportion of correlated channels. Therefore, our findings of correlations between channels and NLP embeddings are more generalizable to NLP models with diverse sizes and families. Our results also support the reviewer's hypothesis that larger models correlate with higher percentage of channels.

We added the Figure below into the Supplementary **Figure S13**:

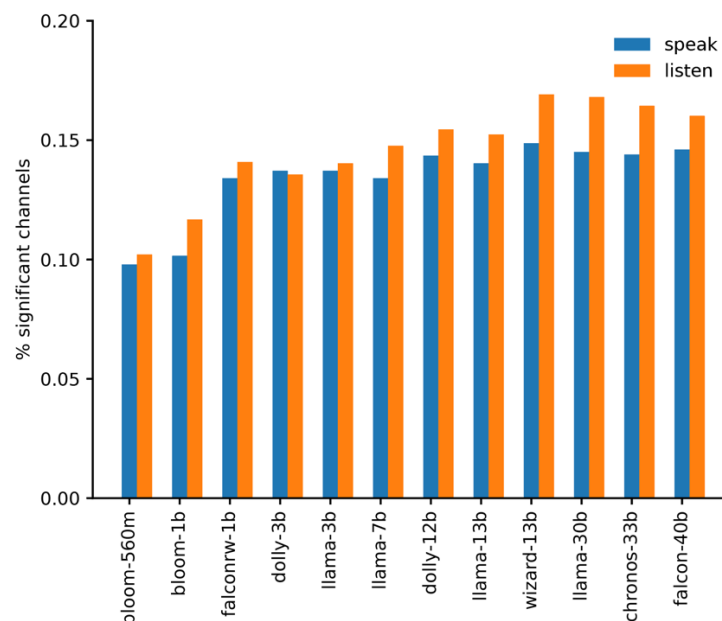


**Figure R3 (S13). Percentage of significant channels examined by various NLP models.** By applying NLP models with various size, number of layers, and families, we found that there were significant percentage of channels correlating to all models for both speech production and comprehension (Chi square proportion test,  $p < 1 \times 10^{-23}$ ), with larger models showing higher percentage of correlated channels ( $R > 0.80$  and  $p < 2 \times 10^{-3}$ ).

We edited the **Results** section:

Thirdly, to confirm that the above findings were generalizable, we compared the neural activity patterns to those of another NLP model trained on a different dataset. Here, we used a BERT (base) model with a bidirectional network architecture and trained on different linguistic materials (46). Again, we found that the proportion of correlated channels was significantly higher than chance (Chi-square test, statistic = 15278,  $p < 10^{-100}$ ), and was greater than or equal to that observed with the GPT-2 model (Fig. S7b). Further, similar findings were observed across a series of NLP models with different layers and architectures (Fig. S13, Chi-square proportion test,  $p < 1 \times 10^{-23}$ ), with an increase of the percentage of correlated channels to the size of NLP models ( $R > 0.80$  and  $p < 2 \times 10^{-3}$ ). These suggest that the relationships between neural activity patterns in the brain and those of the language models are generalizable properties of neuronal responses.

We added the figure in **Methods**, and a paragraph to describe these models:

To confirm the generalizability of our results, in addition to these two NLP models, we examined other models that contain various number of parameters and are trained independently. They majorly composed 4 families of pretrained models, including Bloom (with 560 million and 1 billion parameters), Falcon (1b and 40b), dolly (3b and 12b), and llama (3b, 7b, 13b, wizard-13b, 30b, chronos-33b). These models contain different number of layers and number of hidden embeddings per layer. To account this variation, we took the first 100 principal components and then correlated them with the neural activity. We found that there were significant higher percentage of channels correlated to all models examined (Figure S13, Chi square proportion test,  $p < 1 \times 10^{-23}$ ) during both speaking and listening. Further, the percentage of correlated channels showed an increase trend with the size of NLP models ( $R > 0.80$  and  $p < 2 \times 10^{-3}$ ). By extending NLP models with much diverse sizes and model families, we showed that all of them were significant higher percentage of correlated channels than chance, indicating the generalizability across various of models. In other words, these findings appear to be robust and generalizable property of neural activity.

5) I would appreciate more details about the layers than just saying lower and higher layers.

We appreciate the reviewer's interest in a deeper understanding of the layer-dependent correlations. In NLPs, the processing of language occurs in a hierarchical manner. However, unlike rule-based systems, these models do not have explicit, human-readable rules for each linguistic level. Instead, they learn statistical patterns from vast amounts of text data, which allows them to approximate these linguistic processes. As a result, these layers do not correspond directly to known linguistic categories. The initial layers typically focus on processing low-level features like word-level patterns. Middle layers then start to capture more complex linguistic structures, grammatical relationships, and semantic meanings. The deeper layers begin to model higher-level abstractions and contextual understanding.

To compare the detailed neural activity to NLP layers showing in Figure 2e, we found that similar pattern for speech production and comprehension, with lower percentage of significant channels in the input layer (8% for speech production and 9% for speech comprehension). Then there was a gradual increase with the increase of layers, reaching a peak at the 8<sup>th</sup> layer (11% for speaking and 13% for listening). At higher layers towards the output, we observed a decrease of significant channels with the increase of layers.

We have edited the **Results** section accordingly to include this a detailed description of these correlations across layers:

Neural-to-model correlations during conversation were dependent on the hidden layers within the language models. In NLP, the processing of language occurs in hierarchical layers, which extract the statistical patterns from vast amounts of text data. Although layers do not correspond directly to known linguistic categories, lower (input) network layers preferentially reflect information about individual words independent of their context, whereas middle layers capture the more complex linguistic structures, and higher (output) layers reflect integrated compositional sentence-level information (47-51). Therefore, to further examine the relationship between neural activity patterns and that of the NLP model, we calculated the percentage of channels correlating with each network layer, and found that neural activities preferentially aligned with higher network layers for both speech planning and comprehension (12% and 14% for the first and last 6 layers during speak, T-test statistic = -3.2,  $p = 1 \times 10^{-3}$ ; 14% and 18% for comprehension, statistic = -4.3,  $p = 1.8 \times 10^{-5}$ , **Fig. 2e**). Similar layer-dependent patterns have been observed for speech production and comprehension, with lower percentage of significant channels in the input layer (8% for speech production and 9% for speech comprehension), followed by a gradual increase, reaching a peak at the 8th layer (11% for speaking and 13% for listening). At higher layers towards the output, we observed a decrease of significant channels with the increase of layers. This neural distribution across layers was significantly different from those when the participants were passively listening versus articulating the sentences they heard (repeated measurement ANOVA,  $F(12) = 23.3$ ,  $p = 4 \times 10^{-4}$  for listening,  $F(12) = 5.6$ ,  $p = 0.036$  for speaking). Further, the specific layer distribution was dependent on the brain area (repeated measurement ANOVA,  $F(96) = 4.2$ ,  $p < 1 \times 10^{-5}$  for speak;  $F(96) = 6.7$ ,  $p < 1 \times 10^{-5}$  for listen, **Fig. S9**), with those from temporal lobe exhibiting higher layer-dependence compared to those in frontal cortex (temporal: 11.8% for first 6 layers, 15.0% for last 6 layers during production planning; frontal: 9.1%, 9.9%). In short, the majority of neural activity patterns likely reflect higher order contextual information rather than low level early processing or pre-speech motor planning.

6) "Multiple Comparisson" in Figure 1F has an extra "s"

Thank you for pointing it out. We corrected the spelling.

7) Manning et al.'s 2020 paper in PNAS would be important to cite, not just the Belinkov papers

Thank you for suggesting this paper! Yes, it is very relevant to our work, and we have cited in our main manuscript.

**Reviewer #1** (Remarks to the Author):

I appreciate the authors' revision. However, I believe that not all of my comments have been fully or appropriately addressed.

The authors mentioned in their rebuttal letter that their multiple tests were independent. However, these tests were NOT independent. For example, as described below, after conducting 17 tests, the authors used the results of these tests to extract representative regions of interest for their subsequent analysis.

-----  
Further, across 17 brain areas with a high number of channels implanted ( $n \geq 25$ , Method), there were 9 areas that showed significantly high percentage of channels ( $p < 0.01$ ) that correlated to NLP embeddings, including areas in the temporal and frontal cortex, thalamus and limbic system structure (Fig. 2a). Thus, in subsequent analyses, we used these 9 areas as representative regions of interest.

-----  
In general, there are cases where multiple comparison correction is unnecessary. However, this is true only when the tests are entirely independent. The methods the authors employed may risk yielding false positives.

Thank you for your thoughtful feedback regarding the multiple comparison correction. As you pointed out, it is important to clearly describe and correctly apply the appropriate statistical corrections.

We have therefore now (1) clarified that all analyses were performed using all channels from all areas; (2) we implemented an even stricter Bonferroni correction across all 17 brain areas to limit the possibility of false positive detections and to further confirm that our results were indeed statistically robust; (3) we tested additional NLP models to ensure that our results were consistent and generalizable; and (4) we implemented new tests to confirm that our results were reliable across statistical corrections. Specifically:

**First**, we realize that the statement from our original manuscript “*in subsequent analyses, we used these 9 areas as representative regions of interest*” was misleading, and we apologize for resulting confusion. To clarify, **the selection of 9 areas was made solely for visualization purposes**. In our original manuscript, we limited the number of displayed brain areas to avoid introducing an excessive variety of colors, which might have made the visualization less clear for readers, and to highlight the most prominently represented areas. Therefore, we clarify that we did not limit our analyses to these 9 areas but rather used **all contacts and areas**: In other words, all analyses were conducted using the **complete set** of 1,910 channels across **all brain areas**, and these results are what is presented in both the main manuscript and supplementary materials. We have now made this more explicit in the text.

**Second**, following your suggestion, **we implemented a stricter Bonferroni correction across all 17 brain areas** to ensure that our results were indeed statistically robust and to further limit the possibility of false positive detections. The significance threshold was therefore set here for Bonferroni correction of 17 brain areas:

$$p < \frac{0.01}{17} = 6 \times 10^{-6}.$$

This threshold was used to determine whether any of these areas had a higher percentage of responding channels. Beyond this, prior to this step, we also applied a Bonferroni correction to account for multiple comparisons across the NLP nodes, layers, and channel frequency bands to determine whether a specific channel showed significant correlation with NLP embeddings ( $p < 0.05 / 768 / 13 / 5 = 1 \times 10^{-6}$ ). Using this highly conservative updated threshold, we found that 4 areas in the left hemisphere (24% of all areas) showed significantly higher than chance percentage of channels responding to NLP embeddings during language production, and 5 areas in left hemisphere (29% of all areas) during language comprehension. We have now illustrated these updated results in **Figure R1** and display the ratio of significant versus non-significant channels for each area, with significant areas further color-coded by hemisphere (blue: left hemisphere, orange: right hemisphere).

**Third**, to further confirm the consistency of our results (**Figure R2 left**), we replicated our analysis to include 12 additional large language models that differed in their architectures. In other words, we wanted to ensure that our results were not simply explained by a particular model but were rather consistent across various language models. These models include Bloom, Falcon, Dolly, Llama models with different numbers of layers and parameters, as well as different training and fine-tuning procedures. Here, we maintained the same strict statistical criteria, including Bonferroni correction to 17 brain areas, and found similar results with a one-way ANOVA test, revealing no significant differences in the percentage of responsive channels across models ( $F = 0.89$ ,  $p = 0.56$ ). These added analyses therefore further confirmed that our results were not only robust but also consistent.

**Finally**, to additionally highlight the stringency of our statistical corrections aimed at limiting false positive detections, we compared our current results (based on Bonferroni correction) to that of a False Discovery Rate (FDR) correction. Here, correcting for 17 brain areas, we find  $4/17 = 24\%$  of areas in the left hemisphere for production with FDR correction (compared to the  $4/17 = 24\%$  of areas with Bonferroni correction). Similar findings were also made for comprehension ( $6/17 = 35\%$  and  $5/17 = 29\%$  for FDR and Bonferroni correction, respectively), suggesting that the percentage of responding areas was reliable across different statistical correction methods. Lastly, as described in the methods and our past response, we performed a power analysis to estimate the possibility of false negative detections. This analysis demonstrated the ability to distinguish brain regions with 20% significant channels from those with 5% at a statistical power of 80% for a total of 25 channels. This further confirmed that our sample size was adequately powered. Therefore, in sum, we used several metrics to ensure the statistical robustness of our results. These include measures to limit the possibility of false positive detections and now also include additional clarification and correction to account for the numbers of areas being tested.

These added analyses and clarifications have now been included in the updated manuscript. Below, we provide a summary of some of the main changes made to the main manuscript. All changes made to the manuscript are highlighted in [blue](#).

We updated part of the results section which now reads “Further, across 17 brain areas with a high number of channels implanted ( $n \geq 25$ , Method), there were 9 areas that showed significantly high percentage of channels ( $p < 0.01$ ) that correlated to NLP embeddings, including areas in the temporal and frontal cortex, thalamus and limbic system structure (**Fig. 2a**). Thus, in subsequent figures, we used these 9 areas for visualization purpose to avoid overlapping of colors. We also examined the percentage of areas with Bonferroni correction for multiple comparison on 17 brain areas using GPT-2 model (**Fig. S12**), as well as showing consistent response across other NLP models (**Fig. S13**).”

Second, we updated Figure S12 with pie charts to better illustrate the specific areas showing higher percentage of NLP-responding channels:

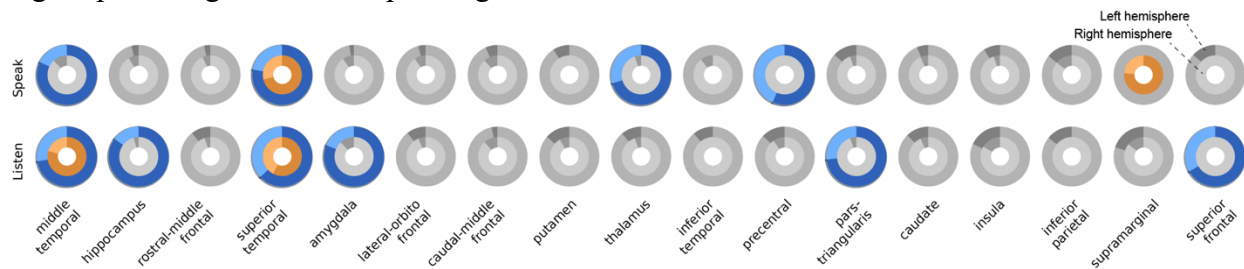


**Figure R1 (S12). Brain areas with significant high percentage of NLP-correlating channels calculated with Bonferroni corrections.** The percentage of NLP-correlating channels were illustrated in each subplot across different brain areas, hemispheres, and modalities. To examine whether there were areas with significant higher percentage of NLP-correlating channels, we performed Bonferroni correction of the brain areas (17 areas), after the Bonferroni correction of NLP nodes, layers and frequency bands. Areas with significantly higher percentage of channels were showing in colors (blue for left hemisphere and orange for right hemisphere).

Finally, we added **Figure S13** in the supplementary materials to show the consistence across various NLP models:

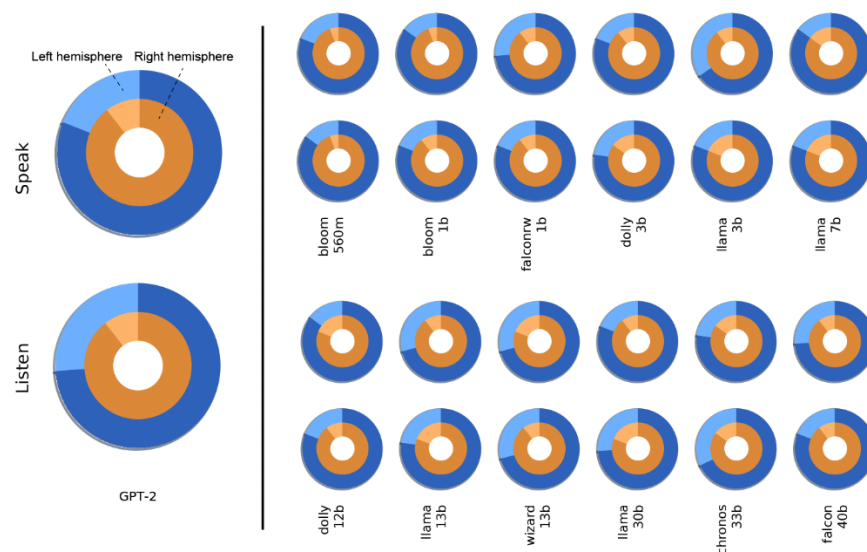


**Figure R2 (S13). Percentage of NLP-responding brain areas across different NLP models after Bonferroni corrections on multiple comparison of brain regions.** *Left*, using GPT-2 model, the percentage of brain areas grouped by hemispheres is plotted during speech production and comprehension. The percentage of brain areas were calculated based on whether they showed significantly higher ratio of NLP-correlated channels with Bonferroni correction for multiple brain areas (17 areas), after Bonferroni correction for NLP nodes, layers, and channel frequency bands. The colors label the hemispheres. *Right*, same analyses were performed on a range of pre-trained state-of-the-arts large language / NLP models with varying sizes, layers, and training datasets. The percentages of NLP-responding brain areas were consistent when using these different NLP models.

Additionally, while the authors created Figure S12, there is no reference to it in the main text.

Thank you for pointing it out. We have added the reference in the main text.

**Reviewer #2** (Remarks to the Author):

My comments have been addressed.

**Reviewer #3** (Remarks to the Author):

I am happy with the revisions to my comments as well as the comments by other reviewers. The changes are extensive, and they significantly elevated the manuscript.

Reviewer #1 (Remarks to the Author):

I appreciate the authors' revision and believe that my comments have been appropriately addressed.

In the text, the supplementary figures appear in the following order: Fig. S1, Fig. S2, Fig. S12, Fig. S13, ...

Thus, the supplementary figures should be re-numbered to ensure sequential order.

Thank you very much for your feedback and for acknowledging that our revisions have addressed your comments. Regarding your observation about the numbering of the supplementary figures, we have ensured that the supplementary figures are re-numbered appropriately to maintain sequential order for clarity and consistency in the manuscript.