nature portfolio

Peer Review File

Progress, Challenges and Future of Linguistic Neural Decoding with Deep Learning

Corresponding Author: Professor Yanfeng Wang

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)
Wang and colleagues take a stab at surveying the literature on neuroscience and AI focused on the particular topic of language in the brain. Overall, it remains unclear what the unique contribution of the article is exactly? How does it set itself apart from other similar papers? In particuar, the AI side of the paper remains largely superficial, which makes this papers mostly a neursocientists' paper with Ai applications, without a strong technical grounding, which is also supported by the kinds of journals that are cited (see point 4 below). This reviewer does not recommend publication of the present work in Communications Biology.

Major
1) Scope: From the abstract and the beginning of the manuscript draft it is difficult to this reviewer to extract whether this is a Neuro-to-AI or AI-to-Neuro paper, such as are they trying to build models to i) simulate the brain or ii) emulate human cognition, arguing that the underlying processing principles are identical or not? 'reviews of neural decoding' is insufficiently precise in this reviewer's eyes.
2) Flow and narrative: 2nd paragraph appears to come out of the blue; why is it suddenly emphasizing time series so much?; please improve this transition as well as paragraph-to-paragraph transitions in the paper in general. Further, the concrete examples and cited papers appear to be often poorly motivated. As it stands, the paper strikes as a stream of largely disconnected isolated examples.
3) Metrics: …which are mentioned appear random, as they do not appear to reflect commonly used performance metrics from the NLP community. So based on what basis where the highlighted metrics exactly chosen for treatment in the present work ?
4) Interdiscplinarity: The paper tries to position itself at the Neuro-AI front. Yet, most citations are from neuroscience venues, rather than AI conferences such as NeurIPS and ICML, which raises doubts about the authenticity of the transdisciplinary grasp of this work.

Minor
1) 'high-parameter ANNs': unusual phrasing in the core technical machine learning literature

2) 'In the textual stimuli classification (TSC) paradigm': never heard about this, perhaps the authors could expand more?

3)'The major languages in the world have their textual form, and communication via text outperforms due to its efficiency and concision': ok but how is this related to Neuro and AI ? Please sharpen.

4)'perception methods': unclear to this reviewer

Reviewer #2

(Remarks to the Author)
As a cognitive scientist, I was hooked by how this review was introduced:

"Despite advancing insights into the human brain, the intricacies of linguistic representation and processing still largely resemble a black-box model, making it challenging to explain intuitively the internal mechanisms and patterns of information transfer. Deep learning presents a viable solution for understanding language in the brain by utilizing large-scale trainable parameters to map the correlation between external stimuli and neural activity."

After reading this, I assumed that the review would aim to draw parallels between how models encode and decode meaningful linguistic signals and how human brains and cognitive faculties support language. However, this is not what the review set out to do, and so that is not what the review achieves. Instead it provides a sweeping overview of recent technical feats of linguistic neural decoding.

The authors have done great work organizing the science as they have for this paper, and it is inspiring to review what has been achieved already and imagine what will come next. However, I feel the review falls short of moving the field toward that future by offering commentary on outstanding questions.

While the review is comprehensive from a technical perspective, it does not seem to have a thesis other than "bigger models are better". The review did not make any particular claims or resolve any open questions. For example, there is little discussion of why some model architectures are better than others at modeling the brain. The review refers to both the brain and the ANNs used to decode the brain's activity "black boxes" at different points, but (I believe) we need to understand the inner workings of both before the kinds of BCI communication tools that the authors envision arriving in the futre will be possible.

I agree with the authors that these models and improvements to brain computer interfaces can do a lot of good in the world, and it is exciting to read about the progress that is being made. The being said, other than providing a curated selection of exciting technical advances, I do not feel that this review offers insight into the way ahead.

There are points in the review where I felt the authors oversimplified or were inattentive to what is most interesting about a connection being drawn between models and the brain. For example, the authors write that "the computations of ANNs account for a significant portion of the variance observed in brain activity", but its not clear which "computations" that the ANN is doing that are relevant to capturing variance in neural activity. Is this just saying that model similarity structure defined among hidden-layer activation patterns correlates with neural similarity structure? Or, is a point being made that Transformers are specifically doing something brain-like? Given that the LLM and the brain are each learning about linguistic structure and word meaning, finding that they both encode similar structures is not inherently interesting---there is plenty of evidence that the concepts learned by LLMs are not quite what humans learn (e.g., https://escholarship.org/uc/item/0mk2t7gz).

The brain recording translation models are interesting, but are presented in a way that feels off. Because of the LLM sitting between the brain features and the output, it seems that part of the prediction ability for subsequent words will be carried by the information the LLM encodes about high-probability sequences of words. In the example where there true text was "At best they tales ..." and the prediction was "At best it is macred...", it seems that "it is" likely emerged from that expectations of the LLM rather than the neural signal. I am surprised that the LLM generated "macred"---that is not an English word as far as I know. I am not sure how this model could produce non-words.

Anyway, it is not clear how much of the brain recording translation performance is driven by identifying and leveraging neural representations that truly encode the lexical content being "decoded", because the decoding pipeline seems to allow the LLM itself to infer words and sentence structure. This matters from a biological and psychological perspective--wanting to use these models to understand the brain and mind--than it does from the perspective of someone wanting to build a tool to allow effective communication for someone who has lost the ability to speak and/or write.

There are many models in the paper that integrate neural information over time. I will note that the cognitive neuroscience of semantic knowledge currently has little to say about how conceptual knowledge is encoded in time-varying neural signal. It is largely unknown--there are many reasonable ways to implement a model that uses time-varying signal when attempting to decode semantic knowledge, but that implementation makes important assumptions about how the brain and mind work. When thinking about linguistic decoding from biological and psychological perspectives, these details of model implementation are more interesting than the model performance--or rather, the model performance is difficult to interpret without that context.

The Future Trends section seems motivated by the idea that technical advances, more compute power, and bigger models are all that is needed to unlock the contents of the mind. For example, the authors write:

"Progress in decoding complex neural signals has shown the potential to reconstruct thoughts, images, or even dreams. This could have profound implications for understanding consciousness and developing communication methods."

Decoding words within fixed vocabularies or generating text/speech from a limited space of neural signals corresponding with articulation is not analogous to freely decoding thought and mental imagery or providing readouts of dreams. The leap to discussing consciousness is nonsense without defining the concept. It is a stretch to suggest that models of the kind discussed in the review are bringing us closer to understanding the nature or mechanics of the brain and mind, given that the review treated the brain as black box.

The calls for future directions are general and not particularly insightful. Yes, better non-invasive neuroimaging will of course improve the ability to decode language from neural activity, as will better BCI technology. But what most needs to be improved? Spatial resolution, temporal resolution, signal to noise, alignment across individuals? Likewise, a call for "universal decoding schemas" to address the fact that neural signals and the concepts they correspond to are non-stationary over time restates the problem but does not point in the direction of a solution.

Reviewer #3

(Remarks to the Author)
Summary:

This paper presents a comprehensive overview of recent advancements in linguistic neural decoding with deep learning methods, especially methods related to large language models (LLMs). The authors provide an overview of how language is represented in the brain and can be decoded from the brain in five sections: Brain-Network Alignment, Brain Decoding Evaluation, Stimuli Recognition, Brain Recording Translation, and Speech Neuroprosthesis. Finally, the author discusses future trends in linguistic decoding research.

Major concerns:

1. The author might need to provide a clearer definition of what is linguistics decoding. The paper should explain why decoding methods like imaged handwriting [1], motor imagery controlling methods, and SSVEP-based spelling methods [2] are not involved in this survey (or just involved in additional sections). However, motor-based speech decoding is involved. In my opinion, both methods are linguistic decoding from external human behaviors, i.e., conveying semantic content through speaking, writing, or typing.
2. The purpose of Section 2 is somewhat unclear. Is it solely to support the following three chapters? If so, what should be the metrics for alignment?
3. As a survey paper, it is important to help people understand the current state and ability of linguistic decoding. Providing more examples and explanations of decoding results could be beneficial, especially to readers who are not specialists in the field. We need to avoid exaggerations and show the true existing ability of semantic decoding.
4. The author can consider including more tables that clearly outline the settings and decoding performance of existing work. The settings could involve BCI devices, types and granularities of decoding, type of language, etc. Table 2 is a good example in the translation section, and similar tables could be provided for other sections.
5. The distinctions between these four types of linguistic tasks could be illustrated using a diagram or a table and provide more explanation.
6. Additional insights could be provided to help discoveries and research in related fields. The insights currently mentioned by the author are mostly related to brain decoding, such as lacking of consuming non-invasive, high-quality data, and being subject- and time-invariant. These issues pertain to brain data acquisition and brain signals but are not specifically related to linguistic decoding. In fact, these limitations could apply to any article on brain signal decoding. For example, can the author summarize the existing deep learning architectures specially designed for linguistic decoding and its relevant challenges?
7. Although the author discusses many aspects related to large models and semantic decoding, it is worth exploring the significance of large models beyond serving as representation and providing language knowledge to help context-based translation. Especially compare LLM-related linguistic research to previous neurolinguistics research, which includes how humans reason based on language, their visual and auditory understanding of language, the different granularity of linguistic understanding from lexeme to sentence, and temporal analyses of linguistic understanding.

Minor concerns:

1. The author may need to explain what is teacher forcing addressed in Section 4 (refer to https://github.com/NeuSpeech/EEG-To-Text).
2. There are some typos that need to be corrected, such as the quotation marks and missing commas in the caption of table 1.
3. The author should provide more discussions on the ethical issue related to linguistic decoding. Although I agree that relevant techniques are in an early stage, it will emerge serious ethic problems if we can read people's minds.
4. The author has used many illustrations from previous articles; please ensure that these figures' permission has been obtained.
5. The author can consider more relevant papers. I am listing some suggestions here, but the author does not need to adopt all of them.

[1] Willett F R, Avansino D T, Hochberg L R, et al. High-performance brain-to-text communication via handwriting[J]. Nature, 2021, 593(7858): 249-254.

[2] Chen X, Wang Y, Nakanishi M, et al. High-speed spelling with a noninvasive brain–computer interface[J]. Proceedings of the national academy of sciences, 2015, 112(44): E6058-E6067.

[3] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. Nature human behaviour, 7(3):430–441, 2023.

[4] Tuckute G, Kanwisher N, Fedorenko E. Language in brains, minds, and machines[J]. Annual Review of Neuroscience,

2024, 47.

[5] Artificial neural networks accurately predict language processing in the brain. M Schrimpf, I Blank, G Tuckute, C Kauf, EA Hosseini… - BioRxiv, 2020

[6] Yin C, Ye Z, Li P. Language Reconstruction with Brain Predictive Coding from fMRI Data[J]. arXiv preprint arXiv:2405.11597, 2024.

[7] Abdou M, Gonzalez AV, Toneva M, Hershcovich D, Søgaard A. 2021. Does injecting linguistic structure into language models lead to better alignment with brain recordings

[8] Anderson AJ, Kiela D, Binder JR, Fernandino L, Humphries CJ, et al. 2021. Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. J. Neurosci. 41(18):4100–19

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

[10] Zou S, Wang S, Zhang J, Zong C. 2022. Cross-modal cloze task: a new task to brain-to-word decoding. In Findings of the Association for Computational Linguistics: ACL 2022, ed. S Muresan, P Nakov, A Villavicencio, pp. 648–57. Kerrville, TX: Assoc. Comput. Linguist. https://doi.org/10.18653/v1/2022.findings-acl.54

[11] Warstadt A, Bowman SR. 2022. What artificial neural networks can tell us about human language acquisition. arXiv:2208.07998 [cs.CL]

Reviewer #4

(Remarks to the Author)
Wang et al. 2024 contribute a review on neural decoding, i.e., the decoding of speech/language from the human brain. The strengths of the manuscript is that are: i) it is very timely, ii) the authors have compiled a large set of studies and the taxonomy presented in Figure 1 is great, iii) the figures are generally helpful. I think the main weakness of the paper is a lack of clear writing, definitions, and logical flaws, which I provide examples of below. I think the content of the manuscript is good, but it needs a large revision in terms of clarity. I do not recommend this manuscript for publication in its current form.

Logical flaws / unclear writing
I do not think that the review is motivated well, and I see logical flaws in several places. I will start with the abstract:

“Linguistic neural decoding reconstructs the stimuli information perceived by text or speech, which has shown groundbreaking achievements, thanks to the improvement of neuroimaging, biology and artificial intelligence.”
How has “biology” improved and how can we thank biology for making groundbreaking achievements? Biology has arguably not changed in the past few years and has not enabled advancements in neural decoding?

“In this work, we present a taxonomy of recent neural decoding progress, focusing on deep learning architectures and strategies, especially those implementing large language models (LLMs) for their powerful reasoning capacity.”
I believe that most studies do not leverage “LLM reasoning”, namely the embeddings. It is true that LLM reasoning can be leveraged, but I do not think that is the main focus of the field.

“This article will help brain scientists and deep learning researchers gain a bird’s eye view of language perception, and thus facilitate their further investigation.”
Why is language denoted as “perception”? The first sentence in the introduction says that language is exclusive to “higher-order” organisms? Language—as opposed to speech perception—is not tied to perception given that certain brain regions are active regardless of the perceptual modality (e.g., Deniz et al 2017).

A few other examples throughout the paper include:

Page 4: “When processing natural language, ANNs exhibit patterns of functional specialization similar to those of cortical language networks”.
Which studies have shown this? Functional specialization is a strong claim (see e.g., Kanwisher, 2010).

Page 4: “it has been verified that the brain encoding models and pre-trained LLMs follow the scaling laws, where the model performance increases as the number of parameters grows, indicating the necessity to develop more complex architecture to bridge the brain activity patterns and human linguistic representations”
I don’t think it logically follows that one needs to build novel architectures? If scaling works, then one might argue that it is enough to make the models bigger? Of course, novel architectures would be helpful, but I don’t think it follows from the references provided in these sentences.

The authors should define what they mean by “language”. Does it entail speech/visual perception? Or do they talk about

language as a “higher-level”, amodal process (e.g., Fedorenko et al. 2024)?

Many model types are mentioned. Oftentimes it is not mentioned what those are, e.g., VQ-VAE (page 11) -- it is unclear why that is of relevance. If the authors chose to mention these model types, it needs to be clear why.

References

Deniz, Fatma, et al. "The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality." Journal of Neuroscience 39.39 (2019): 7722-7736.

Fedorenko, Evelina, Anna A. Ivanova, and Tamar I. Regev. "The language network as a natural kind within the broader landscape of the human brain." Nature Reviews Neuroscience (2024): 1-24.

Kanwisher, Nancy. "Functional specificity in the human brain: a window into the functional architecture of the mind." Proceedings of the national academy of sciences 107.25 (2010): 11163-11170.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I still don't believe that this piece is at the level of Communications Biology. However, let the other two reviewers take the lead on the decision.

Reviewer #2

(Remarks to the Author)
I thank the authors for their thoughtful responses to my initial critiques. It seems that the authors and I have fundamentally different views on what a review of neural decoding should aspire to achieve. I am coming at this from the perspective of a cognitive neuroscientist interested in advancing the basic science of semantic cognition who sees the potential for machine learning/AI to aid in that pursuit. This is why I was eager to find connections between the state-of-the-art technology and cognitive neuroscience being drawn throughout the review. I may not be representative of the readership of Current Biology.

While the revised manuscript is more explicit about communicating its intended scope, it remains my opinion that the review would be strengthened by commenting on architectural/computational/representational differences between BrainTranslator, DeWave, UniCoRN, etc. in the brain recording translation section.

Reviewer #3

(Remarks to the Author)
I thank the author group for their responses and revisions. The responses are persuasive, and I have no further major concerns regarding the updated version. I especially appreciate section 6, in which the author provides a much better summary and insights for future work. I believe that this paper contributes significantly to the field of linguistic neural decoding and is directionally a good survey in the field.

Some minor concerns are listed here:

1. The font size within the images should be comparable to the main text font size. For example, some text in Figures 6 and 2c is too small.

2. Some expressions may be controversial, so the author might consider using more cautious expressions. e.g., Section 3.1: The major languages in the world have their textual form, and communication via text outperforms due to its efficiency and concision.

3. The author frequently addresses a moderate, large, or small set of candidates or vocabulary for neural decoding. Are these descriptions comparable, and how many different expressions are there?

4. The author addresses metrics in domains such as machine translation (MT), text-to-speech (TTS), and automatic speech recognition (ASR). Are the metrics different when applied to neural decoding?

5. Some recent research could be considered, I list some suggestions here:

[1] Xinpei Zhao, Jingyuan Sun, Shaonan Wang, Jing Ye, Xiaohan Zhang, and Chengqing Zong. Mapguide: A simple yet effective method to reconstruct continuous language from brain activities. arXiv preprint arXiv:2403.17516, 2024.

[2] Ren Y, Jin R, Zhang T, et al. Do Large Language Models Mirror Cognitive Language Processing?[J]. arXiv preprint arXiv:2402.18023, 2024. COLING 2025.

Reviewer #4

(Remarks to the Author)
Review #2 (Reviewer #4), responses to author.

1. Right, but then “biology” needs to be defined in that particular way, according to the author’s definition. Otherwise, the reader does not know, and the statement seems invalid.
2. I do not understand this response, sorry. I feel like several things mentioned in this response is out of context: what is the 10k sentences and why is a cascade model relevant here? I also think there is a conflicting use of “LLM reasoning”. In parts of the response, you talk about LLMs learning to reason about e.g. text. In the last part of the response you talk about LLMs “reasoning” about neural signals? I don’t think LLMs reason about neural signals, they predict the signals, no?
3. Thanks for the response. However, in the red text, I do not see how you tackle the definition of perception versus language. This should be revised. I think you need to define how you use the terms perception and language in the article?
4. Thank you.
5. I agree with “As shown in Fig. 2c, it has been verified that the brain encoding models and pre-trained LLMs follow the scaling laws, where the model performance increases as the number of parameters grow”, but I do not see how the last part of the sentence about “developing more complex architectures” follows from that?
6. Please see point 3.
7. Thanks for the table – however, I think that beyond a table, it is necessary to understand why a given model is brought up, as a general point to improve the manuscript.

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)
I thank the authors for their revisions. I think the revised manuscript is good. I have no further comments.

Reviewer #4

(Remarks to the Author)
I thank the authors for their replies and changes to the manuscript. On a higher-level note, I think that this review is useful, but the structure of it and the motivation of certain sections is still lacking. Often, there are conflicting statements in different parts of the paper (which has the flavor of LLM logic to me, personally, but of course I cannot say). For instance, I asked the authors to define perception and language and stick with those definitions, but yet, despite definitions, the authors still add text that clearly does not disambiguate the two, for example (there are other cases of this):

"Language serves as the paramount medium facilitating human communication and information exchange. The human brain plays a crucial role in language interaction, enabling complex cognitive processes supporting the perception, comprehension and production of language.
This article aims to provide brain scientists and deep learning researchers with an overarching viewpoint of the significant correlations observed in the human brain during language perception and production from a methodological perspective, and thus facilitate their further investigation."

And also you talk about "image reconstruction" as an example of perception--what does that mean? I believe that we perceive images, but I do not believe we have a high-fidelity reconstruction pipeline in our brains?

Version 3:

Reviewer comments:

Reviewer #4

(Remarks to the Author)
Thanks to the reviewers for replying to my comments. I will leave it up to the authors and the editor to further decide whether any apparent inconsistencies in the manuscript are critical.

Reply to the Reviewers

Re: Manuscript ID COMMSBIO-24-5522
"Linguistic Neural Decoding with Deep Learning: Progress, Challenges and Future"
Authors of the paper
Communications Biology

Dear Editors and Reviewers:

Thank you for your letter and the reviewers' comments concerning our manuscript entitled "Linguistic Neural Decoding with Deep Learning: Progress, Challenges and Future" (ID: COMMSBIO-24-5522). Those comments are all valuable and very helpful for revising and improving our manuscript, as well as the important guiding significance to our research. We have studied comments carefully and have made revisions which we hope meet with approval. It should be noted that neural decoding involving linguistic representation often involves different perspectives, and we have explained our reasons and views if we adhere to our viewpoint. Our response is presented below in blue, with comments and questions from the reviewers in black and quoted reference text from the revised manuscript in red. The content with a strikethrough means that we have deleted the improper statements in the original text, and the reference signs in the quoted text have been omitted to make our response clearer to read. In the revised manuscript, the added content is also presented in red.

Responds to Reviewer #1:

1. Major-1 "Scope": From the abstract and the beginning of the manuscript draft it is difficult to this reviewer to extract whether this is a Neuro-to-AI or AI-to-Neuro paper, such as are they trying to build models to i) simulate the brain or ii) emulate human cognition, arguing that the underlying processing principles are identical or not? 'reviews of neural decoding' is insufficiently precise in this reviewer's eyes.

   We appreciate the Reviewer's feedback regarding the lack of clarity in stating the scope of our manuscript within the abstract and introduction. We apologize for any ambiguity this may have caused and have revised these sections to more clearly define the scope and objectives of our study. Our research is based on the fact that the study of neural decoding involves the theory of neurology and machine learning algorithms, while the disconnect between the two leads to the isolation of previous research. We are committed to unifying tasks with similar goals and methods into a single paradigm, providing a valuable reference for biological researchers and AI programmers to carry out subsequent work. In our initial manuscript, we first introduced some neurological foundations of neural decoding, which mainly involve Section 1, although our introduction is relatively brief due to space limitations. This section answered two questions: firstly, why it is feasible to use artificial neural networks (ANNs) to model the way the brain perceives and responds to external language stimuli and the opposite process - the brain drives vocalization to express with language, and secondly, what kind of ANN structure is reasonable to use. In the following sections, we divide the task form of neural decoding into various paradigms. Regarding the author's question whether our article is Neuro-to-AI or AI-to-Neuro, as far as we know these two phrases are not commonly used expressions. We understand this expression from this perspective: AI-to-Neuro is aimed at neurological tasks, mainly introduces the AI methods used in neurology, and introduces AI strategies in other tasks into neurology; Neuro-to-AI is aimed at AI tasks, draws inspiration from neurology, and thus explores new network architectures and training methods, introducing neurological mechanisms into AI. We consider the former more accurate if this distinction is consistent with the reviewer. From the perspective of a data engineer and AI programmer, a highly detailed neurological foundation is unnecessary (which is more like the goal of neurobiology or brain science). On the contrary, they are more concerned with whether the data association of the modeling has been confirmed, what kind of experimental scheme to use, and whether the structure of the model

is reasonable. As for our goal of building models, it is neither to simulate the brain nor to emulate human cognition. The research on neural encoding is to generate brain activation patterns given external stimuli, which is more consistent with simulating the human brain. We focus on the opposite process, which is to infer external stimuli or inner intention from a determined neural response. Both processes are presented in Fig. 2(b).

This paper summarizes representative solutions and current progress on the mapping from evoked brain response to linguistic stimuli and intention.....It is important to note that the scope of this review encompasses neural decoding processes in which language serves as the medium, specifically in the forms of written text and spoken speech. This study excludes analyses related to handwriting, which involves motor function, as well as visual image reconstruction.

2. Major-2 "Flow and narrative": 2nd paragraph appears to come out of the blue; why is it suddenly emphasizing time series so much?; please improve this transition as well as paragraph-to-paragraph transitions in the paper in general. Further, the concrete examples and cited papers appear to be often poorly motivated. As it stands, the paper strikes as a stream of largely disconnected isolated examples.

   We have re-examined the article and further sorted out our logical flow. For the time series modeling mentioned by the reviewer, we believe that we are not 'emphasizing so much' (at least in terms of length). What we want to emphasize in this section is that under external stimuli, language representations and neural activity patterns in human brain are correlated. Only when this prerequisite is confirmed, can ANN-based neural decoding be established. In addition, such correlation relationships not only exist in the feature dimension but also align in the time series. Our intention was not to emphasize time series, but to point out that two important concepts in neuroscience and machine learning are related - neural prediction and those methods leveraging contextual information. Contextualized embeddings are widely used in the field of natural language processing (NLP) and stand out for their excellent text-understanding ability. The correlation between the two is evidence of the similarities between how the brain and artificial networks understand language, and also suggests that it is beneficial to consider contextual information when building models. In the manuscript, we have organized similar architectures and solutions into distinct categories, typically grouping them within the same paragraph. These categories are introduced in an ascending order of model complexity and effectiveness, often following a chronological progression as well. This structure allows for a coherent and logical flow of information, facilitating a clearer understanding of the evolution and advancement of decoding experiments and models. For a detailed illustration of our organizational logic, please refer to Fig. 1, which visually represents the categorization and progression of the models discussed. This figure serves as a guide to help readers navigate through the various models and understand the rationale behind their arrangement.

   The essence of deep learning lies in leveraging the inherent correlations within data to complete prediction, regression and generation, and the alignment of neural activities and linguistic representations is crucial for enabling these capabilities. Specifically, neural tracking enables the theoretical possibility of achieving temporally continuous decoding from evoked brain activities, while the neural prediction process underscores the benefit of contextual information injunction, which is commonly used in current neural decoding approaches.

   The neural tracking process during verbal communication is shown in Fig.2a. The cortical activity automatically tracks the dynamics of speech as well as various linguistic properties, including surprisal, phonetic sequences, word sequences, and other linguistic representations. A minor time shift has been observed for information transfer and neural response. It ensures the temporal alignment of brain recordings with linguistic representations, facilitating the serialized and temporal modeling of cortical activities. As shown in Fig. 2b, language stimuli are processed into regular evoked brain responses, referred to as cortical encoding, which begins to emerge early in human life. In contrast, linguistic neural decoding aims to reconstruct the stimuli perceived or the intention expressed from high-dimensional brain responses. In natural listening settings, the human brain encodes a wide range of acoustic features and processes external language stimuli temporally through prediction, highlighting the importance of contextual information in cortical perception, even at the level of single neurons. Predictive processing fundamentally forms the comprehension mechanisms in the human brain, occurring hierarchically in

both acoustic and linguistic levels. This phenomenon underscores the profound impact of context on the forecasting and tracking of ongoing speech streams, necessitating the use of contextual representations to investigate cortical responses. This characteristic is similar to language models constructed by neural networks, where the same stimuli presented in varying contexts are mapped onto diverse semantic features.

3. Major-3 "Metrics": ... which are mentioned appear random, as they do not appear to reflect commonly used performance metrics from the NLP community. So based on what basis where the highlighted metrics exactly chosen for treatment in the present work?

   The reviewer mentioned that the evaluation metrics we listed are random and not representative. We strongly oppose that. We would like to know what the reviewer specifically refers to as "commonly used performance metrics from the NLP community". If we have indeed overlooked the more commonly used metrics, we will be glad to supplement and improve our manuscript. Specifically, the vast majority (>99%) of automatic speech recognition (ASR) utilize PER, CER, and WER as evaluation metrics, determined by their decoding granularity, which is for evaluating the differences in text sequences and pursues absolute consistency between decoding hypothesis and annotation sequences. However, in more difficult tasks such as translation, researchers focus on partial consistency and semantic homogeneity, using ROUGE, BLUE, and BERTScore. For speech output tasks, the evaluation system includes objective and subjective metrics, with the former including the frame error (e.g. FFE) and mel cepstral distortion (MCD) and the latter is typically mean opinion score (MOS). In addition, indicators such as PES-Q have not been listed due to their obsolescence, irrationality, and other reasons. We only choose commonly used indicators that can intuitively represent the decoding performance, which is typically the evaluation method used in the work mentioned in the manuscript.

4. Major-4 "Interdisciplinarity": The paper tries to position itself at the Neuro-AI front. Yet, most citations are from neuroscience venues, rather than AI conferences such as NeurIPS and ICML, which raises doubts about the authenticity of the transdisciplinary grasp of this work.

   The reviewer mentioned that we mainly cited articles in neuroscience and lacked AI conference papers. We have classified the citation sources of our papers since 2015 into three categories: neuroscience (e.g. Nature Neuroscience), artificial intelligence (e.g. AAAI, Nature Machine Intelligence), and interdisciplinary (e.g. Nature, Journal of Neural Engineering). The number of papers in the three categories is 32, 36, and 39, respectively. In addition, it is worth mentioning that neural decoding is not the mainstream direction of AI research. Taking INTERSPEECH 2024 as an example, although the conference included the theme of "Speech and Brain", only 6 papers were included. Considering the above factors, we do not believe that we have overlooked conferences in the AI field. We appreciate the suggestions of Reviewer 1 and have added some proper AI papers. We hope that our article can be presented to researchers in multiple fields with a more reasonable citation allocation.

5. Minor-1 "high-parameter ANNs": unusual phrasing in the core technical machine learning literature

   We appreciate the Reviewer's attention to the wording in our manuscript, specifically regarding the phrase "high-parameter ANNs". We apologize for this imprecise expression, which is not standard terminology within the academic community. In response to the Reviewer's feedback, we have revised this language to ensure it is both accurate and consistent with common academic nomenclature. We believe this adjustment improves the clarity and professionalism of our manuscript.

   ...In contrast, linguistic neural decoding aims to reconstruct the stimuli perceived or the intention expressed from high-dimensional brain responses. In natural listening settings,...

6. Minor-2 "In the textual stimuli classification (TSC) paradigm": never heard about this, perhaps the authors could expand more?

   In fact, this is the standardization of neural decoding based on experimental forms in this article. As for textual stimulation classification (TSC), in this paradigm, subjects receive visual textual information (excluding speech like podcasts), and researchers discriminate the category or instance of textual stimuli based on brain responses from various candidates. This is a relatively simple experimental form, considering the small search space and deterministic choices, usually only non-invasive data is

needed without complex model architecture. We provide a more detailed explanation in our revised manuscript, including experiment forms, data formats and decoding targets for each paradigm. This elaboration aims to clarify the methodologies and enhance the overall coherence of our presentation.

Linguistic neural decoding aims to generate the corresponding external stimuli or inner intended speech from the activated brain signals. This field has lacked a fine-grained division, pushing researchers to design experiments, collect recordings, and build models on their own. In this review, previous research has been categorized according to the experiment design, stimuli type and decoding target (Table 1). Stimuli recognition is the simplest form and usually requires a modest candidate set and limited stimuli sequence. For speech stimuli, in addition to identifying the textual content, sometimes researchers consider reconstructing simple speech features or even the speech waveform. These tasks are typically treated as simple classification or regression. Brain recording translation differs in its ability to handle open-vocabulary continuous decoding, which means a sharp increase in the decoding space and results in a deterioration in the accuracy without introducing intrusive signals. This task focuses more on semantic consistency rather than the absolute identicalness of the text. Speech neuroprosthesis aims to generate inner speech from spontaneous neural activation patterns. The subjects do not receive external stimuli but perform pronunciation tasks of imagined speech or attempted speech. Researchers have achieved word-level high-precision continuous decoding with invasive recordings.

7. Minor-3 "The major languages in the world have their textual form, and communication via text outperforms due to its efficiency and concision": ok but how is this related to Neuro and AI? Please sharpen.

   We apologize for any confusion caused to the reviewers. Our intention was to introduce the task format of textual stimulation classification. Our original text is: "The major languages in the world have their textual form, and communication via text outperforms due to its efficiency and concision. Textual stimuli classification...(task description)". Text highly condenses semantic information. An intuitive example is that when we share personal experiences, most people convey information through writing, typing, or speaking, rather than through drawing or performance. So decoding linguistic stimuli and intention has high significance and applicability. We have revised this paragraph to better align with the reader's cognition.

   The major languages in the world have their textual form, and communication via text outperforms due to its efficiency and concision. Language presented in text highly condenses information and avoids the temporal variability of corresponding speech signals. Early work focused on recovering language information from text stimuli. Textual stimuli classification distinguishes the original information provided to the subject within several candidates.

8. Minor-4 "perception methods": unclear to this reviewer

   We apologize for our imprecise language expression. We have re-standardized the wording in the latest manuscript. In this review, perception is regarded as the process in which the subject receives external stimuli and forms corresponding activation patterns in the brain, which is consistent with the experimental setting of textual stimuli classification. We believe that there is a certain complex and high-dimension mapping between external stimuli and brain activities, and the textual stimuli classification task is an important method to understand this process. The difference is that perception is from linguistic representation to neural representation, while neural decoding is the opposite.

   ~~The previous approach included investigating the perception methods of concrete nouns to avoid abstract representations.~~ The previous approach defined a word set of concrete nouns to avoid neural representations of abstract concepts. Classifiers were adopted to distinguish which word had been perceived by the subject.

Responds to Reviewer #2:

We recognize that Reviewer #2's comments are objective and reasonable. We apologize for neglecting to consider outstanding questions in our initial manuscript and focusing on summarizing existing technical routes and achievements. We did subconsciously avoid making claims about future directions, partially because we framed the paper from the perspective of computational engineers, and much of the neuroscience discussion was still unresolved and in debate.

1. "Despite advancing insights into the human brain, the intricacies of linguistic representation and processing still largely resemble a black-box model, making it challenging to explain intuitively the internal mechanisms and patterns of information transfer. Deep learning presents a viable solution for understanding language in the brain by utilizing large-scale trainable parameters to map the correlation between external stimuli and neural activity." After reading this, I assumed that the review would aim to draw parallels between how models encode and decode meaningful linguistic signals and how human brains and cognitive faculties support language. However, this is not what the review set out to do, and so that is not what the review achieves. Instead it provides a sweeping overview of recent technical feats of linguistic neural decoding.

   The scope of this manuscript does not include neural encoding, that is, the work of predicting the neural activation patterns given the external stimuli information. Instead, we provide a comprehensive methodology for obtaining relevant language information from evoked brain recordings. This article does not include in-depth cognitive science, especially the brain's function mechanism, which is limited by the author's research area. We believe that this won't affect the reference value of this review, considering when AI engineers first come into contact with neural decoding, they typically need to be convinced of the strong correlation between the input (neural recording) and output (decoded sequence), which has been explained in Section 1, and available architectures that have been summarized in the rest of the manuscript. For cognitive scientists, the significance of this paper lies in the latest summary of decoding performance and model schemes. We have provided a comprehensive introduction to the task paradigm and model architecture so that they can be used out of the box.

   In this work, we present a taxonomy of recent neural decoding progress, focusing on deep learning architectures and strategies, especially those implementing large language models (LLMs) for their powerful reasoning capacity. We complete with a concise observation of the challenges and potential future directions. This article aims to provide brain scientists and deep learning researchers with an overarching viewpoint of the significant correlations observed in the human brain during language perception and production from a methodological perspective, and thus facilitate their further investigation.

   It is important to note that the scope of this review encompasses neural decoding processes in which language serves as the medium, specifically in the forms of written text and spoken speech. This study excludes analyses related to handwriting, which involves motor function, as well as visual image reconstruction.

2. The authors have done great work organizing the science as they have for this paper, and it is inspiring to review what has been achieved already and imagine what will come next. However, I feel the review falls short of moving the field toward that future by offering commentary on outstanding questions.

   We acknowledge that on outstanding questions, especially those with significant controversy, we avoid discussion and instead state only the definitive correct view. We do not think it is reasonable to present a debatable position in a review article, but we agree that it is necessary to summarize existing limitations and propose possible solutions. We have modified Section 6 to separately list and emphasize the current limitations including ethical issues and potential directions. For more detailed explanations, please refer to this section of the revised manuscript.

3. While the review is comprehensive from a technical perspective, it does not seem to have a thesis other than "bigger models are better". The review did not make any particular claims or resolve any open questions. For example, there is little discussion of why some model architectures are better than others at modeling the brain. The review refers to both the brain and the ANNs used to decode the brain's activity "black boxes" at different points, but (I believe) we need to understand the inner

workings of both before the kinds of BCI communication tools that the authors envision arriving in the future will be possible.

First and foremost, we wish to clarify that this article does not advocate the notion that "bigger models are better." Instead, we emphasize that different network architectures are appropriate for neural recordings of varying quality and decoding granularities. Simply increasing the number of network layers does not necessarily lead to improved performance. Effective model design must also consider factors such as the volume of data available for training and the configuration of the decoding search space. We argue that the choice of model should be guided by the specific characteristics of the neural data and the objectives of the decoding task. This approach ensures that the model architecture aligns with the inherent properties of the neural decoding experiments, leading to more reliable and efficient performance.

Stimuli recognition is the simplest form of neural decoding, involving the differentiation of linguistic stimuli by analyzing the subject's evoked brain responses. For text stimuli reconstruction, decoding is performed at the word or sentence level using classifiers, embedding models, and custom network modules.

In natural listening scenarios, restoring the original speech features and waveform is a more complex task. Regression models (i.e. ridge regression), CNN and RNN-based network modules, and paramount generation models (i.e. GAN) are widely used.

A feasible translation model architecture, including feature extraction, feature transformation and a pre-trained encoder-decoder to generate the decoding sentence. Both the pre-trained language models (i.e. BART) and speech models (i.e. Whisper) have been verified effective.

The cascade speech neuroprosthesis replaces the acoustic model with a brain model and decodes the corresponding phoneme or small-vocabulary word hypothesis before generating the sentences. These works typically adopted the Gaussian mixture model (GMM) to fit the data distribution of invasive brain activities. Such approaches did not make a groundbreaking impact until the replacement from GMM to ANNs contributed to a steady improvement.

In our manuscript, we did not elaborate on how certain architectures exhibit superior performance in simulating brain activity, as this topic extends beyond the primary scope of our current work. We acknowledge that significant research efforts are dedicated to exploring the connections between neural networks and brain representations, aiming to achieve closer hidden representations given the same language stimuli. This area of investigation is closely related to neural encoding and involves understanding how artificial models can mimic the brain's processing of linguistic information. Given the focused objectives of this article, we have limited our discussion to decoding paradigms and methodologies directly relevant to our study. While the relationship between neural network architectures and brain representations is an important and active area of research, it was not the central focus here. Therefore, we either omitted or only briefly mentioned these aspects to maintain clarity and coherence within the defined scope of our manuscript.

It is important to note that the scope of this review encompasses neural decoding processes in which language serves as the medium, specifically in the forms of written text and spoken speech. This study excludes analyses related to handwriting, which involves motor function, as well as visual image reconstruction.

4. I agree with the authors that these models and improvements to brain computer interfaces can do a lot of good in the world, and it is exciting to read about the progress that is being made. The being said, other than providing a curated selection of exciting technical advances, I do not feel that this review offers insight into the way ahead.

   We apologize for the initial lack of insight into future directions in our manuscript. In response to the Reviewer's comments, we have reorganized the last section to highlight the critical methods and challenges that must be addressed to develop universal, efficient, and cost-effective BCI systems. Our revised discussion now emphasizes the significance of neural universal representations, which are essential for creating models that can generalize across different individuals and tasks. We also underscore the current need for extensive consideration of signal denoising and augmentation techniques to improve

data quality and model robustness. These advancements are crucial for enhancing the reliability and performance of BCIs. In addition to the neural decoding solutions, we also address the standardization of the data collection process to ensure the feasibility of a large-scale neural corpus. Furthermore, we propose that ethical constraints in neural data collection, training, and decoding must be improved. Ensuring privacy, informed consent, and equitable access are paramount as BCI technology advances. Addressing these ethical considerations will foster trust and responsible innovation in the field. For more detailed explanations, please refer to Section 6 of the revised manuscript.

Even though we are still a long way from efficient and harmless BCI, some directions have shown bright prospects. A unified brain representation could be the next big breakthrough in neural decoding, which has made a great impact on other modality processing. There is a consensus on the individual and temporal variability of neural signals. Performing individualized data collection and model training is not a feasible solution considering the corresponding recording duration and computational resources. The implementation of a unified neural representation is a promising solution by fine-tuning with limited user-specific recordings to form personalized decoding systems. This requires expanding the previous experiment to a group population and collecting much more dynamic neural recordings, with self-supervised learning providing a strong relationship with semantic information.

.....

As for privacy preservation and technology regulation, strict management and supervision need to cover the entire process of data collection, model training and deployment application [**?**]. The premise is to form clear data usage standards, minimize the dimensions and duration of neural recordings while ensuring decoding performance, and strictly discard potential privacy-aware instances. The dissemination and use of neural data need to ensure that the goal of the corresponding experiment is for human welfare, and data encryption, differential privacy and federated learning are protection measures that need to be considered. As the modality and experimental population of neural decoding expand, we strongly call for the formation of a unified ethical perspective such as human rights guidelines, which requires neural computing companies and related major researchers to assume corresponding scientific responsibilities.

5. There are points in the review where I felt the authors oversimplified or were inattentive to what is most interesting about a connection being drawn between models and the brain. For example, the authors write that "the computations of ANNs account for a significant portion of the variance observed in brain activity", but its not clear which "computations" that the ANN is doing that are relevant to capturing variance in neural activity. Is this just saying that model similarity structure defined among hidden-layer activation patterns correlates with neural similarity structure? Or, is a point being made that Transformers are specifically doing something brain-like? Given that the LLM and the brain are each learning about linguistic structure and word meaning, finding that they both encode similar structures is not inherently interesting—there is plenty of evidence that the concepts learned by LLMs are not quite what humans learn.

   We would like to limit the scope of the manuscript to an overview of language-centered neural decoding, especially the methodology. The functional similarities between artificial neural networks and the brain are such a complex topic that we think it is difficult to explain in detail in the rest of the paper. The summary and comparison within this topic would have been more appropriate as a separate theoretical review, so we ultimately decided not to cover this information extensively. Regarding the similarities between the brain and the model, we do not think that such similarities are reflected mainly in the model structure. After all, the neural network that simulates the transmission of neural signals (spiking neural networks, etc.) has not yet achieved encouraging performance and only performs on primary perception and classification tasks [13, 14]. On the other hand, ANNs and the neural processing in the brain probably share certain encoding similarities mainly in semantics and partly in syntactics [3, 5, 8].

6. The brain recording translation models are interesting, but are presented in a way that feels off. Because of the LLM sitting between the brain features and the output, it seems that part of the prediction ability for subsequent words will be carried by the information the LLM encodes about high-probability sequences of words. In the example where there true text was "At best they tales ..." and the prediction

was “At best it is macred...”, it seems that “it is” likely emerged from that expectations of the LLM rather than the neural signal. I am surprised that the LLM generated “macred”—that is not an English word as far as I know. I am not sure how this model could produce non-words.

This decoding example comes from [17]. The base model is Whisper, but both Whisper and the language model can generate non-words. This is related to the modeling of words - the most basic unit is not a word, but a sub-word unit, or a word piece. The model learns common letter combinations from a massive corpus and forms corresponding sub-words. Each sub-word corresponds to a unique label for model training. This approach is also adopted by large language models. For the non-word “macred”, the model is likely to output two consecutive units - “ma_” and “cred”. The underscore is used here to indicate that the sub-word unit is not the end of a word so that the transformation from sub-word units to words can be achieved. This example comes from the self-supervised speech model, Whisper, also indirectly illustrates that the sequence obtained by this neural decoding does not completely depend on high-probability sequences of words, for the connection probability of “ma_” and “cred” from the model itself should be negligible. We believe that the emergence of this decoding sequence is most likely due to the fact that there is a mismatch between the speech encoded distribution and the neural signal distribution in the embedding space of fine-tuned Whisper, and this neural recording happens to fall at this location, resulting in an output segment with a relatively high probability in the neural dimension but a very low probability in the speech dimension. The causes of this mismatch may be various. One important reason is that the noise of non-invasive recordings causes a large offset between the sampled and the real neural signal, which may result in the sampled signal falling in some strange positions. Language models usually provide a set of probabilities to guide the model to output a sequence that is more like a sentence (consider the difference between “a lit puppy” and “a little puppy”, both are similar in pronunciation and spelling, but the latter has a higher language probability). Language model probability has an impact on various tasks such as speech recognition, neural neuroprosthesis, and large language model generation, but this is generally not the focus of research. First, we are more concerned about decoding accurate content, and second, it has been shown in various tasks that the prior probability of the language model is only auxiliary, not dominant.

7. Anyway, it is not clear how much of the brain recording translation performance is driven by identifying and leveraging neural representations that truly encode the lexical content being “decoded”, because the decoding pipeline seems to allow the LLM itself to infer words and sentence structure. This matters from a biological and psychological perspective–wanting to use these models to understand the brain and mind–than it does from the perspective of someone wanting to build a tool to allow effective communication for someone who has lost the ability to speak and/or write.

   We have provided some proof in the answer to the previous question. It is quite difficult to abandon LLMs, considering that their role is not only to provide a probability distribution that is more consistent with language rules, but also to provide text generation capabilities. Another point that can prove our statement is from the perspective of the autoregressive generation of LLM. Most LLMs use autoregressive methods to generate sequences, that is, the currently generated unit depends on the internal parameters of the model and the output of the previous time step. For brain recording translation, some work uses the teacher forcing trick, that is, when decoding the unit at the current moment, the previous ground-truth unit is given, which may cause the LLM to have a distribution of subsequent units and produce high-probability sequences. However, for work without teacher forcing, the probability distribution provided by LLM is based on the output of the previous moment, which is likely to be an incorrect token in this work. If the inferring ability of LLM plays a dominant role, this will lead to the gradual accumulation of errors, that is, the output sequence gradually deviates from the transcription. Obviously, this phenomenon has not been observed under the current experimental settings.

8. There are many models in the paper that integrate neural information over time. I will note that the cognitive neuroscience of semantic knowledge currently has little to say about how conceptual knowledge is encoded in time-varying neural signal. It is largely unknown–there are many reasonable ways to implement a model that uses time-varying signal when attempting to decode semantic knowledge, but that implementation makes important assumptions about how the brain and mind work. When

thinking about linguistic decoding from biological and psychological perspectives, these details of model implementation are more interesting than the model performance–or rather, the model performance is difficult to interpret without that context.

We thank the reviewer for the cognitive neuroscience information. We believe that the semantic information and time-varying decoding mentioned in our manuscript do not conflict with the neurological basis mentioned by the reviewer. The semantic we want to express is defined in the field of natural language processing and signal processing, which is usually opposed to acoustic. The former contains rich linguistic information, while the latter focuses more on wave changes and flow. As for time-varying decoding, we stress as the subject receives stimulation or tries to speak over time. We understand that this process may involve more complex brain activities (for example, the word puppy may produce vision images or semantic information such as yellow fur), but the most intuitive semantic sequence (that is, the text or speech sequence with accurate ground-truth) is correlated with neural activity, which is sufficient to provide theoretical support for building deep learning models, minimizing loss, and updating models through gradient backpropagation.

9. The Future Trends section seems motivated by the idea that technical advances, more compute power, and bigger models are all that is needed to unlock the contents of the mind. For example, the authors write: "Progress in decoding complex neural signals has shown the potential to reconstruct thoughts, images, or even dreams. This could have profound implications for understanding consciousness and developing communication methods." Decoding words within fixed vocabularies or generating text/speech from a limited space of neural signals corresponding with articulation is not analogous to freely decoding thought and mental imagery or providing readouts of dreams. The leap to discussing consciousness is nonsense without defining the concept. It is a stretch to suggest that models of the kind discussed in the review are bringing us closer to understanding the nature or mechanics of the brain and mind, given that the review treated the brain as black box.

   We understand the reviewer's concerns and admit that our explanation is not rigorous in cognitive science. Our original intention was to show that decoding brain signals for multiple modalities of human perception (such as in dreams) has gradually been proven feasible. This includes text, speech, images and videos [2, 10, 12, 15, 17]. It is indeed unreasonable to explain abstract concepts such as consciousness and dreams without clear definitions, especially considering these concepts might be controversial in the field of neuroscience. We have expanded this section and reformulated our wording and hope the latest manuscript will meet the reviewer's requirements. In addition, we have reorganized the Future Trends section to better summarize the development direction of language-centered BCI systems with neural decoding approaches.

10. The calls for future directions are general and not particularly insightful. Yes, better non-invasive neuroimaging will of course improve the ability to decode language from neural activity, as will better BCI technology. But what most needs to be improved? Spatial resolution, temporal resolution, signal to noise, alignment across individuals? Likewise, a call for "universal decoding schemas" to address the fact that neural signals and the concepts they correspond to are non-stationary over time restates the problem but does not point in the direction of a solution.

    We understand the authors' concerns, but it is often difficult to determine which areas of improvement are most needed for a complex problem - the issues we listed are all important factors limiting neural decoding, and addressing any one of them without the progress of others is almost impossible. We have reorganized the content of this section and summarized the potential development directions as unified representation, front-end data processing, model structure (LLMs), advanced equipment and standard data collection processes, and finally ethical discussions. In addition, we proposed specific trial methods for current challenges. Taking the subject invariant properties of neural signals, important trial directions include training a unified representation in a self-supervised manner and fine-tuning it through a small amount of personalized data. For non-invasive neural signal processing and augmentation, possible solutions include generative methods like GAN and diffusion. The slice of the revised manuscript is listed below as a reference for the comments above.

    Even though we are still a long way from efficient and harmless BCI, some directions have shown bright prospects. A unified brain representation could be the next big breakthrough in neural decoding, which

has made a great impact on other modality processing. There is a consensus on the individual and temporal variability of neural signals. Performing individualized data collection and model training is not a feasible solution considering the corresponding recording duration and computational resources. The implementation of a unified neural representation is a promising solution by fine-tuning with limited user-specific recordings to form personalized decoding systems. This requires expanding the previous experiment to a group population and collecting much more dynamic neural recordings, with self-supervised learning providing a strong relationship with semantic information.

While invasive neural decoding has demonstrated superior performance, the main limitation of non-invasive signals is their significant noise level. It is worth practicing to perform data augmentation and denoising on neural signals, and existing solutions are mainly based on generative models, such as GAN and diffusion. Research on the robustness of model architectures is still in its early stages, especially since LLM has recently demonstrated extraordinary reasoning performance. Considering the significant mismatch between the tokenized text and neural space, robust neural networks with a stable training strategy are possible to boost the generation performance.

Large language models preserve powerful understanding, reasoning, and generation capabilities, and previous studies have shown that LLMs trained on vast amounts of textual corpus enable the ability to be aligned with other modalities through smaller-scale fine-tuning, thereby generating content with strong semantic consistency and vivid details. The same phenomenon also applies to neural data, where the most significant trend in linguistic neural decoding has been implementing a textual LLM as the backend decoder for text generation. As shown in Fig. 6, in the initial attempts, the LLMs were adopted to generate hypothesis candidates with a separate module score for each potential sentence. A more promising approach treats the LLM as the inferring core to generate correlated textual information, and gradually evolves into a unified decoding system with multi-modality inputs and user-specified output. We believe that the update and iteration of LLMs will promote qualitative changes in neural decoding, thereby achieving application levels in the near future.

Parallel to model improvement, in neuroscience, a pressing issue is the precise collection of neural recordings related to language processing, including acoustic and phonological aspects. This requires identifying specific neuronal populations and brain areas involved in language functions. High-resolution scanners, wearable neurotechnology devices and advanced equipment are also necessary, and more reasonable experimental settings need to be explored. An important aspect is to unify the data collection framework to explore the possibility of developing a massive neural corpus from multiple resources, which means diverse stimuli and subject conditions. The neural recordings from a single experiment trial are typically suitable for small network training while pre-training and fine-tuning on a larger scale is likely to process data spanning several orders of magnitude.

As for privacy preservation and technology regulation, strict management and supervision need to cover the entire process of data collection, model training and deployment application. The premise is to form clear data usage standards, minimize the dimensions and duration of neural recordings while ensuring decoding performance, and strictly discard potential privacy-aware instances. The dissemination and use of neural data need to ensure that the goal of the corresponding experiment is for human welfare, and data encryption, differential privacy and federated learning are protection measures that need to be considered. As the modality and experimental population of neural decoding expand, we strongly call for the formation of a unified ethical perspective such as human rights guidelines, which requires neural computing companies and related major researchers to assume corresponding scientific responsibilities.

Responds to Reviewer #3:

1. Major-1: The author might need to provide a clearer definition of what is linguistics decoding. The paper should explain why decoding methods like imaged handwriting [1], motor imagery controlling methods, and SSVEP-based spelling methods [2] are not involved in this survey (or just involved in additional sections). However, motor-based speech decoding is involved. In my opinion, both methods are linguistic decoding from external human behaviors, i.e., conveying semantic content through speaking, writing, or typing.

   The scope of our manuscript mainly covers neural decoding strongly aligned with language, especially human speech, while avoiding the confounding of motion, even though we also mention experiments of attempted speaking involving the activity of the articulators. If we put handwriting and cursor control in a separate chapter, there are much more achievements. As far as I know, it is now possible to play Civilization VI. We redefine the scope of linguistic neural decoding mentioned in this paper in the hope of resolving this ambiguity. We incorporated motor-based speech decoding into our study partially due to the challenges in independently elucidating inner speech decoding and attempted speech decoding. Given that the experimental setups and model architectures are largely identical, the sole distinction lies in whether the subject makes a sound. In the context of attempted speech recordings, it's important to note that participants do not retain the capability for intelligible speech production; consequently, the sounds they produce during these attempts are typically indiscernible. Another consideration lies in the correlation between motor-based speech and phonemes, the simplest speech unit during vocalization. It ensures the limited decoding space while writing does not. The reason for not considering typing is that we believe it is more of a classification task without language-specific features such as semantics and morphology.

   It is important to note that the scope of this review encompasses neural decoding processes in which language serves as the medium, specifically in the forms of written text and spoken speech. This study excludes analyses related to handwriting, which involves motor function, as well as visual image reconstruction.

2. Major-2: The purpose of Section 2 is somewhat unclear. Is it solely to support the following three chapters? If so, what should be the metrics for alignment?

   Our initial purpose was to briefly demonstrate that decoding language information from neural activity is reasonable, which may require explanation for most computational engineers who have a limited understanding of brain science. It is like when we summarize image recognition, we may first explain that image patches contain object features, while introducing that sound pitches contain speaker features and language information if we explain speaker and speech recognition routes. As for the alignment metrics, it depend on the experimental setting, where the correlation between neural activity and language can explain the effect of alignment, while the accuracy of the decoded language information can also confirm the alignment. This section primarily addresses two phenomena: neural tracking and neural prediction. Neural tracking refers to the process by which, following the reception of language stimuli, the brain encodes corresponding semantic and acoustic information in a sequential temporal pattern. This phenomenon provides crucial theoretical underpinnings for neural decoding efforts. On the other hand, neural prediction underscores the importance of contextual information, a concept that is extensively leveraged in the deep network modeling of language processing.

3. Major-3: As a survey paper, it is important to help people understand the current state and ability of linguistic decoding. Providing more examples and explanations of decoding results could be beneficial, especially to readers who are not specialists in the field. We need to avoid exaggerations and show the true existing ability of semantic decoding.

   We understand the reviewer's concerns - non-intuitive experimental results are difficult for non-expert readers to understand, especially the machine translation indicators we specified (e.g. BLUE, ROUGE and BERTScore). We ultimately decided not to add a table to show the decoding examples, but strongly recommend readers to refer to the corresponding papers and code repositories. The primary reason for this decision is that the examples presented in existing works are often not representative. They typically include both successful cases and deviations from accurate semantic decoding

results, which complicates the intuitive presentation of performance through examples. In addition, the sentence-level stimuli corresponding to the experiments have various levels of difficulty. The summary of decoding performance we show usually spans different levels. For instance, a neural decoding model with 75% WER does not represent the general result of decoding sentences but rather reflects a compromise between simpler sentences with a WER of 50% and more challenging sentences with a WER exceeding 100%. To accurately demonstrate decoding performance, a broader range of examples would be necessary, which is impractical to provide within a single article.

4. Major-4: The author can consider including more tables that clearly outline the settings and decoding performance of existing work. The settings could involve BCI devices, types and granularities of decoding, type of language, etc. Table 2 is a good example in the translation section, and similar tables could be provided for other sections.

   To address the Reviewer's suggestion regarding the inclusion of additional tables to clearly delineate the settings and decoding performance of existing work, including BCI devices, types and granularity of decoding, language types, etc., we have considered expanding the content. However, given constraints such as paper length and our focus on presenting a detailed methodology, we have opted for a balanced approach. In the revised manuscript, we have introduced new tables that elucidate the research division within linguistic neural decoding (see Table 1). Regarding BCI devices, we have decided not to incorporate this aspect into our manuscript. Including BCI-related information would necessitate discussions on signal acquisition principles, device parameters, and other specifics that diverge from our primary objective of detailing the methodological aspects of neural decoding.

5. Major-5: The distinctions between these four types of linguistic tasks could be illustrated using a diagram or a table and provide more explanation.

   We have incorporated a new table and provided succinct explanations to highlight the distinctions and interrelations between different tasks (see Table 1). Considering the present absence of standardized frameworks in linguistic neural decoding and the innovative nature of our categorization, we deem these improvements necessary to advance the field. Detailed instructions in the following sections are carefully structured to correspond one-to-one with the table entries, thereby enhancing the clarity and comprehensibility of our methodology.

   In this review, previous research has been categorized according to the experiment design, stimuli type and decoding target (Table 1). Stimuli recognition is the simplest form and usually requires a modest candidate set and limited stimuli sequence. For speech stimuli, in addition to identifying the textual content, sometimes researchers consider reconstructing simple speech features or even the speech waveform. These tasks are typically treated as simple classification or regression. Brain recording translation differs in its ability to handle open-vocabulary continuous decoding, which means a sharp increase in the decoding space and results in a deterioration in the accuracy without introducing intrusive signals. This task focuses more on semantic consistency rather than the absolute identicalness of the text. Speech neuroprosthesis aims to generate inner speech from spontaneous neural activation patterns. The subjects do not receive external stimuli but perform pronunciation tasks of imagined speech or attempted speech. Researchers have achieved word-level high-precision continuous decoding with invasive recordings.

6. Major-6: Additional insights could be provided to help discoveries and research in related fields. The insights currently mentioned by the author are mostly related to brain decoding, such as lacking of consuming non-invasive, high-quality data, and being subject- and time-invariant. These issues pertain to brain data acquisition and brain signals but are not specifically related to linguistic decoding. In fact, these limitations could apply to any article on brain signal decoding. For example, can the author summarize the existing deep learning architectures specially designed for linguistic decoding and its relevant challenges?

   We have restructured Section 6 to incorporate more insightful recommendations and practical solutions. Specifically, regarding deep learning algorithms, we propose: 1)Establishing a unified neural representation through self-supervised learning to enhance the generalizability of distinct subjects. 2)Implementing robust model architectures along with neural signal denoising and augmentation at

the front end to improve data quality and model performance. 3) Leveraging Large Language Models (LLMs) for their advanced reasoning and inference capabilities to enrich the decoding process. These enhancements aim to address current challenges and promote advancements in the field.

A unified brain representation could be the next big breakthrough in neural decoding, which has made a great impact on other modality processing. There is a consensus on the individual and temporal variability of neural signals. Performing individualized data collection and model training is not a feasible solution considering the corresponding recording duration and computational resources. The implementation of a unified neural representation is a promising solution by fine-tuning with limited user-specific recordings to form personalized decoding systems. This requires expanding the previous experiment to a group population and collecting much more dynamic neural recordings, with self-supervised learning providing a strong relationship with semantic information.

While invasive neural decoding has demonstrated superior performance, the main limitation of non-invasive signals is their significant noise level. It is worth practicing to perform data augmentation and denoising on neural signals, and existing solutions are mainly based on generative models, such as GAN and diffusion. Research on the robustness of model architectures is still in its early stages, especially since LLM has recently demonstrated extraordinary reasoning performance. Considering the significant mismatch between the tokenized text and neural space, robust neural networks with a stable training strategy are possible to boost the generation performance.

Large language models preserve powerful understanding, reasoning, and generation capabilities, and previous studies have shown that LLMs trained on vast amounts of textual corpus enable the ability to be aligned with other modalities through smaller-scale fine-tuning, thereby generating content with strong semantic consistency and vivid details. The same phenomenon also applies to neural data, where the most significant trend in linguistic neural decoding has been implementing a textual LLM as the backend decoder for text generation. As shown in Fig. 6, in the initial attempts, the LLMs were adopted to generate hypothesis candidates with a separate module score for each potential sentence. A more promising approach treats the LLM as the inferring core to generate correlated textual information, and gradually evolves into a unified decoding system with multi-modality inputs and user-specified output. We believe that the update and iteration of LLMs will promote qualitative changes in neural decoding, thereby achieving application levels in the near future.

7. Major-7: Although the author discusses many aspects related to large models and semantic decoding, it is worth exploring the significance of large models beyond serving as representation and providing language knowledge to help context-based translation. Especially compare LLM-related linguistic research to previous neurolinguistics research, which includes how humans reason based on language, their visual and auditory understanding of language, the different granularity of linguistic understanding from lexeme to sentence, and temporal analyses of linguistic understanding.

   We appreciate the reviewer's comments on comparing the understanding, reasoning, and generation process of language and neural signals by large models with the multi-granularity processing of language by humans. It is a huge topic and difficult to explain in a limited space. The scope of our paper mainly involves neural decoding, but not the internal understanding mechanism, especially how language information is processed after entering the brain and large language models, the similarities in semantics and syntatics, etc. We are glad to dive as a follow-up research topic, but for the completeness and independence of this manuscript, we decided to limit it to methodology.

8. Minor-1: The author may need to explain what is teacher forcing addressed in Section 4 (refer to https://github.com/NeuSpeech/EEG-To-Text).

   We explained this trick in our revised manuscript. This is a common strategy in speech recognition tasks. During autoregressive decoding, instead of feeding the model's own predictions back into it as inputs during training, the teacher forcing strategy provides the actual target annotation. This method helps the model learn faster and more effectively. Since the annotation is not available during inference, it should be set to the previous decoding result instead.

   However, their models were highly estimated with teacher-forcing schema during evaluation, which means instead of feeding the model's own previous predictions as inputs for the next time step, the

actual target values were used during inferring. This prevented them from generating meaningful sentences in real-life applications.

9. Minor-2: There are some typos that need to be corrected, such as the quotation marks and missing commas in the caption of table 1.

   We thank the reviewer for pointing out your typos, and we have corrected them in the revised manuscript.

10. Minor-3: The author should provide more discussions on the ethical issue related to linguistic decoding. Although I agree that relevant techniques are in an early stage, it will emerge serious ethic problems if we can read people's minds.

    We acknowledge the importance of ethical considerations in the development of linguistic decoding technologies and agree that this aspect deserves more thorough discussion. The discussions on the ethical issues are listed below.

    Privacy preservation: Ethical debates regarding collecting and decoding neural signals from the human brain remain an important limiting factor. Invasive data collection is only carried out on a small population due to its surgical risk and usually requires ensuring the necessity of craniotomy for medical treatment. Non-invasive data has much more promotional potential but also carries significant risks of privacy leakage - a more comprehensive data usage convention may be necessary, including the standardization of the collection process, the requirements of experiment subjects, the decoding granularity and vocabulary, and necessary solutions to avoid violation of personal privacy. To have widespread potential, BCI systems must be privacy-preserving and ethically sound. The system should be clearly aware of what information can be accessed, displayed, or made public, and choose to conceal or ignore when it comes to personal privacy or inner thoughts. A responsible stance within society that firmly opposes the misuse of neurodata would serve as the ethical guide for the future advancement of neurotechnology.

    As for privacy preservation and technology regulation, strict management and supervision need to cover the entire process of data collection, model training and deployment application. The premise is to form clear data usage standards, minimize the dimensions and duration of neural recordings while ensuring decoding performance, and strictly discard potential privacy-aware instances. The dissemination and use of neural data need to ensure that the goal of the corresponding experiment is for human welfare, and data encryption, differential privacy and federated learning are protection measures that need to be considered. As the modality and experimental population of neural decoding expand, we strongly call for the formation of a unified ethical perspective such as human rights guidelines, which requires neural computing companies and related major researchers to assume corresponding scientific responsibilities.

11. Minor-4: The author has used many illustrations from previous articles; please ensure that these figures' permission has been obtained.

    The images included in our manuscript consist primarily of icons and components that are licensed for non-commercial use. For explanatory purposes, we have also utilized feature maps derived from our own experiments. When incorporating experimental results from other studies into our figures, we have ensured to properly cite the original papers. We are committed to adhering to copyright regulations and ethical standards. Should any issues arise concerning the use of these images, we will promptly make the necessary corrections and adjustments to ensure compliance.

12. Minor-5: The author can consider more relevant papers. I am listing some suggestions here, but the author does not need to adopt all of them.

    We are grateful to the reviewers for their valuable suggestions, which have helped us enhance our manuscript. Most of the relevant papers fit our In response to their feedback, we have added appropriate references to extend and better explain our work in the revised manuscript.

Responds to Reviewer #4:

1. "Linguistic neural decoding reconstructs the stimuli information perceived by text or speech, which has shown groundbreaking achievements, thanks to the improvement of neuroimaging, biology and artificial intelligence." How has "biology" improved and how can we thank biology for making groundbreaking achievements? Biology has arguably not changed in the past few years and has not enabled advancements in neural decoding?

   We think the development of biology has indeed promoted neural decoding. This perspective is closely tied to our definition of biology, which encompasses fields such as brain science and neuroimaging—domains that fall under the broader umbrella of biological sciences. Specifically, high-resolution neuroimaging techniques and more precise signal collection methods have played a crucial role in advancing neural decoding capabilities.

2. "In this work, we present a taxonomy of recent neural decoding progress, focusing on deep learning architectures and strategies, especially those implementing large language models (LLMs) for their powerful reasoning capacity." I believe that most studies do not leverage "LLM reasoning", namely the embeddings. It is true that LLM reasoning can be leveraged, but I do not think that is the main focus of the field.

   For the role and potential of LLMs in neural decoding, we hold conflicting views. The capabilities of LLM reasoning have been verified in the past few years. It was first demonstrated in text understanding and generation, and it should take time to expand in other areas. Among the work we have summarized, most of the recently published papers have utilized LLMs, and they have shown superior performance in non-invasive continuous decoding, which may be an ideal paradigm for future brain-computer interfaces. In our revised manuscript, we additionally include experiments in the past few months, majorly utilizing an LLM backbone. Even for invasive neural decoding, we noticed that fine-tuning with about 10k sentences could achieve comparable results to the cascade model [4, 16]. Considering the complexity of the LLM backbone, further data expansion is likely to bring more stable performance improvements. In addition, this also shows that LLMs trained on a vast amount of text preserve the ability to understand and reason about neural signals, which is an interesting finding.

3. "This article will help brain scientists and deep learning researchers gain a bird's eye view of language perception, and thus facilitate their further investigation." Why is language denoted as "perception"? The first sentence in the introduction says that language is exclusive to "higher-order" organisms? Language—as opposed to speech perception—is not tied to perception given that certain brain regions are active regardless of the perceptual modality (e.g., Deniz et al 2017).

   We apologize for any confusion caused by our original wording. To clarify, by "language perception", we refer to the process by which the human brain receives and interprets external language stimuli. Achieving high-precision neural decoding would indirectly demonstrate our understanding of the brain activities elicited by language stimuli, thereby enhancing our comprehension of the language perception process. In response to feedback from multiple reviewers, we have revised this expression in the revised manuscript to ensure clarity and precision. Additionally, we acknowledge the controversy surrounding the attribution of language capabilities to advanced organisms alone. Considering that various species have been confirmed to transmit information through vocalization [7, 11], we have modified this statement to reflect a more inclusive and accurate representation of communication across different species.

   Language serves as the paramount medium facilitating human communication and information exchange. The human brain plays a crucial role in language interaction, enabling complex cognitive processes supporting the perception, comprehension and production of language.

   This article aims to provide brain scientists and deep learning researchers with an overarching viewpoint of the significant correlations observed in the human brain during language perception and production from a methodological perspective, and thus facilitate their further investigation.

   Language serves as the paramount medium facilitating human communication and information exchange. The human brain plays a crucial role in language interaction, enabling complex cognitive processes supporting the perception, comprehension and production of language.

4. "When processing natural language, ANNs exhibit patterns of functional specialization similar to those of cortical language networks". Which studies have shown this? Functional specialization is a strong claim (see e.g., Kanwisher, 2010).

   When processing language information, both neural networks and the human brain exhibit certain similarities. Both possess neural units that are sensitive to specific signals or functions. Research in this area has evolved from examining relatively simple network structures to exploring LLMs [1, 6, 9]. Despite limitations imposed by differences in network architectures and training resources, there is evidence of functional specialization within these models. While the specific neural networks responsible for particular functions may vary across different models due to structural inconsistencies, functional and causal evidence supports the specialized processing.

5. "it has been verified that the brain encoding models and pre-trained LLMs follow the scaling laws, where the model performance increases as the number of parameters grows, indicating the necessity to develop more complex architecture to bridge the brain activity patterns and human linguistic representations". I don't think it logically follows that one needs to build novel architectures? If scaling works, then one might argue that it is enough to make the models bigger? Of course, novel architectures would be helpful, but I don't think it follows from the references provided in these sentences.

   We understand the reviewer's concern about whether it is possible to conclude from the research we cited that simply expanding the model is enough to achieve stable performance improvements. We stated a more complex architecture is a necessary factor for neural decoding to achieve considerable performance in the manuscript, but not the only factor. The effect of this task is also closely related to the amount and quality of data, etc. In the article we cited, the original expression is "increasing scale in both models and data will yield incredibly effective models of language processing in the brain, enabling better scientific understanding as well as applications such as decoding", and we emphasized the scale in models. We believe that it is reasonable to conclude that it is necessary to increase the complexity of the models, but the argument that "it is enough to make the models bigger" is a misunderstanding of the content of our manuscript.

   As shown in Fig. 2c, it has been verified that the brain encoding models and pre-trained LLMs follow the scaling laws, where the model performance increases as the number of parameters grows, indicating the necessity to develop more complex architecture to bridge the brain activity patterns and human linguistic representations.

6. The authors should define what they mean by "language". Does it entail speech/visual perception? Or do they talk about language as a "higher-level", amodal process (e.g., Fedorenko et al. 2024)?

   Our manuscript only includes neural decoding related to language, and therefore excludes visual reconstruction and generation of body movements. Language can be presented in various forms, including receiving text and speech stimuli, as well as spontaneously imagining text or trying to speak. Since handwriting is more related to motion, we do not include this part. We have explicitly explained our scope in the revised manuscript.

   It is important to note that the scope of this review encompasses neural decoding processes in which language serves as the medium, specifically in the forms of written text and spoken speech. This study excludes analyses related to handwriting, which involves motor function, as well as visual image reconstruction.

7. Many model types are mentioned. Oftentimes it is not mentioned what those are, e.g., VQ-VAE (page 11) – it is unclear why that is of relevance. If the authors chose to mention these model types, it needs to be clear why.

   We agree with the comments that the exclusion of the details of the model architecture mentioned in the review would interfere with the reader's understanding. We have added a brief introduction in the appendix to summarize the model structure. We believe that such changes are reasonable and will make the paper more reader-friendly.

References

[1] Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. The llm language network: A neuroscientific approach for identifying causally task-relevant units. arXiv preprint arXiv:2411.02280, 2024.

[2] Miguel Angrick, Shiyu Luo, Qinwan Rabbani, Daniel N Candrea, Samyak Shah, Griffin W Milsap, William S Anderson, Chad R Gordon, Kathryn R Rosenblatt, Lora Clawson, et al. Online speech synthesis using a chronically implanted brain–computer interface in an individual with als. Scientific reports, 14(1):9617, 2024.

[3] Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Bridging the semantic latent space between brain and machine: Similarity is all you need. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 11302–11310, 2024.

[4] Sheng Feng, Heyang Liu, Yu Wang, and Yanfeng Wang. Towards an end-to-end framework for invasive brain signal decoding with large language models. In Interspeech 2024, pages 1495–1499, 2024.

[5] Abraham Jacob Fresen, Rochelle Choenni, Micha Heilbron, Willem Zuidema, and Marianne de Heer Kloots. Language models that accurately represent syntactic structure exhibit higher representational similarity to brain activity. In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 46, 2024.

[6] Jiři Hammer, Robin Tibor Schirrmeister, K Hartmann, Petr Marusic, Andreas Schulze-Bonhage, and Tonio Ball. Interpretable functional specialization emerges in deep convolutional networks trained on brain signals. Journal of neural engineering, 19(3):036006, 2022.

[7] Bubacarr Jobarteh, Madalina Mincu, Gavojdian Dinu, and Suresh Neethirajan. Multi modal information fusion of acoustic and linguistic data for decoding dairy cow vocalizations in animal welfare assessment. arXiv preprint arXiv:2411.00477, 2024.

[8] Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas, and Evelina Fedorenko. Lexical-semantic content, not syntactic structure, is the main contributor to ann-brain similarity of fmri responses in the language network. Neurobiology of Language, 5(1):7–42, 2024.

[9] Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. Shared functional specialization in transformer-based language models and the human brain. Nature communications, 15(1):5523, 2024.

[10] Xuanhao Liu, Yan-Kai Liu, Yansen Wang, Kan Ren, Hanwen Shi, Zilong Wang, Dongsheng Li, Baoliang Lu, and Wei-Long Zheng. Eeg2video: Towards decoding dynamic visual perception from eeg signals. In The Thirty-eighth Annual Conference on Neural Information Processing Systems.

[11] Venkatraman Manikandan and Suresh Neethirajan. Decoding poultry vocalizations-natural language processing and transformer models for semantic and emotional analysis. bioRxiv, pages 2024–12, 2024.

[12] Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. Psychometry: An omnifit model for image reconstruction from human brain activity. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 233–243, 2024.

[13] Dayong Ren, Zhe Ma, Yuanpei Chen, Weihang Peng, Xiaode Liu, Yuhan Zhang, and Yufei Guo. Spiking pointnet: Spiking neural networks for point clouds. Advances in Neural Information Processing Systems, 36, 2024.

[14] Ana Stanojevic, Stanisław Woźniak, Guillaume Bellec, Giovanni Cherubini, Angeliki Pantazi, and Wulfram Gerstner. High-performance deep spiking neural networks with 0.3 spikes per neuron. Nature Communications, 15(1):6793, 2024.

[15] Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11333–11342, 2024.

[16] Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, et al. A high-performance speech neuroprosthesis. Nature, 620(7976):1031–1036, 2023.

[17] Yiqian Yang, Yiqun Duan, Qiang Zhang, Renjing Xu, and Hui Xiong. Decode neural signal as speech. arXiv preprint arXiv:2403.01748, 2024.

Reply to the Reviewers

Re: Manuscript ID COMMSBIO-24-5522
"Linguistic Neural Decoding with Deep Learning: Progress, Challenges and Future"
Authors of the paper
Communications Biology

Dear Editors and Reviewers:

Thank you for your letter and the reviewers' comments concerning our manuscript entitled "Linguistic Neural Decoding with Deep Learning: Progress, Challenges and Future" (ID: COMMSBIO-24-5522). Those comments are all valuable and very helpful for revising and improving our manuscript, as well as the important guiding significance to our research. We have studied the comments carefully and have made revisions which we hope meet with approval. Our response is presented below in blue, with comments and questions from the reviewers in black and quoted reference text from the revised manuscript in red. The content with a strikethrough means that we have deleted the improper statements in the original text, and the reference signs in the quoted text have been kept aligned with the revised manuscript to make our response clearer to read. In the revised manuscript, the added content is also presented in red. In addition to the use of color content, the font size of Fig. 2 and Fig. 6 in the revised manuscript has been increased in response to the suggestion of Reviewer #3.

Responds to Reviewer #1:

I still don't believe that this piece is at the level of Communications Biology. However, let the other reviewers take the lead on the decision.

Thank you very much for your continued feedback and insights on our manuscript. We understand and respect your perspective regarding the suitability of our paper for Communications Biology. We are grateful for the opportunity to improve our research. We respect the decision-making process involving all reviewers and will await their collective judgment. Thank you once again for your valuable time and consideration.

Responds to Reviewer #2:

I thank the authors for their thoughtful responses to my initial critiques. It seems that the authors and I have fundamentally different views on what a review of neural decoding should aspire to achieve. I am coming at this from the perspective of a cognitive neuroscientist interested in advancing the basic science of semantic cognition who sees the potential for machine learning/AI to aid in that pursuit. This is why I was eager to find connections between the state-of-the-art technology and cognitive neuroscience being drawn throughout the review. I may not be representative of the readership of Communications Biology.

Thanks to reviewer #2 for the suggestions. Regarding the implementation of neural decoding, our views are actually not contradictory. If neurocognition can better model the cognitive and expression processes of the brain, machine learning/AI can avoid more noise and approximation. We look forward to expanding this aspect in the future, but we think the current volume of the article is detailed enough.

While the revised manuscript is more explicit about communicating its intended scope, it remains my opinion that the review would be strengthened by commenting on architectural/computational/representational differences between BrainTranslator, DeWave, UniCoRN, etc. in the brain recording translation section.

We understand your concerns. In the section on brain recording translation, we illustrated the task setup through Fig. 4 and summarized the primary works in Table 3, including their model architectures, data sources, and achieved performance. We have revised this section to provide more detailed explanations for some of the models, and the added explanations or new content are displayed in the following red paragraphs. However, regarding models like DeWave and UniCoRN, which opted for a teacher forcing approach during testing, their practical applicability remains uncertain. Consequently, we have chosen not to overly emphasize these solutions.

Alternatively, a solution with implementation potential was proposed to generate text directly from MEG recordings without teacher forcing [54]. The proposed architecture, NeuSpeech, utilized MEG instead of EEG or fMRI and incorporated a Whisper model. During the training process, only the AdaLora module and the convolutional layers within the encoder were fine-tuned, while the Transformer layers in the encoder and the entire decoder remained frozen. Compared to other approaches, NeuSpeech corroborated the association between brain activity and speech signals.

Under this paradigm, more work emerged that implements LLMs to translate brain signals in large vocabularies, including schemes using contrastive learning and curriculum learning [51]. By constructing positive and negative sample pairs from EEG of different subjects exposed to identical or different sentence stimuli, the method aimed to pull closer the representation distances of semantically similar inputs, while pushing apart dissimilar ones. More similar sample pairs are considered challenging, and the strategy followed a progression from easy to difficult.

Under the same experimental context, [132] employs encoders and projectors to align the distributions between fMRI and text. An external large model, GPT, samples candidate words before selecting the option with the closest distribution with predicted fMRI signal. This process completes the sequence decoding in an autoregressive manner.

Responds to Reviewer #3:

I thank the author group for their responses and revisions. The responses are persuasive, and I have no further major concerns regarding the updated version. I especially appreciate section 6, in which the author provides a much better summary and insights for future work. I believe that this paper contributes significantly to the field of linguistic neural decoding and is directionally a good survey in the field.

Thank you very much for your positive feedback and for acknowledging the revisions we have made. We are delighted to hear that you found our responses persuasive and that you have no further major concerns regarding the updated version of the manuscript. Your recognition is highly encouraging and motivates us to continue refining our work in this area.

1. Minor-1: The font size within the images should be comparable to the main text font size. For example, some text in Figures 6 and 2c is too small.

   Thank you very much for your careful review and for pointing out the issue with the font size in the figures. We appreciate your attention to detail, which helps improve the clarity and readability of our manuscript. We agree that the text within the images should be appropriate to ensure consistency and ease of reading. Specifically, regarding Figures 6 and 2c, we have increased the font size of the text. Additionally, we have reviewed all other figures to ensure uniformity across the manuscript.

2. Minor-2: Some expressions may be controversial, so the author might consider using more cautious expressions. e.g., Section 3.1: The major languages in the world have their textual form, and communication via text outperforms due to its efficiency and concision.

   We appreciate your attention to the sensitivity required when discussing potentially controversial expressions. Regarding Section 3.1, we have revised the content to ensure a more balanced and nuanced discussion.

   The major languages in the world have their textual form, and communication via text is often valued for its efficiency and conciseness.

3. Minor-3: The author frequently addresses a moderate, large, or small set of candidates or vocabulary for neural decoding. Are these descriptions comparable, and how many different expressions are there?

   We used these terms ("moderate", "large", "small") to qualitatively describe the size of candidate sets or vocabularies based on their relative sizes within the context of specific studies. A moderate or small vocabulary typically contains less than ten, at most a few dozen candidates. Specifically, in [139], the decoding result includes 6 words and 2 pseudowords, a total of 8 candidates. In [70-72], the experiments are conducted on a vocabulary of 5, 15, and 3 candidates respectively. This setting is usually suitable for experiments on small amounts of non-intrusive data, where only a few phonemes or words are considered. The opposite is conducted on a large vocabulary, even open-vocabulary decoding. In [80], the decoding scope has been extended to 125,000 words. This is achieved by first implementing full-set phoneme decoding and then implementing phoneme-to-word sequence conversion through an external

language model. Similarly, in some work, a large language model is used to align neural features with the text distribution, thereby achieving a decoding range equal to the large language model vocabulary. In our manuscript, we include detailed descriptions of the vocabulary size and provide some examples below.

Larger vocabularies bring greater difficulties. In [25], a network module with dense layers and a regression-based decoder was implemented to directly classify an fMRI scan over a 180-word vocabulary.

Following these achievements, in [26] the researchers predicted masked words and phrases. The proposed approach utilized an encoder-decoder paradigm and achieved 18.20% and 7.95% top-1 accuracy over a 2,000-word vocabulary on the two tasks respectively.

Due to the small vocabulary (typically involving several, a dozen, or several dozen candidates), such approaches employed classifiers.

Through this work, researchers achieved a 25.8% WER on a vocabulary of 125,000 words within the acceptable bound of performance [144], with a recognition rate of 62 words per minute.

4. Minor-4: The author addresses metrics in domains such as machine translation (MT), text-to-speech (TTS), and automatic speech recognition (ASR). Are the metrics different when applied to neural decoding?

   These metrics all measure the similarity between the predicted and ground-truth sequences. Those derived from TTS are mainly used in brain-to-speech tasks, for the decoding outputs of both are speech waves. The decoding targets of ASR and MT (machine translation) tasks are both text sequences, and the difference lies in the requirements for accuracy. The MT task is to translate the same semantics from the source language to the target language. Since the same sentence can be expressed in multiple text sequences, the evaluation metrics only involve semantic consistency, which are widely used in brain recording translation. Brain recording translation uses non-intrusive data and cannot achieve high-precision decoding of text sequences (usually with extremely high WER), so the decoding requirements need to be weakened to the semantics consistency. When the data is replaced with invasive data, the indicators of the ASR task are widely used. In summary, the evaluation of neural decoding needs to be selected according to the decoding outputs and difficulty, and metrics that can intuitively reflect the experimental results are proper ones.

   As for sequential decoding, ASR and MT tasks generate text sequences, with the primary distinction being the requirement for accuracy. The MT task involves translating identical semantics from a source language to a target language, allowing for multiple valid text sequence representations of the same sentence. Consequently, evaluation metrics for MT focus on semantic consistency, which are extensively employed in brain recording translation. Given that brain recording translation utilizes non-invasive data and cannot achieve high-precision text sequence decoding, the evaluation criteria must prioritize semantic consistency over exact textual matches. When invasive data is used, ASR metrics become more applicable.

   In natural listening and speaking scenarios, reconstructing the speech wave can be necessary. Those derived from TTS are mainly used in brain-to-speech tasks, for the decoding outputs of both are speech waves.

5. Minor-5: Some recent research could be considered, I list some suggestions here.

   We appreciate your effort in providing these suggestions, which will undoubtedly strengthen the relevance and depth of our work. We have reviewed the references you provided and incorporated them into our manuscript.

   A recent study has indicated that, in addition to model scaling, the amount of data utilized during the training process positively influences the similarity of representations between the brain and neural networks. Furthermore, alignment training is deemed an effective approach to enhancing this similarity [104].

   Under the same experimental context, [132] employs encoders and projectors to align the distributions between fMRI and text. An external large model, GPT, samples candidate words before selecting the

option with the closest distribution with predicted fMRI signal. This process completes the sequence decoding in an autoregressive manner.

in the initial attempts, the LLMs were adopted to generate hypothesis candidates with a separate module score for each potential sentence [56,132]

Responds to Reviewer #4:

Thanks to reviewer #4 for further pointing out and explaining the drawbacks in the revised manuscript. The opinions from the first revision and this round are listed together for better illustration.

1. "Linguistic neural decoding reconstructs the stimuli information perceived by text or speech, which has shown groundbreaking achievements, thanks to the improvement of neuroimaging, biology and artificial intelligence." How has "biology" improved and how can we thank biology for making groundbreaking achievements? Biology has arguably not changed in the past few years and has not enabled advancements in neural decoding?

   We think the development of biology has indeed promoted neural decoding. This perspective is closely tied to our definition of biology, which encompasses fields such as brain science and neuroimaging—domains that fall under the broader umbrella of biological sciences. Specifically, high-resolution neuroimaging techniques and more precise signal collection methods have played a crucial role in advancing neural decoding capabilities.

   Right, but then "biology" needs to be defined in that particular way, according to the author's definition. Otherwise, the reader does not know, and the statement seems invalid.

   Thanks for your response. The statement about "biology" has already been removed in the previous revised version. This change aims to enhance the accuracy and rigor of the paper.

2. "In this work, we present a taxonomy of recent neural decoding progress, focusing on deep learning architectures and strategies, especially those implementing large language models (LLMs) for their powerful reasoning capacity." I believe that most studies do not leverage "LLM reasoning", namely the embeddings. It is true that LLM reasoning can be leveraged, but I do not think that is the main focus of the field.

   For the role and potential of LLMs in neural decoding, we hold conflicting views. The capabilities of LLM reasoning have been verified in the past few years. It was first demonstrated in text understanding and generation, and it should take time to expand in other areas. Among the work we have summarized, most of the recently published papers have utilized LLMs, and they have shown superior performance in non-invasive continuous decoding, which may be an ideal paradigm for future brain-computer interfaces. In our revised manuscript, we additionally include experiments in the past few months, majorly utilizing an LLM backbone. Even for invasive neural decoding, we noticed that fine-tuning with about 10k sentences could achieve comparable results to the cascade model. Considering the complexity of the LLM backbone, further data expansion is likely to bring more stable performance improvements. In addition, this also shows that LLMs trained on a vast amount of text preserve the ability to understand and reason about neural signals, which is an interesting finding.

   I do not understand this response, sorry. I feel like several things mentioned in this response is out of context: what is the 10k sentences and why is a cascade model relevant here? I also think there is a conflicting use of "LLM reasoning". In parts of the response, you talk about LLMs learning to reason about e.g. text. In the last part of the response you talk about LLMs "reasoning" about neural signals? I don't think LLMs reason about neural signals, they predict the signals, no?

   Thank you for your comments. We apologize for the confusion caused by our wording, and it was indeed our negligence and misunderstanding. We have replaced all instances of "LLM reasoning about neural signals" with "prediction" or "inferring" as these terms are more commonly used within the domain of large language models (LLMs). The term "reasoning" generally implies a deeper engagement than mere prediction; it involves understanding the underlying meanings and making inferences based on this understanding. Utilizing LLMs for neural decoding, however, has not yet reached this level of sophistication. This modification will ensure our terminology aligns more closely with established

practices in the field and accurately reflects the current state of research. In addition to "prediction", "inferring" is often associated with the application phase of trained models, where these models are used to make predictions or generate sequences based on the learned patterns from training data. Given this context, using "inferring" and "prediction" in place of "reasoning" when discussing tasks related to neural signals is both reasonable and appropriate.

Regarding 10k sentences and cascade models, we aim to highlight the significant potential of the approaches with LLM backbone. Our assertion is that using LLMs as the core component of neural decoding represents a major and enduring direction in the field. This approach enables more nuanced interpretations of neural signals by leveraging the predictive and inferential capacities of LLMs.

In this work, we present a taxonomy of recent neural decoding progress, focusing on deep learning architectures and strategies, especially those implementing large language models (LLMs) for their powerful information understanding, processing, and generation capacity.

Another work proposed simultaneously leveraged the inferring ability of LLMs and implemented an fMRI encoder to learn a suitable prompt in an auditory-decoding setting.

Another recent approach introduced an E2E framework with pre-trained LLMs for decoding invasive brain signals, leveraging the comprehensive inferring capability of GPT-2, OPT and Llama2.

3. "This article will help brain scientists and deep learning researchers gain a bird's eye view of language perception, and thus facilitate their further investigation." Why is language denoted as "perception"? The first sentence in the introduction says that language is exclusive to "higher-order" organisms? Language—as opposed to speech perception—is not tied to perception given that certain brain regions are active regardless of the perceptual modality (e.g., Deniz et al 2017).

   We apologize for any confusion caused by our original wording. To clarify, by "language perception", we refer to the process by which the human brain receives and interprets external language stimuli. Achieving high-precision neural decoding would indirectly demonstrate our understanding of the brain activities elicited by language stimuli, thereby enhancing our comprehension of the language perception process. In response to feedback from multiple reviewers, we have revised this expression in the revised manuscript to ensure clarity and precision. Additionally, we acknowledge the controversy surrounding the attribution of language capabilities to advanced organisms alone. Considering that various species have been confirmed to transmit information through vocalization, we have modified this statement to reflect a more inclusive and accurate representation of communication across different species.

   Language serves as the paramount medium facilitating human communication and information exchange. The human brain plays a crucial role in language interaction, enabling complex cognitive processes supporting the perception, comprehension and production of language.

   This article aims to provide brain scientists and deep learning researchers with an overarching viewpoint of the significant correlations observed in the human brain during language perception and production from a methodological perspective, and thus facilitate their further investigation.

   Thanks for the response. However, in the red text, I do not see how you tackle the definition of perception versus language. This should be revised. I think you need to define how you use the terms perception and language in the article?

   Thank you for your suggestions. We agree with your perspective that it is necessary to define the terms related to language and perception within the scope of this article. Accordingly, we have added the following content at the appropriate locations, hoping that this meets your requirements.

   In this paper, the process by which the brain receives external language stimuli is referred to as perception, which primarily involves the neural encoding process. During this process, external stimuli are transformed into specific neural response patterns, ...

   It is important to note that the language discussed in this paper is a synthesis of semantic and syntactic information, featuring specific content presented in a defined format, primarily including text and speech forms. Visual image reconstruction is excluded, as it contains semantic content but lacks linguistic syntactic presentation. Similarly, motions such as handwriting are not considered, given their involvement with bodily movements and minimal relevance to semantic and syntactic aspects of language.

4. "When processing natural language, ANNs exhibit patterns of functional specialization similar to those of cortical language networks". Which studies have shown this? Functional specialization is a strong claim (see e.g., Kanwisher, 2010).

   When processing language information, both neural networks and the human brain exhibit certain similarities. Both possess neural units that are sensitive to specific signals or functions. Research in this area has evolved from examining relatively simple network structures to exploring LLMs. Despite limitations imposed by differences in network architectures and training resources, there is evidence of functional specialization within these models. While the specific neural networks responsible for particular functions may vary across different models due to structural inconsistencies, functional and causal evidence supports the specialized processing.

   Thank you.

5. "it has been verified that the brain encoding models and pre-trained LLMs follow the scaling laws, where the model performance increases as the number of parameters grows, indicating the necessity to develop more complex architecture to bridge the brain activity patterns and human linguistic representations". I don't think it logically follows that one needs to build novel architectures? If scaling works, then one might argue that it is enough to make the models bigger? Of course, novel architectures would be helpful, but I don't think it follows from the references provided in these sentences.

   We understand the reviewer's concern about whether it is possible to conclude from the research we cited that simply expanding the model is enough to achieve stable performance improvements. We stated a more complex architecture is a necessary factor for neural decoding to achieve considerable performance in the manuscript, but not the only factor. The effect of this task is also closely related to the amount and quality of data, etc. In the article we cited, the original expression is "increasing scale in both models and data will yield incredibly effective models of language processing in the brain, enabling better scientific understanding as well as applications such as decoding", and we emphasized the scale in models. We believe that it is reasonable to conclude that it is necessary to increase the complexity of the models, but the argument that "it is enough to make the models bigger" is a misunderstanding of the content of our manuscript.

   As shown in Fig. 2c, it has been verified that the brain encoding models and pre-trained LLMs follow the scaling laws, where the model performance increases as the number of parameters grows, indicating the necessity to develop more complex architecture to bridge the brain activity patterns and human linguistic representations.

   I agree with "As shown in Fig. 2c, it has been verified that the brain encoding models and pre-trained LLMs follow the scaling laws, where the model performance increases as the number of parameters grow", but I do not see how the last part of the sentence about "developing more complex architectures" follows from that?

   Thank you for your response. There appear to be different perspectives on the concept of "more complex" structures. In our perspective, even if there is no significant enhancement in the model architecture, merely increasing the number of layers or enlarging the dimensions of the latent space would render the model more complex. This is due to the fact that decoding identical content would necessitate a higher computational cost. Indeed, in practical scenarios, the quantity of model parameters reaches a bottleneck due to various factors, with the most critical aspect being data. Firstly, there exists a trade-off between data quality and quantity. Acquiring large-scale, clean invasive data is challenging, whereas non-invasive data, which is more accessible, often suffers from noise issues. Secondly, neural signals can vary with subjects and over recording times, potentially necessitating specific distributions for accurate modeling. Should these two issues be adequately addressed, we believe that it would become feasible to introduce models of a larger scale. Considering that our wording might have caused confusion among readers, we have revised our explanation to be more straightforward, focusing on the aspect of parameter quantity and avoiding any emphasis on complexity.

   As shown in Fig. 2c, it has been verified that the brain encoding models and pre-trained LLMs follow the scaling laws, where the model performance increases as the number of parameters grows, indicating the necessity to develop models with larger parameter scales to bridge the brain activity patterns and human linguistic representations if given sufficient data and other necessary conditions.

6. The authors should define what they mean by "language". Does it entail speech/visual perception? Or do they talk about language as a "higher-level", amodal process (e.g., Fedorenko et al. 2024)?

   Our manuscript only includes neural decoding related to language, and therefore excludes visual reconstruction and generation of body movements. Language can be presented in various forms, including receiving text and speech stimuli, as well as spontaneously imagining text or trying to speak. Since handwriting is more related to motion, we do not include this part. We have explicitly explained our scope in the revised manuscript.

   It is important to note that the scope of this review encompasses neural decoding processes in which language serves as the medium, specifically in the forms of written text and spoken speech. This study excludes analyses related to handwriting, which involves motor function, as well as visual image reconstruction.

   Please see point 3.

   Thank you for your suggestions. The definition of language as used in this article is presented in the red response under point 3.

7. Many model types are mentioned. Oftentimes it is not mentioned what those are, e.g., VQ-VAE (page 11) – it is unclear why that is of relevance. If the authors chose to mention these model types, it needs to be clear why.

   We agree with the comments that the exclusion of the details of the model architecture mentioned in the review would interfere with the reader's understanding. We have added a brief introduction in the appendix to summarize the model structure. We believe that such changes are reasonable and will make the paper more reader-friendly.

   Thanks for the table – however, I think that beyond a table, it is necessary to understand why a given model is brought up, as a general point to improve the manuscript.

   Thank you for your suggestions. The rationale behind the application of model architectures in neural decoding can indeed be complex to explain. Many models have been validated on other tasks, such as VQ-VAE, which was originally applied to extract discrete representations and perform information compression for modalities including images, speech, and video. Although neural signals are distinct from these modalities, they exhibit characteristics similar to audio, particularly speech, when the signal source is confined to listening, speaking, or inner speech. Therefore, it is reasonable to adapt and apply these methods in the context of neural decoding. Ultimately, models focus on maintaining associative mappings rather than distinguishing between whether they are processing speech or neural recordings.

Reply to the Reviewers

Re: Manuscript ID COMMSBIO-24-5522
"Linguistic Neural Decoding with Deep Learning: Progress, Challenges and Future"
Authors of the paper
Communications Biology

Dear Editors and Reviewers:

Thank you for your letter and the reviewers' comments concerning our manuscript entitled "Linguistic Neural Decoding with Deep Learning: Progress, Challenges and Future" (ID: COMMSBIO-24-5522). Those comments are all valuable and very helpful for revising and improving our manuscript, as well as the important guiding significance to our research. We have studied the comments carefully and have made revisions, which we hope will meet with approval. Our response is presented below in blue, with comments and questions from the reviewers in black and quoted reference text from the revised manuscript in red. In the revised manuscript, the added content is also presented in red. In addition to the revisions following the reviewers' comments, we modify the graphs to comply with the editorial policies.

In addition to addressing the reviewer's comments, we have also revisited the figures in our manuscript to align with editorial policies. In the original version, given that our article is in a review format, we used schematic diagrams for illustration purposes, which did not prioritize data accuracy. We have revised this section, with specific modifications including:

- Fig. 2c: The original image was reconstructed from the reference paper [1, 4]. Since the original paper did not provide the raw data, this image contains inaccuracies. We have replaced it with another image showing the perception, comprehension, and generation phases in communications.
- Fig. 3a: The image of the Semantic Space is a schematic representation intended to illustrate that different words are mapped into distinct spaces, with semantically similar words positioned closer. We replace it with the figure generated using word2vec model.

The rest of the figures comply with editorial policies. For the above modifications, we provide the original code and figure instances at the end of this rebuttal letter.

Responds to Reviewer #3:

I thank the authors for their revisions. I think the revised manuscript is good. I have no further comments.

Thank you very much for your positive feedback and for acknowledging the revisions we have made. We are delighted to hear that you found our responses persuasive and that you have no further concerns regarding the updated version of the manuscript. Your recognition is highly encouraging.

Responds to Reviewer #4:

I thank the authors for their replies and changes to the manuscript. On a higher-level note, I think that this review is useful, but the structure of it and the motivation of certain sections is still lacking. Often, there are conflicting statements in different parts of the paper (which has the flavor of LLM logic to me, personally, but of course I cannot say). For instance, I asked the authors to define perception and language and stick with those definitions, but yet, despite definitions, the authors still add text that clearly does not disambiguate the two, for example (there are other cases of this):

"Language serves as the paramount medium facilitating human communication and information exchange. The human brain plays a crucial role in language interaction, enabling complex cognitive processes supporting the perception, comprehension and production of language. This article aims to provide brain scientists and deep learning researchers with an overarching viewpoint of the significant correlations observed in the human brain during language perception and production from a methodological perspective, and thus facilitate their further investigation."

And also you talk about "image reconstruction" as an example of perception–what does that mean? I believe that we perceive images, but I do not believe we have a high-fidelity reconstruction pipeline in our brains?

Thanks to reviewer #4 for further pointing out and explaining the drawbacks in the revised manuscript. We consider some of the comments to be reasonable, and have modified and emphasized them in the updated manuscript. For those we hold different points, we have also explained our views and hope this will address the concerns of Reviewer #4.

1. Often, there are conflicting statements in different parts of the paper (which has the flavor of LLM logic to me, personally, but of course I cannot say).

   We appreciate the reviewer's careful reading of the manuscript and their observation regarding perceived inconsistencies in certain sections. We acknowledge that some passages may require further clarification to ensure coherence, and we have revised those sections for greater precision. However, we would like to assure the reviewer that this work is entirely original, and any apparent contradictions are unintentional—likely a result of transitioning between different aspects of the argument rather than an artifact of automated generation. We welcome specific feedback on which passages raised concerns, as this would help us make targeted improvements.

2. I asked the authors to define perception and language and stick with those definitions, but yet, despite definitions, the authors still add text that clearly does not disambiguate the two, for example (there are other cases of this):

   "Language serves as the paramount medium facilitating human communication and information exchange. The human brain plays a crucial role in language interaction, enabling complex cognitive processes supporting the perception, comprehension and production of language."

   We think the reviewer's interpretation has likely led to some misunderstanding. We have compiled and explained all instances where these terms are mentioned throughout the manuscript, with the original unedited text highlighted in red and our explanations provided in blue. For ambiguous content, we have made modifications and indicated our proposed revisions.

   For "Perception":

   This article aims to provide brain scientists and deep learning researchers with an overarching viewpoint of the significant correlations observed in the human brain during language perception and production from a methodological perspective, and thus facilitate their further investigation.

   Language perception refers to the process by which external textual or auditory signals are received by the brain, generating specific neural representations. Some tasks addressed in this manuscript, including stimuli recognition and brain recording translation, involve decoding the original stimuli based on these neural representations.

   The human brain plays a crucial role in language interaction, enabling complex cognitive processes supporting the perception, comprehension and production of language.

   In a typical interactive scenario, the brain undergoes the following processes: Perception converts external linguistic stimuli into specific neural patterns; comprehension involves steps such as semantic extraction, understanding, and reasoning; production entails outputting responses in a specific form, for example, by guiding the vocal organs to produce speech. In response to the reviewer's concerns, we have added Fig. 2c in the revised version to better illustrate this process.

   In this paper, the process by which the brain receives external language stimuli is referred to as perception, which primarily involves the neural encoding process.

   This sentence clarifies the definition of perception, emphasizing in our revised manuscript that perception not only involves receiving external stimuli but also requires their conversion into specific neural representations. In the original manuscript, it can be inferred from the involvement of neural encoding in perception.

   In natural listening settings, the human brain encodes a wide range of acoustic features and processes external language stimuli temporally through prediction, highlighting the importance of contextual information in cortical perception, even at the level of single neurons.

A correct mention of perception. Within a history context, the brain encodes acoustic stimuli through a predictive mechanism. The same stimulus may be converted into different neural representations across varying histories, and this process involves perception.

Despite ample evidence supporting the predictive characteristics of human language processing, it is recently suggested the benefits actually come from the capacities of models to predict brain responses. Regardless of the mechanisms in the perception process, language models have the potential to understand and infer neural responses.

A correct reference to perception. The preceding discussion highlights the ability of language models to simulate the process of converting linguistic stimuli into neural representations, which corresponds to the process of perception. The second half then mentions the capability of language models to predict the stimuli or intention based on neural responses.

Some work treated the sentence-level responses as a combination of latent word effects, bridging the relationship of text perception in words and sentences.

The term "perception" has been removed from this sentence. The original intent was to convey that this work considers the neural representations corresponding to sentence stimuli as the accumulation of the neural representations of individual words.

The speech envelope refers to the variations in amplitude and intensity of a speech signal over time. It plays a crucial role in speech perception and understanding, for our brains are tuned to these variations, helping recognize speech sounds, syllables, and words.

The speech envelope contains both semantic and acoustic information, and different envelopes as external stimuli will elicit distinct neural representations. Therefore, this statement is also evident.

Previous studies have provided evidence for the neural representation of phonemes and other acoustic features during the perception of speech.

Although this statement appears in the section of inner speech recognition, it is intended to illustrate a prerequisite: "Speech perception requires the rapid and effortless extraction of meaningful phonetic information from a highly variable acoustic signal." [2, 6] There is a clear association between phonemes and neural representations. These two papers were originally cited in the manuscript; however, to enhance clarity, we omitted the specific citation markers in the above red text.

For "Language": The term "language" is mentioned 54 times in the main text. For the sake of brevity, we have only listed the most significant instances. However, it is important to note that we have thoroughly reviewed the entire manuscript to ensure there are no conflicting statements.

Language is the primary medium through which humans achieve information transfer and exchange. Through complex systems of symbols and rules, language enables the conveyance of ideas, concepts, and messages, thereby playing an indispensable role in social interaction and knowledge dissemination.

Regardless of whether language is in textual or spoken form, it serves as an essential means of communication.

Linguistic neural decoding aims to obtain outstanding language information from the evoked human brain during information interaction of both textual and spoken formats.

"Language information" refers to information related to language, which in our manuscript includes textual elements (such as phonemes, words, and text sequences) and speech waveforms.

This article aims to provide brain scientists and deep learning researchers with an overarching viewpoint of the significant correlations observed in the human brain during language perception and production from a methodological perspective, and thus facilitate their further investigation.

Language perception involves decoding the corresponding linguistic information from neural signals under natural reading and listening. This process encompasses the transmission from external stimuli to neural representations, including the perception process. Language production, on the other hand, includes speech neuroprostheses, which decode relevant linguistic information from neural signals during inner speech and attempted speech.

Language serves as the paramount medium facilitating human communication and information exchange.

Regardless of whether language is in textual or spoken form, it serves as an essential means of human communication.

The human brain plays a crucial role in language interaction, enabling complex cognitive processes supporting the perception, comprehension and production of language.

The concepts of language perception, comprehension, and production have already been explained in the explanation of perception. These three components represent crucial processes in language interaction, with the human brain playing a critical role.

Deep learning presents a viable solution for understanding language in the brain by utilizing large-scale trainable parameters to map the correlation between external stimuli and neural activity.

"Language in the brain" refers to the process by which linguistic information is transmitted to the human brain and subsequently processed and generated.

It is important to note that the language discussed in this paper is a synthesis of semantic and syntactic information, featuring specific content presented in a defined format, primarily including text and speech forms. Visual image reconstruction is excluded, as it contains semantic content but lacks linguistic syntactic presentation. Similarly, motions such as handwriting are not considered, given their involvement with bodily movements and minimal relevance to semantic and syntactic aspects of language.

This sentence is the definition of language in our manuscript: Language should encompass semantic information and be presented with a specific syntactic format, integrating both semantics and syntactic. Therefore, it does not include images (contain semantics but lack syntactic) or motion (unclear semantics and syntactic). Instead, we generally consider textual and spoken forms.

Brain recording translation... This task is analogous to machine translation, treating brain activity as the source language and translating it into human-understandable text.

Machine translation translates text from a source language into a target language. Here, "language" refers to language types, such as English, Chinese, and German, which are considered distinct languages.

In this paper, the process by which the brain receives external language stimuli is referred to as perception, which primarily involves the neural encoding process.

Similarly, external language stimuli encompass both textual and auditory forms, typically received by humans through natural reading and listening. The latter parts of the manuscript include multiple mentions of "language stimuli", and for the sake of brevity, we won't explain them repeatedly.

During this process, external stimuli are transformed into specific neural response patterns, with neural tracking ensuring the association between language and neural representations.

"language" refers to language-related features here including surprisal, phonetic sequences, word sequences, and other linguistic representations. It has been explained in the manuscript.

inner speech recognition... This task is highly correlated with ASR, for they both: 1) model the relation between diverse temporal features and deterministic textual information; 2) correlate with pronunciation and acoustics; 3) aim to generate language-compliant text.

"Language-compliant text" refers to text that contains semantic information and adheres to syntactic rules, neither meaningless token sequences nor random word combinations. For ASR and inner speech recognition, it is typically achieved by adding a language model score, which is also explained in the manuscript.

In theory, multiple elements of building a digital human can be obtained from invasive brain activity, including textual sentences, speech waves, facial movements, as well as body movements not related to language.

Body movements not related to language refers to movements unrelated to speech production. Broadly speaking, these include movements that go beyond facial motion and encompass other types of motion contributing to the overall representation of a digital human.

Current experiments on multiple tasks have shown that language, speech, and visual reconstruction of neural signals have achieved the ability to restore semantic features.

There is indeed an imprecise expression here. According to common usage in the literature, "language" includes both text and speech. This sentence has been revised in the manuscript.

3. And also you talk about "image reconstruction" as an example of perception–what does that mean? I believe that we perceive images, but I do not believe we have a high-fidelity reconstruction pipeline in our brains?

   Our manuscript merely explains why visual image reconstruction was not included as part of the study's scope and did not delve into any detailed inferences. We are at a loss to understand how the reviewer arrived at the conclusion regarding a "high-fidelity reconstruction pipeline in our brains"—our content does not warrant such conclusions.

   Visual image reconstruction refers to the attempt to decode brain activity signals to reconstruct visual images perceived, remembered, or imagined by humans. This process is closely related to the linguistic neural decoding addressed in our paper: both involve semantic information reconstruction. Recent advances in visual image reconstruction have achieved some level of detail [3, 5, 7, 8]. It should be noted that visual image reconstruction entails semantic details like object categories, events, colors, etc., but lacks well-defined syntactic information and has a weaker connection with language (without strong alignment with a corresponding text sequence). Considering these points, recent advancements in visual image reconstruction are beyond the scope of our current work. Moreover, it is important to highlight that our manuscript mentions visual image reconstruction only once, emphasizing that this aspect is not covered in our work. We are unsure why the reviewer places significant emphasis on this term.

4. On a higher-level note, I think that this review is useful, but the structure of it and the motivation of certain sections is still lacking.

   We are interested in specific suggestions from the reviewer to address this perceived shortcoming. Considering that we have provided detailed explanations for all other comments, we hope the reviewer can further reconsider our motivations and structure, or provide concrete and detailed suggestions for improvements.

   In the current version, our manuscript primarily focuses on linguistic neural decoding, offering methods for neuroscientists and computer engineers to decode language stimuli or intentions from neural activities. We begin by introducing the definitions and scope related to linguistic neural decoding, followed by a brief overview of the theoretical foundations enabling the construction of language information from neural representations. We then discuss widely adopted metrics for experiment evaluation, enumerate various experimental paradigms along with their development and latest methodologies, and conclude with potential future directions. Additionally, we have included a content flow diagram for better illustration, as shown in Fig. 1 of the manuscript.

Figs:

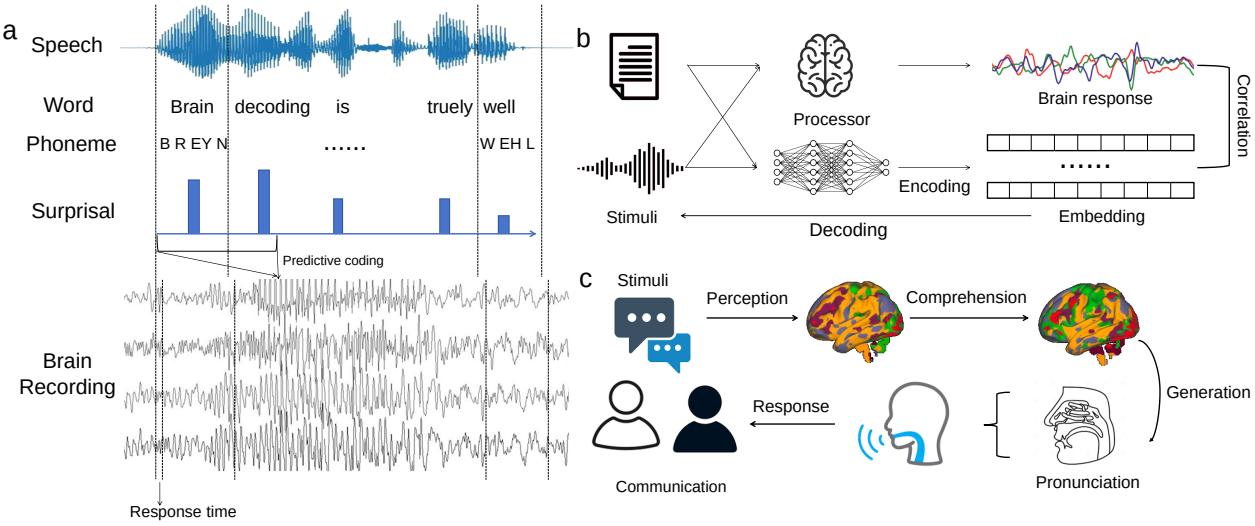

Figure 2: **The formation of linguistic representation in the human brain.** a. The human brain tracks the dynamic flow of speech and linguistic properties with minor response delay, and the neural response is performed in a continuous predictive manner. b. The human brain and the neural networks can both encode textual or verbal stimuli into specific representations, and the decoding process aims to reconstruct the linguistic information. c. In vocal communications, the brain processes perception, comprehension, and generation, which is performed by guiding the movements of articulators.

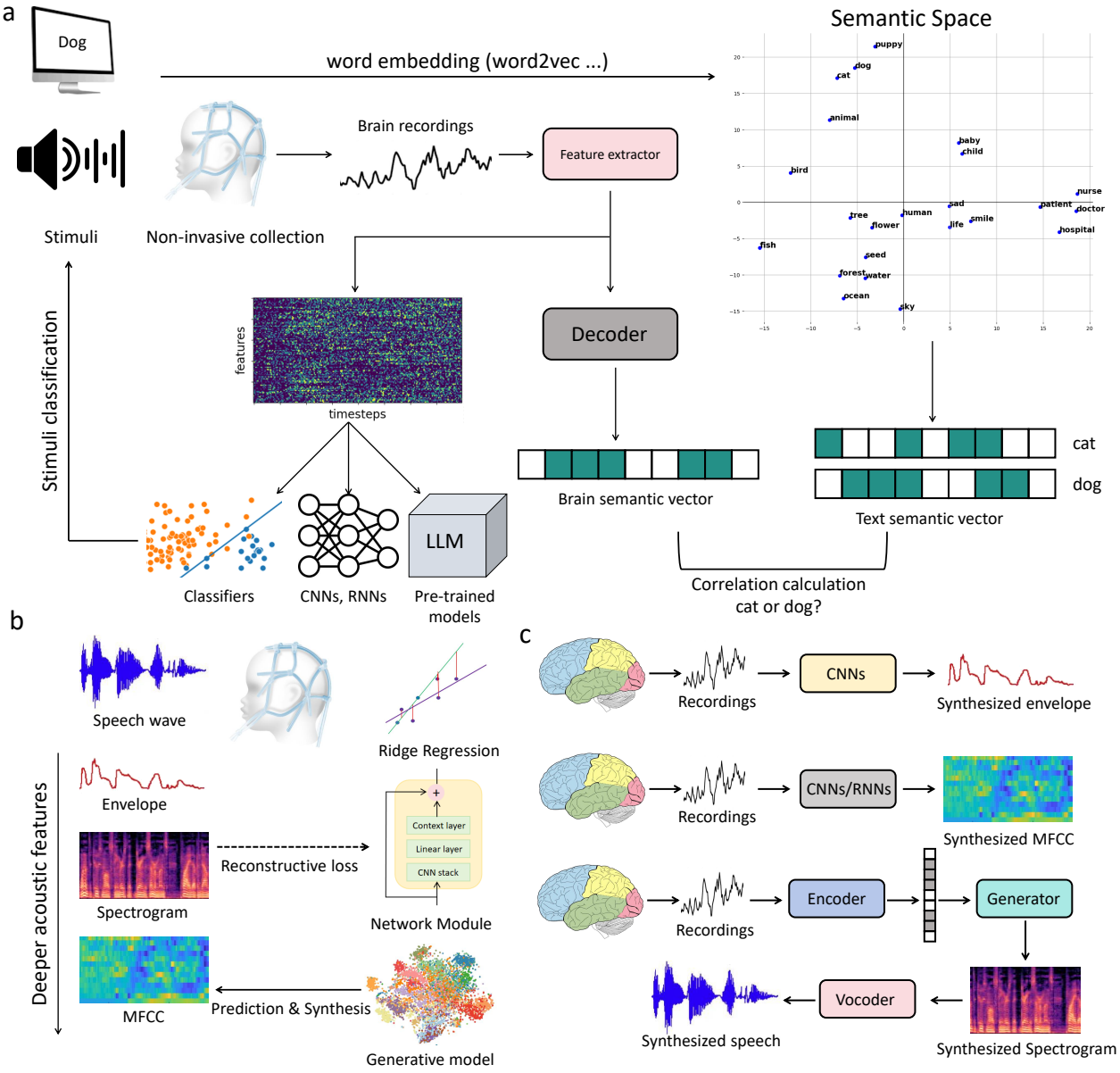

Figure 3: **Stimuli recognition of evoked brain activity.** a. An overview of the stimuli recognition task. The subject receives textual or vocal information while the active brain signals are collected. The raw brain recordings are processed into feature space, followed by classifiers, networks or pre-trained models to distinguish the original stimuli based on the complexity and candidate size. Several approaches adopted word embeddings (i.e. word2vec) to compare the decoded vector in a semantic space. b. In natural listening scenarios, restoring the original speech features and waveform is a more complex task. Regression models (i.e. ridge regression), CNN and RNN-based network modules, and paramount generation models (i.e. GAN) are widely used. c. The decoding architecture for various speech-related targets. The speech envelope can be easily reconstructed with CNNs while more complex networks are necessary for the decoding of MFCC. The most difficult task is to synthesize the stimuli wave, where an encoder-generator-vocoder architecture has been verified effective.

Code:

To reproduce the semantic space diagram in Fig.3, you need to download the word2vec model from `https://huggingface.co/fse/word2vec-google-news-300` and then run the following code. Other texts and images used in the manuscript were not generated using code.

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from gensim.models import KeyedVectors

model_path = 'word2vec-google-news-300.model'  # Update with your actual model path
model = KeyedVectors.load(model_path)

words = ['dog', 'puppy', 'fish', 'seed', 'patient', 'cat', 'animal', 'tree', 'flower',
         'sky', 'bird', 'fish', 'ocean', 'water', 'forest', 'human', 'doctor', 'nurse', '
    hospital', 'baby', 'child', 'life', 'sad', 'smile']

word_vectors = np.array([model[word] for word in words])*10

pca = PCA(n_components=2)
reduced_vectors = pca.fit_transform(word_vectors)

plt.figure(figsize=(12, 10))
plt.scatter(reduced_vectors[:, 0], reduced_vectors[:, 1], color='blue')

for i, word in enumerate(words):
    plt.text(reduced_vectors[i, 0], reduced_vectors[i, 1], word, color='black', fontsize=15,
     fontweight='bold')

plt.axhline(0, color='black',linewidth=1)
plt.axvline(0, color='black',linewidth=1)

ax = plt.gca()
ax.spines['top'].set_color('none')
ax.spines['right'].set_color('none')
ax.spines['left'].set_color('none')
ax.spines['bottom'].set_color('none')

plt.title('Semantic Space', fontsize=18, fontweight='bold', pad=20)
plt.grid(True)
plt.show()
```

References

[1] Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. Advances in Neural Information Processing Systems, 36, 2024.

[2] Edward F Chang, Jochem W Rieger, Keith Johnson, Mitchel S Berger, Nicholas M Barbaro, and Robert T Knight. Categorical speech representation in human superior temporal gyrus. Nature neuroscience, 13(11):1428–1432, 2010.

[3] Xiangtao Kong, Kexin Huang, Ping Li, and Lei Zhang. Toward generalizing visual brain decoding to unseen subjects. arXiv preprint arXiv:2410.14445, 2024.

[4] Haowei Lin, Baizhou Huang, Haotian Ye, Qinyu Chen, Zihao Wang, Sujian Li, Jianzhu Ma, Xiaojun Wan, James Zou, and Yitao Liang. Selecting large language model to fine-tune via rectified scaling law. arXiv preprint arXiv:2402.02314, 2024.

[5] Yizhuo Lu, Changde Du, Chong Wang, Xuanliu Zhu, Liuyun Jiang, Xujin Li, and Huiguang He. Animate your thoughts: Reconstruction of dynamic natural vision from human brain activity. In The Thirteenth International Conference on Learning Representations.

[6] Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang. Phonetic feature encoding in human superior temporal gyrus. Science, 343(6174):1006–1010, 2014.

[7] Ruijie Quan, Wenguan Wang, Zhibo Tian, Fan Ma, and Yi Yang. Psychometry: An omnifit model for image reconstruction from human brain activity. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 233–243, 2024.

[8] Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11333–11342, 2024.

Reply to the Reviewers

Re: Manuscript ID COMMSBIO-24-5522
"Linguistic Neural Decoding with Deep Learning: Progress, Challenges and Future"
Authors of the paper
Communications Biology

Dear Editors and Reviewers:

Thank you for your letter and the reviewers' comments concerning our manuscript entitled "Linguistic Neural Decoding with Deep Learning: Progress, Challenges and Future" (ID: COMMSBIO-24-5522). Our response is presented below in blue, with comments and questions from the reviewers in black.

Responds to Reviewer #4:

Thanks to the reviewers for replying to my comments. I will leave it up to the authors and the editor to further decide whether any apparent inconsistencies in the manuscript are critical.

Thank you very much for your continued feedback and insights on our manuscript. We understand and respect your perspective regarding the suitability of our paper for Communications Biology. Thank you once again for your valuable time and consideration.