Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Alignment of Large Language Models and Brain Activity: Exploring Language Processing through sEEG in a Multimodal Syntactic Task

Master's Thesis

Victor Gillioz

vgillioz@ethz.ch

Kreiman Laboratory
Harvard Medical School

Supervisors:
Prof. Gabriel Kreiman
Prof. Julia Vogt

June 14, 2024

Acknowledgements

I would first like to thank Professor Gabriel Kreiman for giving me the opportunity to complete this research project in his lab and for his support throughout this time. My gratitude extends to the members of the lab. Thank you for your warm welcome and the enriching discussions on scientific and non-scientific topics. They have greatly contributed to my growth on an academic and personal level.

I am particularly thankful to my friends and family. Your unconditional support, in Boston and from abroad, has been invaluable. Your encouragement and love have made all the difference during this significant phase of my life.

The countless meetings I have had during my time in Cambridge have been profoundly enriching, both professionally and personally, and I will keep the memories of these interactions close to my heart as I move forward.

Thank you all for your integral roles in my journey.

Abstract

Recent advances in large language models (LLMs) and cognitive neuroscience have opened up new avenues for understanding the neural basis of language processing. Building on the observation that LLMs can effectively be used as predictors of neural activity during language tasks, we investigated the alignment between GPT-2 XL's inner representations and neural activity from participants engaged in a multimodal linguistic task. Subjects were presented with sentences containing semantic and syntactic violations through audio and visual modalities. Their neural activity was recorded using stereo-encephalography (sEEG), allowing us to explore precise spatial and temporal dynamics of brain responses to language processing.

We identify the impact of hidden factors, such as word position, on the model alignment score with neural activity and warn against the risk of drawing incorrect conclusions from observed correlations. We explore spatial and temporal dynamics of neural responses through additional paradigm control. Our findings reveal that GPT-2 XL representations align with neural activity patterns across sentence structures and modalities and suggest that LLM-based approaches can provide insights into specific neural correlates of language processing, such as syntactic violations. This study contributes to the growing body of research at the intersection of artificial intelligence and cognitive neuroscience, extending our understanding of the neural underpinnings of language processing.

Contents

Chapter 1

Introduction

Language is a core aspect of human communication, enabling the exchange of information and ideas. It is not only central to everyday interactions but also a complex cognitive function that engages multiple regions of the brain. Historically, the neural basis of language processing has been thoroughly studied using various neuroimaging techniques, such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG). These studies have provided valuable insights into the neural mechanisms underlying language processing, including the involvement of specific brain regions in different aspects of language comprehension and the temporal dynamics of neural activity during language processing tasks.

The advent of artificial intelligence, especially through the development of advanced deep neural networks, has introduced new methodologies for exploring cognitive processes. Recent research leveraging state-of-the-art models has demonstrated the potential of comparing deep neural network activity to neural activity for studying cognitive tasks. Notably, recent large language models (LLMs) have been used to study the neural basis of language processing, revealing intriguing correlations between the computational processes in these models and biological language processing in the human brain. This line of research has highlighted the potential of model-to-brain comparison as a powerful tool for gaining deeper insights into the neural mechanisms that underlie language, suggesting shared computational principles between language models and the brain.

1.1 Motivation

Building upon previous studies, this study proposes to explore the neural basis of language processing using a novel multimodal stereoencephalography (sEEG) dataset containing sentences with semantic and syntactic violations. Our research is driven by the hypothesis that neural responses vary with changes in sentence construction and that these variations can be observed by comparing neural activity to large language model representations. By examining these dynamics under controlled experimental conditions, this work aims to enhance our understanding of how different linguistic modalities and sentence structures affect brain function during language tasks.

1.2 Research Objectives

The primary objective of this study is to empirically validate existing literature across visual and auditory modalities and different sentence structures and provide insights into the factors contributing to a high alignment between models and the brain. Initially, our research will focus on how different combinations of linguistic conditions from our dataset impact the alignment scores between neural signals and model predictions. Subsequently, we aim to identify electrodes with predictable neural activity in specific modalities and analyze how variations in sentence type may affect the temporal alignment scores associated with these electrodes. Through this rigorous analysis, this study seeks to contribute valuable new insights into the modulatory effects of sentence structure on neural language processing, leveraging recent advances in using large language models in cognitive neuroscience research.

Chapter 2
Background

2.1 Language in the Brain

The human brain processes language through a complex interplay of various regions and mechanisms, each specializing in different aspects of linguistic comprehension and production. Central to this complex system is the role of syntax, the set of rules, principles, and processes that govern the structure of sentences in a given language.

2.1.1 Neuroanatomy

The neuroanatomical basis of language processing involves both distributed and specialized brain areas. Language processing primarily engages areas in the left hemisphere, notably Broca's area in the left inferior frontal gyrus and Wernicke's area in the left superior temporal gyrus, although both hemispheres contribute to different extents depending on the linguistic task [1].

The anterior regions are generally associated with processing the grammatical structure of sentences. Within these regions, Broca's area, which includes Brodmann areas 44 and 45, is traditionally associated with syntactic processing and speech production. This region plays a role in manipulating the structure of sentences and maintaining syntactic information in working memory. Posterior regions, including Wernicke's area, are more involved with semantic processing and are linked with language comprehension. These regions are connected by a network of fibers, enabling rapid communication necessary for language processing. These findings are supported by neuroimaging studies showing different activation patterns when subjects process syntactic and semantic anomalies [1, 2].

The division of labor within the language network is reflected in Friederici's neurocognitive model of language comprehension, which posits that early stages of sentence processing involve the rapid identification of phonological and lexical elements by temporal regions such as the superior temporal gyrus (STG) and middle temporal gyrus (MTG), followed by the engagement of frontal regions, including the inferior frontal gyrus (IFG), in syntactic structuring and semantic processing. This model underscores the importance of temporal dynamics in understanding the neurobiology of language and proposes that different types of linguistic information are processed at different speeds and brain regions, culminating in the integration of both types of information [2, 1].

2.1.2 Syntax Processing

Syntax refers to the rules and principles that govern the structure of sentences in a language. It includes the arrangement of words and phrases to create well-formed sentences. Understanding syntax is crucial because it shapes language's meaning and communicative intent [3]. From a cognitive perspective, syntactic processing involves the identification of word categories, their relationships, and the application of grammatical rules.

The processing of syntax has been modeled in various ways in cognitive neuroscience. Syntax-first model proposes that syntactic analysis precedes semantic interpretation during sentence comprehension [4], while interactive models suggest that syntactic and semantic processes co-occur and influence each other dynamically throughout sentence comprehension [1, 5].

fMRI studies have shown that while lexical processing primarily engages the left temporal lobe, syntactic processing involves both the left frontal and temporal lobes, and suggest that these processes are interdependent, supporting a model where lexical and syntactic information is integrated during language comprehension [6]. Stereo-electroencephalography (sEEG) research confirmed the regions' significance, highlighting the middle temporal gyrus, superior temporal gyrus, inferior frontal gyrus, and the frontal part of the cingulate gyrus, as crucial for syntactic processing [7]. However, while fMRI can capture broad syntactic processing, it lacks the necessary temporal resolution to distinguish fine-grained syntactic computations [8].

Electrophysiological studies using event-related potentials (ERPs) have identified specific markers associated with syntactic processing, including the early left-anterior negativity (ELAN) and the P600 components. The ELAN, typically observed around 250 milliseconds post-stimulus, is linked to the initial stages of syntactic structure building and is associated with phrase structure violations. The P600 component, which appears between 500 and 700 milliseconds, is related to syntactic reanalysis and repair, marking a second-pass parsing mechanism [9, 10, 1]. Neuroimaging studies complement these findings by showing that more complex syntactic structures result in greater activation in language-related brain regions, particularly in Broca's area and neighboring frontal regions, highlighting the neural basis of syntactic processing complexities [3].

2.2 Language Models

Language models are a fundamental component of natural language processing (NLP) and understanding. They are designed to capture the statistical properties and patterns of human language, enabling machines to generate and comprehend text. Various architectures have been developed for language modeling, ranging from traditional n-gram models to modern neural network-based models. Language models have many applications, including machine translation, speech recognition, sentiment analysis, and text generation.

In particular, neural network models, such as recurrent neural networks (RNNs) [11], long short-term memory (LSTM) networks [12], and transformers [13], have shown superior performance in capturing complex linguistic patterns and generating coherent text.

2.2.1 Word Embeddings

Word embeddings are dense vector representations of words, capturing semantic meanings based on the context in which words appear. Traditional methods like one-hot encoding represent words as sparse, high-dimensional vectors but do not capture semantic relationships. Embeddings like those from Word2Vec [14] and GloVe [15] aim to represent words in a lower-dimensional continuous vector space where semantically similar words are close to each other.

Popular approaches to generate word embeddings include count-based methods, where word co-occurrence statistics are used to learn word vectors, and prediction-based methods, where a model predicts a word based on its context.

GloVe

GloVe (Global Vectors for Word Representation) [15] combines the benefits of count-based and prediction-based methods. It begins with constructing a co-occurrence matrix, recording the frequency of word pairings within a defined window. By focusing on the ratios of these co-occurrence probabilities, GloVe effectively captures significant semantic relationships.

The learning process uses gradient descent to minimize the difference between the dot product of word vectors and the logarithm of their co-occurrence counts. This approach allows GloVe to learn embeddings that efficiently capture broad contextual information.

GloVe is known for its semantic coherence, computational efficiency, and versatility. Due to its ability to produce rich, semantically meaningful word representations, it is widely used in various NLP tasks, such as text classification, machine translation, information retrieval, and sentiment analysis.

2.2.2 Large Language Models

Large language models (LLMs) are a class of neural network models trained on massive amounts of text data to generate coherent and contextually relevant text. Models such as BERT [16] or the GPT models [17, 18] have demonstrated their ability to generate human-like text and achieve state-of-the-art performance on complex language processing tasks.

LLMs' extensive training allows them to capture intricate linguistic patterns and effectively perform a wide range of NLP tasks, including text generation, completion, question answering, translation, and summarization.

GPT-2

In particular, GPT-2 (Generative Pre-trained Transformer 2) [17] is a large language model (LLM) developed by OpenAI that utilizes a transformer architecture [13]. This architecture relies on self-attention mechanisms to generate text by weighing the importance of different words in a sequence, thus capturing relationships between them. GPT-2 is a decoder-only Transformer model, which generates text autoregressively by predicting the next word in a sequence given the previous words.

GPT-2 is pre-trained on a vast corpus of internet text using unsupervised learning, with the objective of minimizing the negative log-likelihood of predicting the next word. The model exists in various sizes, from small to extra-large, with larger models having more parameters and capturing more linguistic knowledge. GPT-2 XL is an extended version of GPT-2. It consists of 1.5 billion parameters and leverages an increased number of transformer block layers. Its larger size makes it notable for its ability to generate coherent and contextually relevant text.

2.2.3 As Models of Human Language

In addition to their practical applications, LLMs have been studied as models of human language. Some research suggests the potential for LLMs to implicitly learn symbolic and grounded representations, aligning their internal conceptual spaces with human cognitive models. This underscores the importance of understanding the internal representations of these models beyond their task performance alone [19], and is supported by recent evidences that LLMs can learn grounded representations such as space and time [20].

However, other studies highlight the division between formal and functional linguistic competence in these models, suggesting the need for modular architectures to integrate both competencies effectively in a manner akin to human cognition. [21].

Syntax Acquisition

Researchers have explored LLMs' syntactic processing capabilities as an important aspect of language. Experiments on BERT revealed how the model processes language in a structured, layered manner that mimics traditional linguistic pipelines [22]. However, many aspects of BERT's internal workings and the reasons behind its impressive performances remain unclear [23].

Further investigations across language models and syntactic tasks found that LLMs could capture many syntactic phenomena through learning on large text corpora. However, their perplexity scores did not necessarily correlate with human-like syntactic understanding [24]. Instead, LLMs often relied on heuristics rather than full syntactic competence [25]. They suggest, however, that integrating linguistic principles into LLM architectures and examining their internal processes can provide insights into both artificial and human language understanding [25].

Notably, recent findings highlight a sudden drop in training loss in masked language models that coincides with the acquisition of syntactic attention structures essential for grammatical capabilities [26].

The study of LLMs as models of human language not only advances our understanding of these models but also offers insights into the nature of language processing. The continued exploration of their internal processes and alignment with human cognitive models offers valuable insights into the nature of language processing.

2.3 Neural Networks and Neuroscience

Artificial neural networks (ANNs) were initially inspired by the structure and function of the human brain. Early work by Warren McCulloch and Walter Pitts proposed mathematical models of neurons that could perform logical operations [27]. This foundation was further developed with the introduction of the perceptron by Frank Rosenblatt [28] and the backpropagation algorithm by Rumelhart, Hinton, and Williams [29]. These advancements paralleled discoveries in neuroscience, such as Hebbian learning, which describes synaptic strengthening with use [30]. Initially deemed biologically implausible, recent research in deep learning and neuroscience points towards backpropagation-like mechanisms in the brain through feedback connections [31] and cortical dendrites microcircuits [32, 33].

ANNs, inspired by the structure and function of biological neural networks, have shown remarkable capabilities in various cognitive tasks. Recent advancements in deep learning, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer models, have achieved significant success in tasks such as image recognition, speech recognition, and natural language processing. These achievements often draw comparisons to the hierarchical processing found in the brain's visual and auditory cortices [34]. CNNs, for example, were greatly influenced by the receptive fields and hierarchical feature extraction mechanisms of the visual cortex [35]. Moreover, the more recent and highly succesful attention mechanism [36], prevalent in transformer models such as BERT [16] and the GPT models [17, 18], has also found connections to the neuroscience literature through its relation to Sparse Distributed Memory (SDM) [37, 38], a plausible model of the cerrebelum [39], and the observation that neuron-astrocyte networks, widly present across the brain, can perform Transformer-like operations [40].

2.4 Comparing Neural Networks and the Brain

Drawing from these parallels, the comparison between artificial neural networks and the human brain has long been a topic of interest in neuroscience and artificial intelligence research. The impressive successes of ANNs have led to an ever-growing interest in understanding the similarities and differences between artificial and biological neural systems.

Several research studies have explored the relationships between the internal representations learned by ANNs and those occurring in biological brain systems across different species and various tasks. It has been observed that ANNs models can capture hierarchical processing similar to those in the human brain, leading to insights into sensory information processing and integration. To allow comparison between ANNs and the brain, neural data is collected through methods such as functional magnetic resonance imaging (fMRI), electrocorticography (ECoG), or magnetoencephalography (MEG), providing high-resolution measures of brain activity. This neural activity can then be compared or predicted from the representations learned in the artificial neural networks.

Multiple comparison metrics have been developed and widely used to assess the alignment between artificial neural networks and the brain, including Representational Similarity Analysis (RSA) [41], Canonical Correlation Analysis (CCA) [42], Centered Kernel Alignment (CKA) [43], or more recently Soft Matching Distance [44] and Sparse Positive Alignment [45].

2.4.1 Probing

In comparing artificial neural networks (ANNs) and the brain, probing neural networks to predict neural activity has emerged as a powerful approach. Probing is an essential field at the intersection of computational neuroscience and artificial intelligence. Initially introduced in neuroscience, probing allows the investigation of the internal representations of a computational model [46, 47]. It was later extended to the study of ANNs to understand the inner workings of trained models and gain insight into their representations [48, 49].

Probe Design

Different probing designs, including linear probing, are employed to study both ANNs and biological systems. Probes are evaluated based on their ability to decode meaningful information without themselves learning the task, such as decoding part of speech in language models [50]. This concern, often called "decoding from the retina" in neuroscience, underscores the importance of aligning probing techniques with specific research goals [51].

Linear probing, in particular, is favored for its simplicity and high selectivity; they have a reduced risk of memorizing the task rather than revealing genuine structure [49]. It has been used to study representations of deep neural networks in various tasks, such as image classification [49] or decoding linearly encoded information from language models [50, 52, 53].

Probing Neural Activity

By training a linear probe on the hidden representations of a pre-trained neural network to predict neural responses to various stimuli, researchers can evaluate the alignment between the internal representations of the ANN and the brain. Evaluating this alignment on novel stimuli allows to asses the model's ability to capture underlying neural processes. Predicting neural response to stimuli not only provides an alignment metric but also facilitates the identification and creation of specific stimuli for targeted control of neural activity in domains like vision and language processing [54, 55].

The capacity to infer neural activity from ANNs is crucial for advancing sensory and cognitive brain studies, impacting the development of neurotechnological applications such as brain-machine interfaces, neural prosthetics, and brain-computer interfaces. By comparing ANNs representations to empirical neural data, researchers can uncover and refine computational principles underlying neural processes, deepening our understanding of brain functions.

2.4.2 Vision

Initial research on the human and primate visual systems revealed that deep convolutional neural networks (CNNs) could effectively predict neural activity within the visual cortex. This demonstrates their capability to mirror complex biological processes in both static and dynamic visual tasks, capturing essential aspects of brain visual processing [56, 57, 58, 59].

Performance-Optimized Models

Performance-optimized hierarchical models, unlike unsupervised ones [57], showed a strong correlation between model categorization performance and neural response predictability in the inferior temporal (IT) and V4 cortices [56, 58]. Optimized for object recognition, these models presented intermediate representations resembling those in the visual cortex, effectively capturing critical aspects of visual processing [58, 57]. Moreover, DNNs revealed a gradient in representation complexity mirroring the hierarchical organization of the visual cortex, with lower layers corresponding to early visual areas (V1) and higher layers to higher visual areas (IT) [60].

Unsupervised Learning

While deep supervised models align well with human and primate IT recordings, highlighting the importance of task-driven learning, recent studies reveal that novel unsupervised learning methods like SimCLR [**?**] also predict neural responses effectively across the visual cortex. This challenges the necessity of supervised learning and underscores the potential of unsupervised models in mimicking sensory learning [61].

Integrative Benchmark

The introduction of the Brain-Score benchmark in 2018 is crucial for evaluating the brain-likeness of artificial neural networks, emphasizing the need for multiple evaluation criteria and a standard benchmark. It also revealed that beyond a certain accuracy threshold, model performance might diverge from brain patterns [62].

Control of Neural Activity

Using the prediction ability of linear probing, ANN-driven methods have also been utilized to control neural activity in the primate V4 cortex by generating specific luminous power patterns, showcasing potential applications in neuroscience research and therapeutic interventions [54].

Language-Vision Integration

Recent studies on multimodal training involving vision and language indicate that while CLIP models excelled in predicting high-level visual cortex responses [63], controlled comparisons of language-aligned models showed they did not significantly outperform unimodal vision models in predicting ventral stream activity [64]. These findings suggest that, while they are beneficial for behavioral alignment, language models are effective to the extent that they capture object information in images [64, 65].

Findings in Vision

Overall, these studies illustrated that ANN comparison provided a comprehensive framework for predicting neural responses and understanding visual processing in the brain, offering valuable insights into visual perception mechanisms.

2.4.3 Audition

ANNs have also been used to predict the auditory cortex's response. A task-optimized neural network was able to replicate human auditory behavior and predict cortical responses, revealing a hierarchical processing structure within the auditory cortex [66]. Additionally, self-supervised learning models were shown to more accurately reflect both behavioral and brain responses to speech as measured by fMRI, with layers aligning well with the cortical hierarchy of speech processing [67].

2.4.4 Language

Research has extended deep neural networks' success from visual and auditory to language processing. Studies revealed that word embeddings were linearly correlated with brain activity [68, 69], and that models like LSTM [12] offered enhanced predictivity through better embeddings and contextual information [70]. The exploration has advanced into large language models (LLMs), particularly transformer models, assessing their alignment with human brain activity across varied tasks and measurement techniques.

Model Comparison

Research comparing 43 language models, including embedding models, recurrent neural networks, and transformers like BERT and GPT, found that advanced models like GPT-2 XL closely approached the noise ceiling in predictivity with brain data from fMRI and ECoG recordings across auditory and visual language responsive areas. These models also accurately predicted human reading times, indicating a strong relationship between neural responses and observable behaviors [71].

Another comparative study involving CNNs, word embeddings, and transformers with fMRI and MEG data showed peak correlations in different brain regions, CNNs in early visual areas, word embeddings in the left temporal and frontal cortices, and transformers in regions associated with sentence comprehension [72].

Both groups observed a strong correlation between a model's ability to predict the next word in a sequence and its alignment with the brain, suggesting a shared predictive objective [71, 72].

Insights with GPT-2

Additional research demonstrated GPT-2's high correlation with brain responses and its ability to uncover aspects of language processing challenging to study in model-free approaches [73], making it a preferred model for probing language processing in the brain [74, 75, 76, 77, 78, 55]. Studies showed that GPT-2's intermediate layers were particularly effective in predicting brain activity [72, 76].

Moreover, the model's prediction score was strongly aligned with the subject's comprehension of stimuli, particularly in areas associated with high-level language processing and the deeper layers of the GPT-2 model [74]. Its embeddings could predict activity in both speakers and listeners during face-to-face conversations, emphasizing a shared linguistic space [77].

Predictive Coding in the Brain

Goldstein et al. expanded upon predictive processing in the human language network, suggesting that the brain processes language similarly to autoregressive models by anticipating stimuli to minimize surprise, a concept rooted in findings from earlier studies [79]. They utilized ECoG recordings from participants listening to spoken narratives, comparing static word embeddings like GloVe with contextual embeddings from GPT-2, and observed distinct patterns of pre-onset prediction and post-onset surprise, aligning closely with prediction-error signals in LLMs [71, 74, 75]. Notably, neural activity encoded upcoming words up to 800 ms before their onset, with a marked increase in activity 400 ms after onset for unpredictable words, particularly in the IFG, which exhibited enhanced neural patterns when predicting linguistic information.

Caucheteux et al. further supported these findings using fMRI data from participants listening to short stories, demonstrating that long-range predictions enhance model correspondence with brain activity, especially in higher cognitive regions like the frontoparietal cortex, suggesting a hierarchical predictive coding system within the brain [78].

However, Antonello et al. challenged the centrality of predictive coding, suggesting that the efficacy of LLM models might also stem from their capacity to capture a broad array of linguistic features, not solely their predictive properties. They argue that while predictive information might be present in the brain, it may not be the primary driving factor [80].

Hierarchical Similarities

In addition to the predictive coding work, Goldstein et al. suggest that the layered hierarchy of GPT-2 XL mirrored the temporal structure of language processing in the human brain. They observed a strong correlation between the depth of the model layers and the timing of neural activity across language-related brain regions, suggesting that the brain processes language hierarchically, with different cortical areas activated at specific times during comprehension, in a similar way to large language models [76].

Building on this observation, Mischler et al. further investigated the correspondence between LLMs' layer hierarchy and human brain structure. They examined 12 high-performance LLMs, finding that models with better benchmark performance showed higher activity correlation and mapped onto brain pathways within fewer layers [81].

Control of Neural Activity

Like in vision, studies using GPT-2 XL have shown potential for modulating neural activity in language networks through aligned model responses to diverse linguistic stimuli, revealing the control potential over neural activity [55]. Surprisal and linguistic well-formedness significantly influenced the strength of neural responses.

Syntax and Semantics

Model-based studies, like those using recurrent neural network grammars (RNNGs), have shown that they can predict brain responses to syntactic and semantic violations, featuring early and P600-like peaks [82], reflecting the brain's processing capabilities [10, 1].

Caucheteux et al. used fMRI data from narrated texts to examine how GPT-2 activations across lexical, compositional, syntactic, and semantic classes correlate with brain activity. They discovered that compositional representations involved a broader cortical network than lexical ones and that syntactic and semantic processing shared common neural substrates, supporting a distributed approach to brain syntactic processing [83].

Kauf et al. also analyzed the impact of lexical semantic content versus syntactic structure on brain activity predictions using GPT2-XL, finding that lexical semantic content, predominantly carried by content words, played a more crucial role than syntactic form, underscoring the human language system's focus on meaning derivation and the potential of deep learning models to mimic this process [84].

Findings in Language

Just as in vision and audition, these studies collectively underscore the potential of LLMs in revealing local and temporal structures in the brain's language processing and comparing large language models' coding mechanisms and layered hierarchy to the brain's language processing mechanisms. This model-based approach, allowed by the recent advances in LLMs, offers a promising approach to understanding language processing dynamics that were challenging to explore with traditional methods.

2.4.5 Additional Modalities

Recent studies have expanded this line of work to include proprioception in monkeys [85] and multimodal integration in the human brain [86], highlighting the versatility and broad potential of ANNs in modeling diverse neural processing aspects with high fidelity.

2.4.6 Future Advances and Challenges

The use of ANNs extends beyond machine learning and artificial intelligence, serving as crucial tools for empirical exploration and theoretical understanding of brain mechanisms. The ongoing development of ANN models is expected to enhance our understanding of the brain's complex processes, providing deeper insights into sensory processing, cognitive functions, and advances in neurotechnological applications.

Han et al. note, however, that the efficacy of artificial neural networks (ANNs) in modeling brain computational mechanisms does not necessarily mean that the underlying brain architectures correspond to those of ANNs. By comparing various architectures, they found significant variability in performance, influenced more by stimuli type than the model's architecture, underscoring the need for a refined evaluation of model alignment and a deeper investigation across architectures [87].

Chapter 3

Methods

3.1 Task Paradigm

3.1.1 Design

The proposed task paradigm involves recording neural activity using stereoencephalography (sEEG) in epileptic patients as they process short sentences either auditorily or visually. Participants are exposed to three sentence structures, some containing a semantic or syntactic violation through the inversion of two words.

Each sentence is four words long and exists in three different structures: GS, GnS, and nGnS. Each category explores different manipulations of sentence components. Each sentence under consideration is constructed from four words and presented in all three formats to highlight distinct grammatical and semantic characteristics.

GS (Grammatical Sentence) adheres to conventional grammar rules and semantic coherence, presenting sentences that are both syntactically correct and meaningful, typically following the Subject-Verb-Object (SVO) order. For example, "The girls ate cakes" is a standard, coherent sentence in this format.

GnS (Grammatical Nonsensical Sentence) retains grammatical correctness but disrupts semantic coherence by inappropriately swapping the roles of nouns within the structure. This leads to structurally sound but semantically illogical sentences, such as "The cakes ate girls".

nGnS (Non-Grammatical Nonsensical Sentence) breaks both grammatical norms and semantic logic, resulting in sentences that neither conform to standard syntax nor make logical sense. An example would be "The ate girls cakes", which challenges comprehension and interpretation due to its scrambled structure.

There are 151 different sentences presented through audio and visual modalities in a randomized order to avoid order effects. The task is divided into blocks, each containing a set of sentences. The auditory stimuli are presented through headphones, while the visual stimuli are displayed on a screen. Each word is presented with an 875ms interval, and a 1000ms pause follows the sentences.

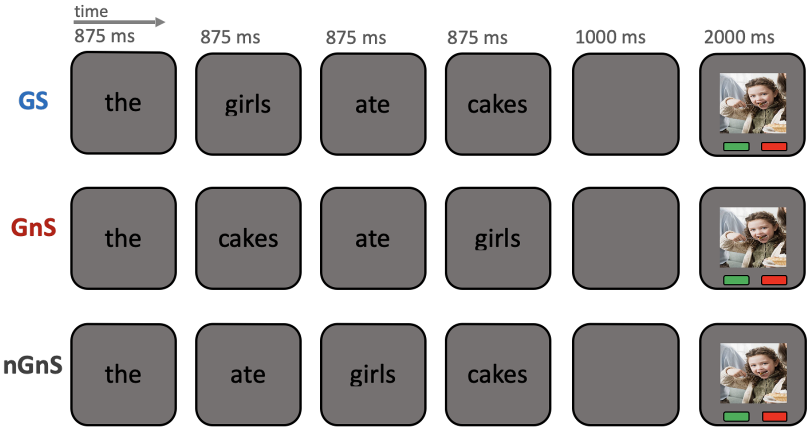


Figure 3.1: Task Paradigm: Example of Sentence Structures

The participants are instructed to listen to or read the sentences carefully and, after an interval of 1s, are presented with an image. If the sentence falls under the GS (Grammatical Sentence) category and accurately matches the image's content, subjects are instructed to press a green button, otherwise a red button. This step evaluates the subjects' comprehension of a grammatically correct and semantically coherent sentence and its correspondence to a visual representation.

3.1.2 Importance of the Study

The proposed task paradigm is important for understanding the neural mechanisms underlying language processing in the human brain. By presenting each sentence in three different formats, the task is designed to elicit distinct neural responses, allowing the identification of specific components associated with semantic and syntactic processing.

Employing sEEG offers high-resolution insights into the dynamics of syntactic processing across different sentence structures and modalities, allowing fine temporal and local analysis of neural activity. This information is essential for developing models of language processing and understanding how the brain comprehends language.

This work uses large language models to predict neural activity recorded during the task. Extensive literature has shown that neural activity can be predicted from word and sentence embeddings. We aim to build on these findings by predicting neural activity in stimuli containing semantic and syntactic violations. We hypothesize that neural activity might differ between the different sentence structures and that a model-based approach will allow us to understand better the interactions of semantic and syntactic processing in the brain, identifying temporality and regions involved in semantic and syntactic processing.

3.2 Data

3.2.1 Collection

Stereo-electroencephalography (sEEG) is an invasive neurophysiological method that records electrical activity directly from the cerebral cortex. It is particularly used for patients with intractable epilepsy, who are candidates for surgical treatment. sEEG involves the implantation of electrodes into targeted areas of the brain to identify the regions responsible for the onset of epileptic seizures, helping differentiate epileptic foci from essential brain areas involved in important functions like speech, motor skills, and sensory processing. By providing high temporal and spatial resolution, sEGG is highly suitable for studying neural activity during specific cognitive tasks, such as language processing.

This study collected data from 926 electrodes in 10 English-speaking patients performing the language processing task. The sEEG recordings provide high temporal and spatial resolution, allowing precise brain activity mapping during language processing. The electrodes were implanted in various brain regions according to clinical criteria for each participant, and the signals were recorded continuously during the task at a sampling rate of 1000 Hz. Experiments were conducted at the Cleveland Clinic, where the study was approved by the hospital's institutional review board and carried out with the participants' informed consent.

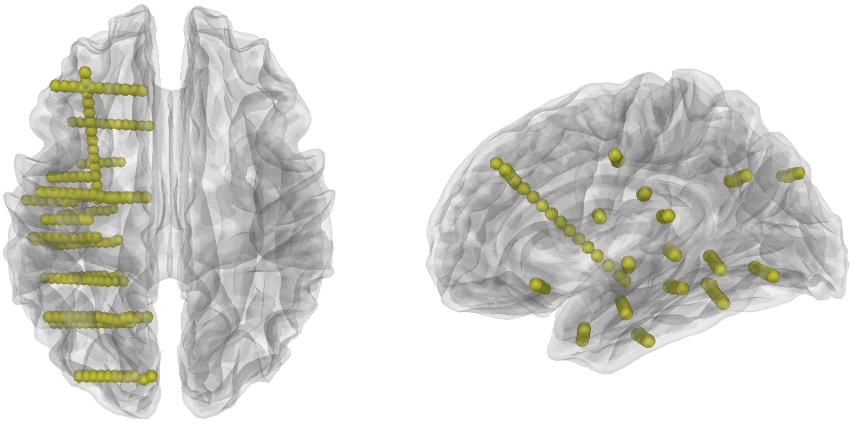
Event markers were used to synchronize the onset and offset of stimuli with the recordings. Distinct electrical spikes were used to mark fixation, words, and image onsets, allowing for precise alignment of neural activity with the task events.

3.2.2 Preprocessing

Initially, electrodes positioned at locations identified with epileptic activity are excluded to avoid biased or distorted data. Additionally, electrodes showing flat signals and those contaminated with artifacts are removed from the dataset.

A bipolar montage is employed to enhance signal quality and minimize external noise. This setup uses adjacent electrodes to measure potential differences, effectively reducing common noise and enhancing the local field potential signal's clarity. Once these procedures are applied, the ECoG signal is cropped to the task period to focus the analysis on relevant data.

We follow the ECoG processing procedure of Goldstein et al. [75] for the following steps. We first remove outlier values, identified as data points lying beyond three interquartile ranges (IQR) from the 25th to the 75th percentile, and interpolate their values using Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) from the Scipy library to maintain continuity in the time series.



(a) Superior View (b) Left Lateral View

Figure 3.2: Electrode Locations in the Brain

Using MNE [88], we filter out the 60 Hz line noise and its harmonics using a fifth-order Butterworth filter, typical interference frequencies in electrical recordings. This is followed by a bandpass filter between 70 and 150 Hz to isolate the high gamma range. This frequency band has been shown to correlate positively with local neural firing rates, suggesting active brain regions involved in the task [89]. The signal's power spectral density is computed using Morlet wavelets to analyze the dynamics and amplitude of brain activity during the task. We additionally take the logarithm of the power to stabilize the signal variance and finally normalize the high-gamma power across the entire recording [90, 75].

3.2.3 Feature Extraction

Neural Activity

In this exploratory work, we study the recordings from a single patient in which electrodes have been identified to be responsive to language processing from audio and visual modalities by Misra et al. (manuscript in preparation). After bipolar referencing and removing faulty electrodes, the dataset consists of 190 electrodes, providing extensive coverage in the left hemisphere. Coverage encompasses the frontal, parietal, and temporal lobes, including the amygdala and hippocampus. It also covers the insula and cingulate cortex. (Fig. 3.2).

Electrical spikes corresponding to task event onsets are used to precisely align the sEEG recordings and obtain event-locked neural activity. To allow for temporal analysis of the neural activity, the ECoG signal is segmented into 200 ms time windows, shifted by 100 ms, from 200 ms before the onset of the word to 1000 ms after the onset. This results in 12 time points per sentence, which we aim to predict from the sentence embeddings. This integration over time windows is consistent with previous studies, and the choice of window size and shift has been shown to have minimal impact on the regression results [75, 86]. Through this temporal analysis, we expect to capture the dynamics of semantic and syntactic processing in the brain happening from word onset.

3.3 Linear Probing

Following the extensive literature on probing neural activity from language models, we adopt a linear probing approach [74, 75, 55, 84], in particular ridge regression [70, 77, 55, 86, 81].

3.3.1 Glove Embeddings

Our first predictor variables are Glove embeddings, which are already trained and available in the Glove library [15]. Importantly, all the words in the task have a corresponding embedding. Following Goldstein et al. [91], we use the 50-dimensional embeddings. As shown through multiple studies, the different embedding sizes do not significantly impact the model's performance [75, 86].

3.3.2 GPT-2 Embeddings

For the model-based analysis, we use the GPT-2 XL model from the GPT-2 model family [17], which has shown high alignment with neural data [71, 74, 75, 55]. We use the Hugging Face Transformer library to extract the embeddings from the GPT-2 model [92].

Following literature showing that middle layers of transformer models perform best in predicting neural activity [74, 76, 81], we use the 22nd layer of the GPT-2 XL model [55] and extract the embeddings corresponding to the last sentence token [71, 55]. These embeddings are 1024-dimensional, which we use as our predictor variables to regress neural activity.

From the GPT-2 XL model, we define two settings. The first consists of single words only, where we give the model the current stimulus word. The second consists of the sentence up to the current stimulus word, providing context. As observed in the literature, the sentence embeddings perform better than word-only embeddings [75, 74].

To improve computational cost, we cache the activations from the model to avoid inferring the embeddings at each iteration. We cache the activations for the 151 sentences in the dataset, in each of the three orders, and for each word, allowing us to quickly access the embeddings when fitting the probe.

3.3.3 Ridge Regression

Following widely used methods in the literature, we use ridge regression to predict neural activity from the defined predictor embeddings [75, 86, 81]. We expect the neural activity to be linearly related to the embeddings, and ridge regression accounts for the high dimensionality of the sentence embeddings and the potential multicollinearity between the features.

We predict neural activity using three types of embeddings: the Glove embeddings, the GPT-2 embeddings with the current word only, and the GPT-2 embeddings with the sentence context.

The response variable is the previously extracted high gamma power at each word onset, integrated over 12 time points from 200ms before the onset to 1000ms after the onset. This multi-time point approach means we fit 12 models per embedding type, one for each time point. We note that correlation is individually estimated at each time point and does not account for the correlation in time of the signal.

Ridge regression minimizes the following loss function:

$$\min_{\beta} ||y - X\beta||^2 + \alpha||\beta||^2 \tag{3.1}$$

where X is the word or sentence embeddings, y is the high gamma band power at a time point, and α is the regularization parameter controlling the amount of shrinkage applied to the coefficients. For α, we explore 8 values from 10^0 to 10^6 in a logarithmic scale.

To select an optimal value α, we use a 5-fold cross-validation procedure and keep the value that minimizes the average mean squared error (MSE) of the predictions before fitting the model on the whole training set and predicting the activations on a test set.

We normalize predictor variables X and response variables y on the training set statistics before fitting each ridge model to ensure that all variables contribute equally to the penalty term and prevent bias towards variables with larger scales.

These procedures are implemented in Python using the PyTorch library [93] to benefit from GPU acceleration and improve computational time on the large number of regressions we fit.

3.3.4 Performance Metrics

In selecting the optimal α value, we use MSE as the performance metric, in line with the ridge regression loss function.

To evaluate the model's performance, we use the Pearson correlation coefficient between the predicted and actual activations. This follows the literature that has shown that the Pearson correlation is a good and comparable measure of the alignment between a model and neural activity [62, 71, 72, 75, 86, 81].

The Pearson correlation coefficient measures the linear relationship between two variables, ranging from -1 to 1. It is calculated as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.2}$$

A value of 1 indicates a perfect positive linear relationship, -1 a perfect negative linear relationship, and 0 no linear relationship. This allows a comparable measure of the alignment between the model and the neural activity across different conditions and studies.

3.3.5 Electrode Selection

Preselection of language-responsive electrodes can be approached with different methods, such as comparing sentences to random word lists and non-word lists [71], which we do not have access to in our dataset. However, identification of stimulus-responsive is also employed through a comparison of neural activity during word presentation to a fixation or silence baseline [71, 81]. Goldstein et al. use a different approach by preselecting electrodes that show alignment with the Glove embeddings [75]. Despite showing a higher alignment with the GPT-2 model across electrodes, this selection method could, however, present the risk of a biased electrode selection aligned with the Glove embeddings.

In this exploratory work with a limited number of electrodes, we follow Subramaniam et al. and avoid preselecting electrodes explicitly [86]. We aim to find significant electrodes solely based on the model's performance and draw conclusions from the observed alignment.

3.3.6 Assessing Significance

Assessing the significance of the model performance is crucial to ensure meaningful results are observed, but it is not a straightforward task.

On fMRI and MEG data, the noise ceiling, i.e., the maximum alignment we can expect on a subject from an external predictor, is usually assessed by predicting the neural response of a patient from the neural response of all other patients [74, 71]. However, with ECoG or sEEG data, predicting neural response at the electrode level from subjects is not feasible due to the fine spatial resolution of the data and the disparity in the placement of the electrodes between subjects.

The ceiling is usually estimated at the electrode level [71] or not assessed [75, 76, 86].

Multiple specificities in our dataset and research objectives make the computation of a noise ceiling difficult. This includes limited word repetition under each task condition. Moreover, using temporal windows instead of a single average activity value at the word level makes the computation of a noise ceiling more complex. Additionally, we observed high variability in the model's performance depending on the test set, likely due to the small dataset size under certain conditions. These observations motivate us to explore resampling methods instead to estimate the significance of the model alignment, similar to Subramaniam et al. [86].

Following Subramaniam et al., the first approach consists of resampling the test set for each prediction. We rerun 1000 iterations of the model to get a stable estimate of the model's performance and a confidence interval for the performance. For each iteration, we randomly select 80% of the data for training and 20% for testing. Following Subramaniam et al., we identify interesting electrodes based on the mean performance across the 1000 iterations and the 95% confidence interval, evaluating an electrode as significant if the 95% confidence interval does not include 0 [86].

3.3.7 Train and Test Sets

For each data split, the test set contains 80% of the data for training and 20% for testing. To offer comparable results between models, modalities, and task conditions, we cache the 1000 random splits of the test set to ensure reproducibility. Moreover, we match the sentences between the task conditions to ensure each split contains corresponding sentences across three different structures. This allows us to control the sentences' effect on the regression performance. Additionally, when training on mixed task conditions, the same sentence will not be present in both the training and test set, which risks artificially inflating the estimate of the model's generalization performance.

We note that control is not done at the word level, with the same word potentially present in both the training and test set. The limited dataset size and the imbalanced representation of words in the sentences make defining representative splits at the word level challenging, potentially introducing a bias in the model performance. This procedure, which is not found to be controlled for in the literature [71, 91], could, however, represent a risk of overestimating the model generalization performance.

Chapter 4
Results

4.1 Combined Analysis

Our first analysis aims to investigate the ability of the models to predict neural activity based on various combinations of words and sentence types. We aim to understand how well the models can perform depending on the data combination. This approach echoes Schrimpf et al. work on ECoG data [71], where they predict neural activity at each word onset in the context of independent eight-word sentences from Fedorenko et al. [94]. In our setting, given that the first sentence word is always "The" and offers no variation in word or sentence embeddings, we do not include it in the analysis.

Moreover, given possible differences in time display and processing for the vision and audio stimuli and the models' inability to differentiate between them, we keep the two modalities separated. We aim to identify modality-specific alignment and offer a comparison between the two.

The explored combinations are the following:

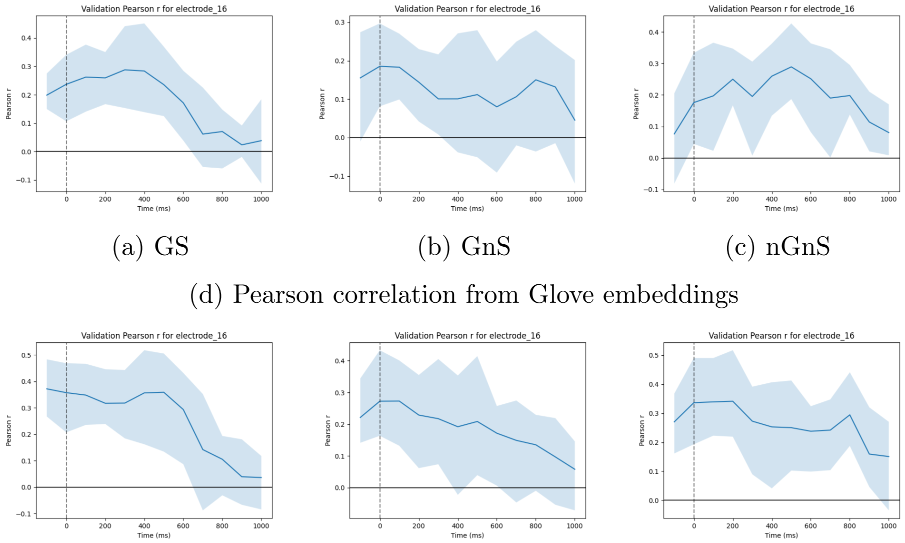
- W2 (Word 2), W3 (Word 3), W4 (Word 4), W2 $\cup$ W3 $\cup$ W4
- GS, GnS, nGnS, GS $\cup$ GnS $\cup$ nGnS
- AUD (Audio), VIS (Visual)

This represents 32 combinations for the 190 electrodes and the three models. Following Goldstein et al. [75] and the resampling approach from Subramaniam et al. [86], we preselect electrodes from the Glove embeddings. In this initial exploratory analysis, the computational needs resulting from the many combinations motivate us to perform only 100 bootstrap samples. At this point, we identify multiple electrodes with significant activation for some condition according to our definition (the 95% percentile of the predicted test sets does not intersect with 0).

We observe that models trained on combined words often showed a very high correlation with the neural activity, providing encouraging results in line with previous studies in the setting of fixed-length sentences [71]. Aiming to understand the underlying mechanisms driving the alignment, we explore some electrodes in more detail.

4.1.1 Electrode 16 - Left Middle Frontal Gyrus

We start our analysis with electrode 16 (Fig. A.1), located in the left middle frontal gyrus, which plays an important role in literacy [95]. As can be seen in the case of audio stimuli in 4.1, the correlation score between the predicted and actual activations is particularly high for the combined word condition.



(a) GS (b) GnS (c) nGnS

(d) Pearson correlation from Glove embeddings

(e) GS (f) GnS (g) nGnS

(h) Pearson correlation from GPT-2 XL sentence embeddings

Figure 4.1: Electrode 16, correlation with neural activity on combined W2, W3, W4 audio stimuli.

Exploring the average neural activity of the electrode at stimuli onset, we observe a ramping up of the activation throughout the trials, both in the case of audio and visual stimuli (Fig. 4.2). This suggests that the area integrates information over the trial duration, coherent with the functions of the middle frontal gyrus. This also aligns with observations in the literature, in particular, the eight-word sentences dataset from Fedorenko et al. [94] used by Schrimpf et al. [71]. Comparing the average activity at each word onset, we observe a clear difference in mean activity resulting from the ramping behavior (Fig. 4.3).

To investigate the impact of position versus content in driving the alignment between models and neural activity, we perform a permutation test designed to remove the semantic and syntactic content provided to the model while keeping

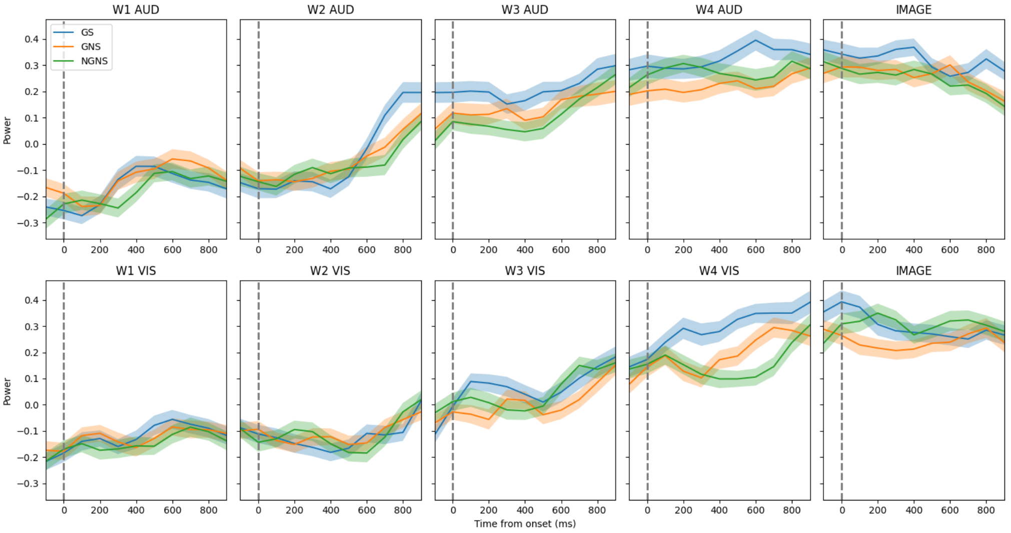


Figure 4.2: Electrode 16, mean and standard estimate of the mean at stimuli onsets.

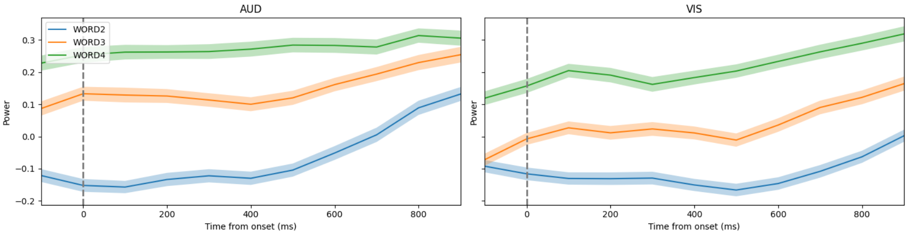
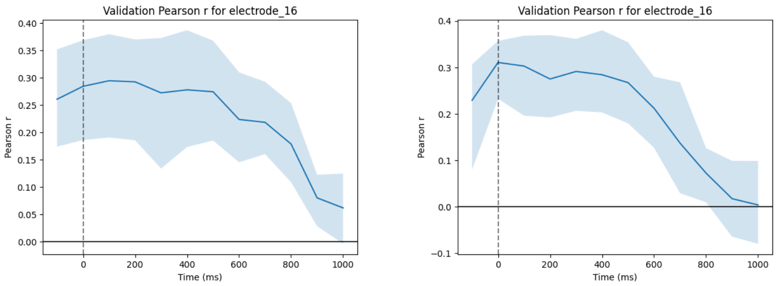


Figure 4.3: Electrode 16, mean and standard estimate of the mean between words.

consistency between the stimuli within a trial and preserving the word order. The procedure consists of shuffling sentences between the trials, where the context for W2, W3, and W4 of a trial is replaced by W2, W3, and W4 of another trial.

In Fig. 4.4, we observe that the correlation is mostly preserved, suggesting that word position in the sentence is a strong predictor of neural activity. However, this also means that the correlation driven by the sentence content might be minimal. However, the minor decrease in correlation observed, particularly in the late period, might be linked to this semantic and syntactic information or word-level self-consistency.



(a) Normal embeddings

(b) Shuffled across trials

Figure 4.4: Electrode 16, correlation with neural activity on combined W2, W3, W4 and combined GS, GnS, and nGnS audio stimuli from GPT-2 XL sentence embeddings.

4.1.2 Electrode 127 - Left Precentral Gyrus

This ramping behavior can also impact model alignment across modalities. Analyzing electrode 127 (Fig. A.2) in the left precentral gyrus, we observe that neural activity is connected to the visual modality, where spikes in activity can be observed at visual stimuli onsets, including words and image presentation (Fig. 4.5). Inspection also reveals an increase in activity prior to the presentation of visual stimuli, coherent with the role of the precentral gyrus in motor planning and execution, demonstrated to induce eye movements [96].

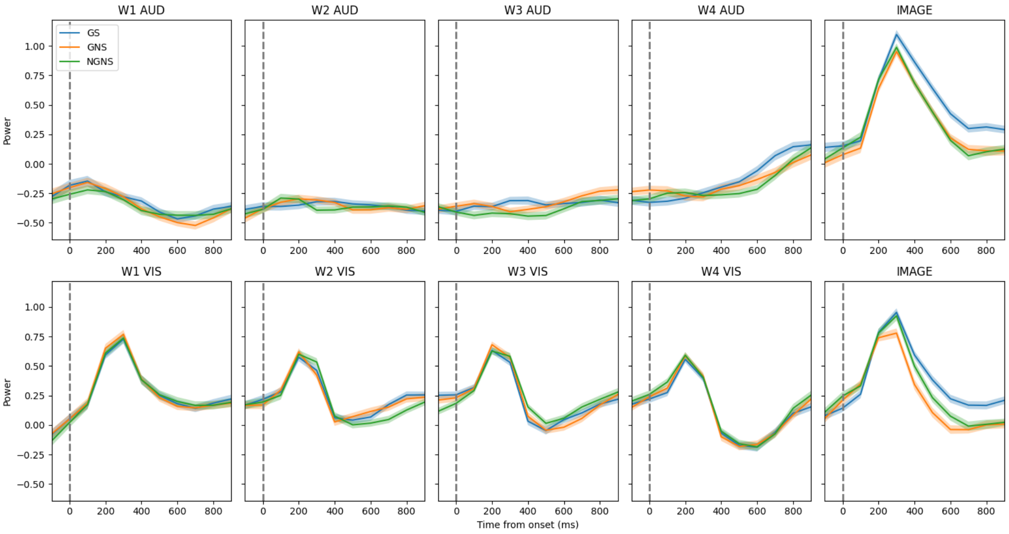


Figure 4.5: Electrode 127, mean and standard estimate of the mean at stimuli onsets.

This preparation behavior significantly increases the activity in word 4, compared to the other words, as shown in Fig. 4.6. Comparing the alignment score between the models and the neural activity, we observe a strong correspondence between the distance in mean neural activity in each word and the correlation score (Fig. 4.7). This correlation is again preserved in the sentence shuffling test, suggesting the correlation derives again from the position of the words and the observed ramping behavior, with an impact across modalities resulting from the task design.

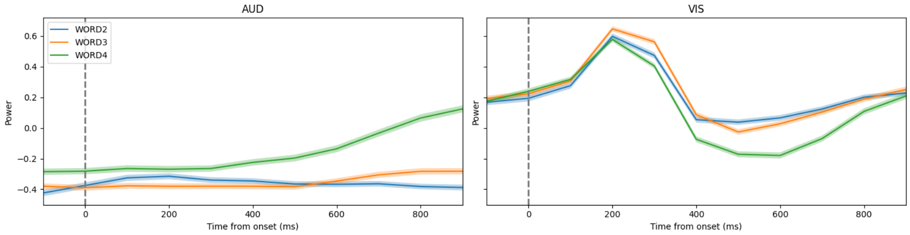
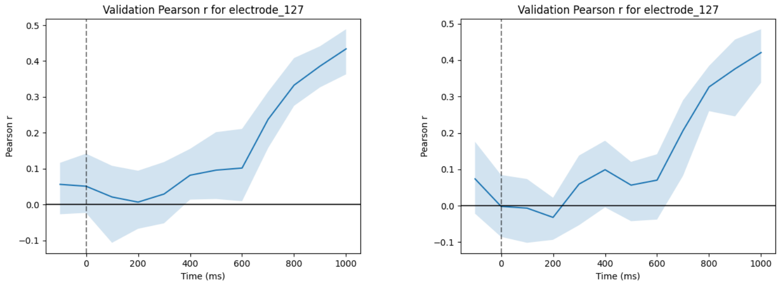


Figure 4.6: Electrode 127, mean and standard estimate of the mean between words.



(a) Normal embeddings (b) Shuffled across trials

Figure 4.7: Electrode 127, correlation with neural activity on combined W2, W3, W4 and combined GS, GnS, and nGnS audio stimuli from GPT-2 XL sentence embeddings.

4.1.3 Electrode 131 - Left Precentral Gyrus

Present on the same sEEG strip as electrode 127, electrode 131 (Fig. A.3) is also located in the left precentral gyrus. In addition to displaying a high alignment score in the combined word analysis, this electrode also shows a strong correlation with neural activity at the word level, implying limited importance of word position on the alignment score. Exploring neural activity, we observe that both modalities contain spikes in activity during word presentation but not during image presentation (Fig. 4.8). This suggests that the precentral gyrus at this location might be involved in high-level language processing, which aligns with findings that the precentral gyrus is involved in silent reading [97].

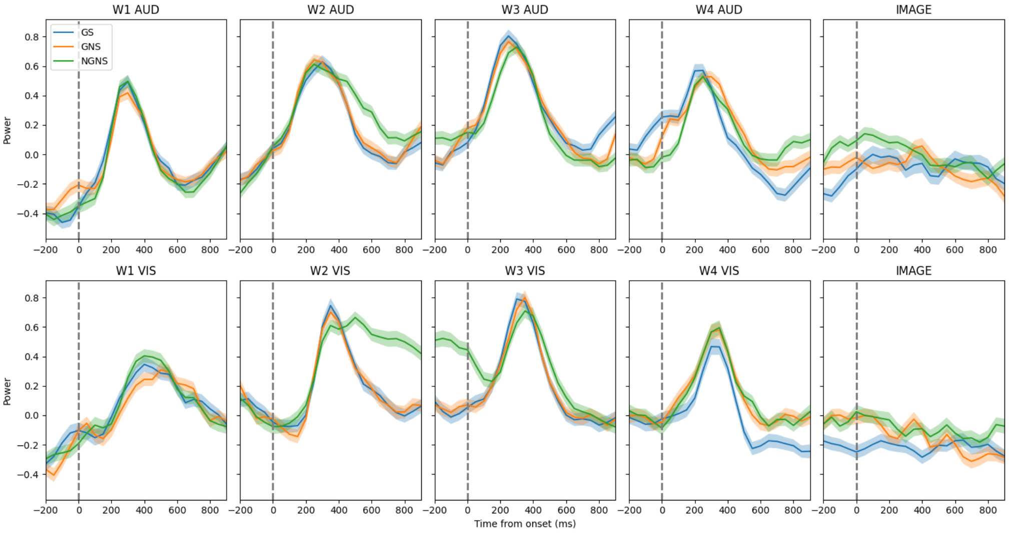
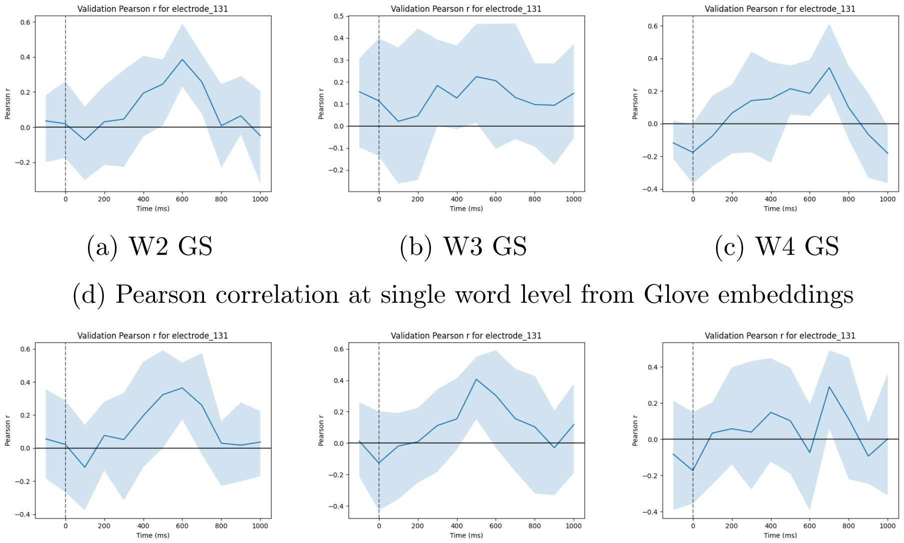


Figure 4.8: Electrode 131, mean and standard estimate of the mean at stimuli onsets.

Interestingly, we observe a clear difference in neural activity at W2 in the nGnS sentences. As a reminder, the nGnS condition is the only grammatically incorrect, where the subject noun at W2 is inverted with the verb at W3. This difference in neural activity in both modalities, when presented with a lexical violation in the form of a misplaced verb, bears strong similarities with the P600 effect, a positive deflection in the event-related potential (ERP) waveform typically observed in response to syntactic violations, such as an anomalous verb. It starts around 500 ms, peaks around 600 ms, and lasts at least 500 ms [98]. The P600 effect is thought to reflect the reanalysis of the sentence structure, suggesting that the left precentral gyrus might be involved in syntactic processing and reinforcing our hypothesis that the precentral gyrus is involved in high-level language processing. We leave further investigation of this effect for future work.

In this electrode, we also show a significant alignment score in the single-word, single-sentence type analysis, particularly under the GS condition (Fig. 4.9). In this particular setting, sentence structure is identical across every trial. Thus, correlation cannot derive from word position nor part of speech or syntax processing, suggesting that the linguistic content itself might drive the alignment. In this scenario, we further account for the self-correlation of neural activity at the word level by partitioning the data to ensure that a word is not included in both the training and test sets. Despite an observed drop in correlation (Fig C.1), which might be due to imbalanced word representation in the data and variability of the split size, the alignment score remains significant.

This provides strong evidence that the precentral gyrus is involved in high-level language processing and that the models can capture the variance of its activity during language processing.



(a) W2 GS (b) W3 GS (c) W4 GS

(d) Pearson correlation at single word level from Glove embeddings

(e) W2 GS (f) W3 GS (g) W4 GS

(h) Pearson correlation at single word level from GPT-2 XL sentence embeddings

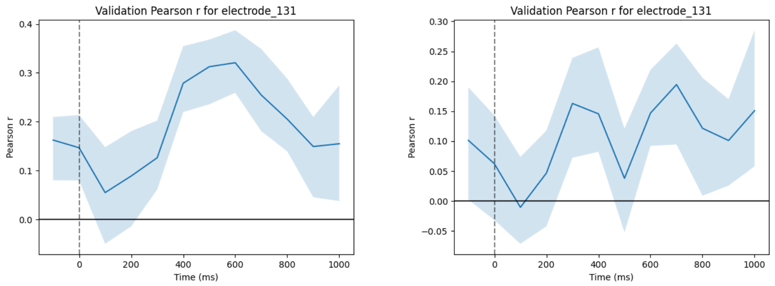
Figure 4.9: Electrode 131, correlation with neural activity at single word level.

Additional verification with the sentence shuffling test shows that the correlation drops significantly when the sentence content is shuffled across trials. This suggests that the sentence content drives a significant part of the alignment (Fig 4.10). We note, however, that the correlation is not entirely lost, suggesting that the word position in the sentence still plays an important role in the alignment score.

4.1.4 Motivation for Further Controlled Analysis

Our observations suggest that, in alignment with previous studies, the models can capture the neural activity corresponding to the processing of the sentence content. In particular, the alignment scores observed in electrode 131 suggest that the models can do so using information from the word and sentence semantic content.

Despite these encouraging results, we observe that the alignment score often derives from factors other than linguistic content. This comes in addition to the high number of possible combinations and the high variability in the data, proving challenging to draw clear conclusions. The illustrated ramping dynamic observed in electrodes 16 and 127, sometimes even impactful across modalities,



(a) Normal embeddings

(b) Shuffled across trials

Figure 4.10: Electrode 131, correlation with neural activity on combined W2, W3, W4 and combined GS, GnS, and nGnS audio stimuli from GPT-2 XL sentence embeddings.

would not be captured across models and conditions equally and could lead to misleading conclusions. In particular, the observed ramping behavior in the visually responsive electrode 127 is critical, as it could lead us to wrongly conclude that the electrode is language-responsive in the audio modality.

This motivates us to explore alignment in a more controlled setting, where we can isolate hidden factors in the alignment score and draw more robust conclusions from the results.

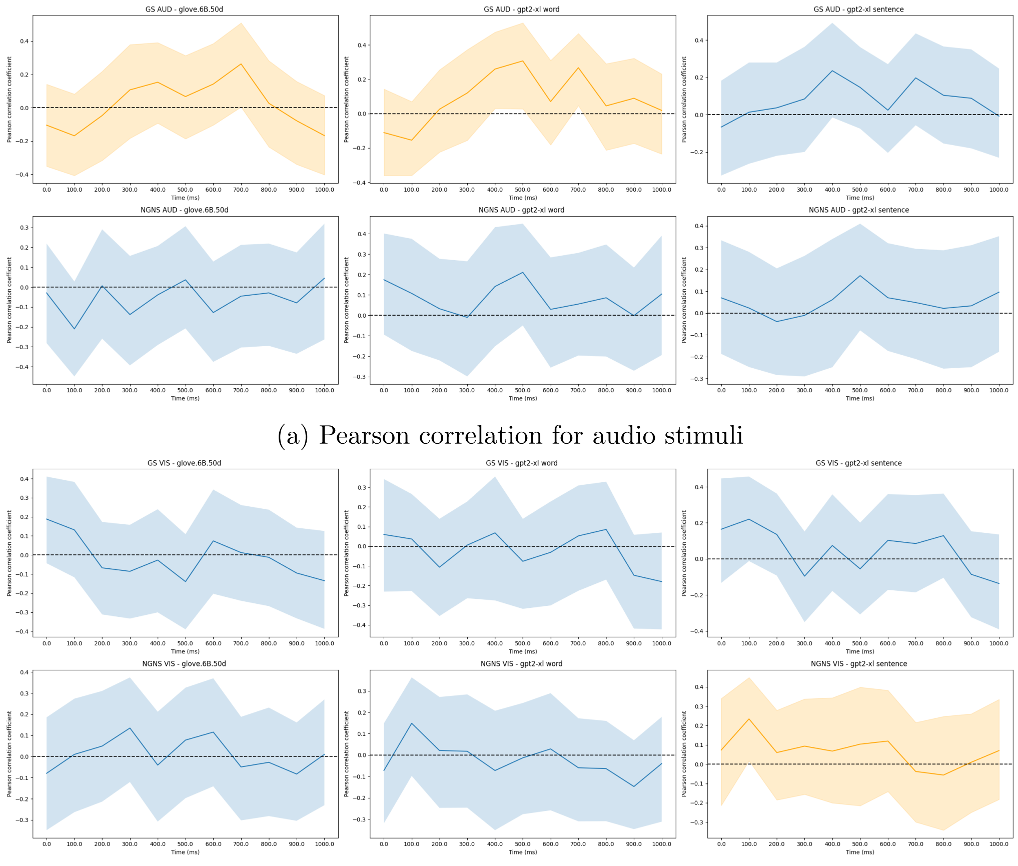
4.2 Word 4 Analysis

We focus on the fourth word of the GS and nGnS sentences to allow for better control over our experimental conditions. Under those two conditions, the fourth word is identical, and any observed difference in alignment can be attributed to the impact of sentence structure and its integration on neural activity. This setting also effectively controls for the impact of word position in the sentence that we observed in the previous section. Moreover, as the final word in the sentence, W4 provides the longest and most comprehensive context for the GPT model to predict neural activity. This setting should allow for a more controlled analysis of the impact of the predictor model and the sentence type on the alignment score.

In this section, we compare predictions from the Glove and GPT-2 XL embeddings, word-only and with sentence context, at W4 in separated GS and nGnS sentences. We aim to identify local and temporal differences in the prediction of neural activity resulting from models and sentence types.

4.2.1 GS and nGnS Comparison

The first step in our analysis compares the prediction of the neural activity for the fourth word for the GS and nGnS sentences separately. We again follow Subramaniyam et al. approach to assess the significance of the correlation [86]. This allows us to identify electrodes for which models significantly predict neural activity under some conditions. This gives us 12 correlation scores for each electrode across the two modalities, two sentence types, and three models. Under this procedure, electrode 131, theorized to be involved in high-level language processing, is detected as significant across both modalities (Fig. 4.11). We note that the correlation scores are not identical to the previous section as we use 1000 bootstrap samples instead of 100 here to provide a more robust correlation estimate. This allows to account for the small sample size and high variability in the data and helps reduce the risk of false positives.



(a) Pearson correlation for audio stimuli

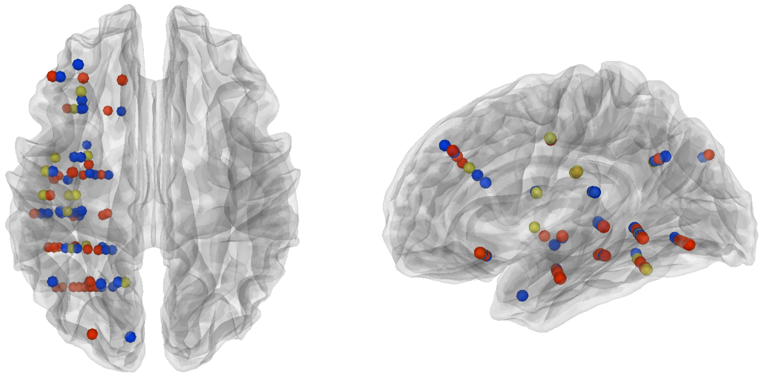
(b) Pearson correlation for visual stimuli

Figure 4.11: Correlation for electrode 131. The shaded area represents the 90% confidence interval. Orange indicates that more than 95% of the bootstrap samples are above zero at any timestamp.

Fig. 4.11 illustrates that in some cases, the GPT-2 XL model with sentence embedding does not outperform the Glove or GPT-2 XL model with word embedding. Correlation across timestamps seems to follow common trends across models, but high variability in the results is expected to make fine-grained comparisons challenging.

4.2.2 Modality Responsive Electrodes

Based on our results, we identify modality-responsive electrodes, for which the models significantly predict the neural activity in some conditions. This provides us with 33 audio-only responsive electrodes, 32 visual-only responsive electrodes, and 13 electrodes responsive to both modalities (Fig. 4.12).



(a) Superior View

(b) Left Lateral View

Figure 4.12: Modality responsive electrodes. Blue indicates audio-only responsive electrodes, red indicates visual-only responsive electrodes, yellow indicates electrodes responsive to both modalities.

The location of modality-responsive electrodes is the following:

- Audio only: Amygdala, Insula, Medial Orbitofrontal Cortex, Middle Temporal Gyrus, Orbital Part of the Inferior Frontal Gyrus, Precentral Gyrus, Rostral Middle Frontal Gyrus, Supramarginal Gyrus, Temporal Pole.
- Vision only: Amygdala, Hippocampus, Fusiform Gyrus, Inferior Temporal Gyrus, Lateral Orbitofrontal Cortex, Middle Temporal Gyrus, Orbital Part of the Inferior Frontal Gyrus, Precentral Gyrus, Rostral Middle Frontal Gyrus, Superior Frontal Gyrus, Superior Temporal Gyrus.
- Audio and Vision: Inferior Temporal Gyrus, Middle Temporal Gyrus, Orbital Part of the Inferior Frontal Gyrus, Precentral Gyrus, Rostral Middle Frontal Gyrus.

Again, we note that the small dataset and high variability in the results limit our ability to conclude with certainty that every detected electrode is indeed involved in language processing in the given modality. Our significance threshold assessment is likely to still result in false positives and, on the contrary, fail to detect some brain areas involved in language processing. Thus, further analysis is required to confirm the role of individual electrodes in modality processing during our language task and meticulously assess the coherence of our results with the current literature.

A more robust threshold, set at 97.5% of the bootstrap samples above zero, reduces the number of electrodes to 13, 9, and 1 for audio only, visual only, and both modalities, respectively. With this threshold, we strongly reduce the risk of false positives. The resulting electrodes are shown in Fig. A.4. The only detected multimodal electrode is located in the Frontal Lobe, which is involved in high-level language processing [99]. A visualization of the electrodes confirms a build-up in neural activity throughout the sentence (Fig. 4.8), aligned with previous observations [94]. This provides strong evidence of language integration across modalities and the ability of language models to predict this multimodal neural process.

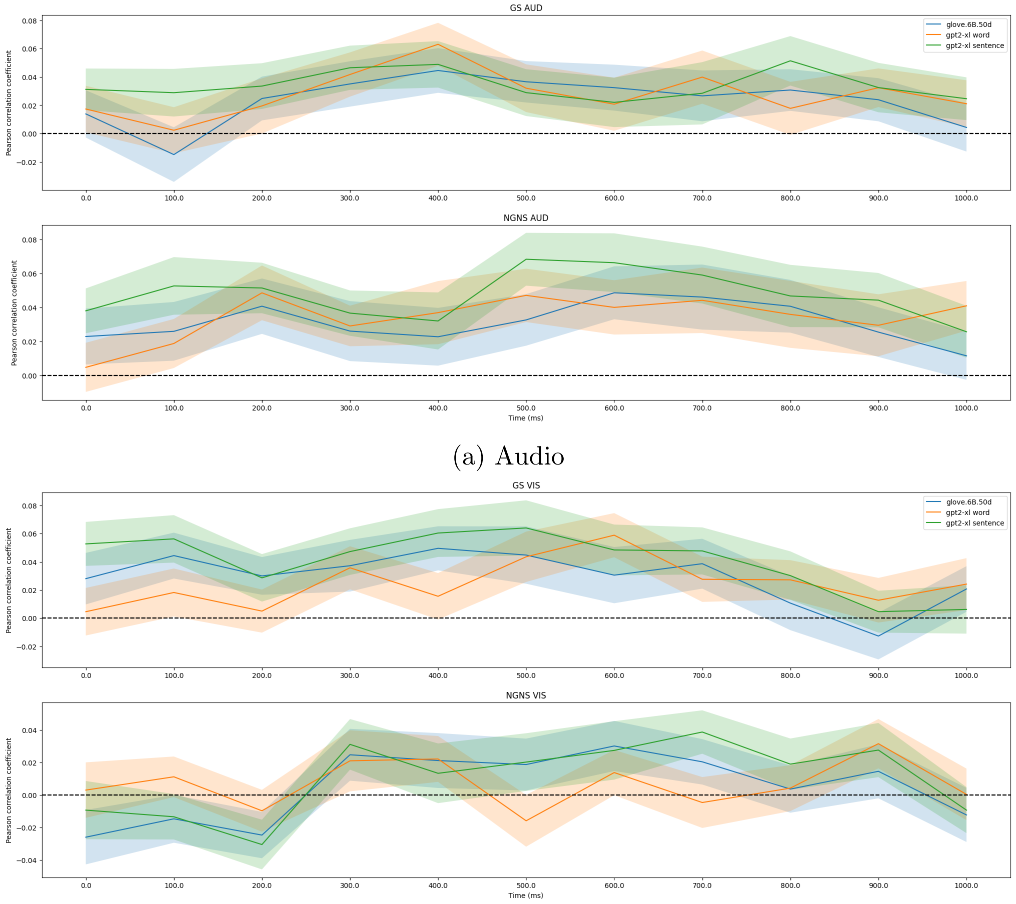
4.2.3 Average Alignment Analysis

To investigate the overall alignment dynamic, we average the correlation scores across modality-responsive electrodes. Averaging across electrodes is expected to lose some fine-grained spatial and temporal information. However, it allows us to obtain a general picture of the impact of the sentence type and predictor model on the alignment, following standard practice in the field [71, 75, 81].

We average the correlation scores across the 46 audio-responsive and 45 visual-responsive electrodes. We compare this mean average score across the different models (Fig. 4.13) and across the sentence types (Fig. 4.14).

The small size of our data and its high variability make it difficult to draw strong conclusions from these observations. However, we note some minor trends observed in these comparisons. In the case of model comparison, there seems to be an overall minor improvement in the alignment score from using the GPT-2 XL model with sentence embedding in both modalities, in line with previous findings [71, 75].

In the case of sentence type comparison, we observe an overall similar alignment score between the GS and nGnS sentences for the audio stimuli. However, the alignment score seems to be timed differently between the two types of sentences. Indeed, for the GS sentences, there appear to be spikes in activity around the 200-400ms window and again the 600-800ms window for GPT-2 XL sentence embeddings. We remind here that values at any time stamp correspond to the average activity over the previous 200ms. In comparison, nGnS sentences present



(a) Audio

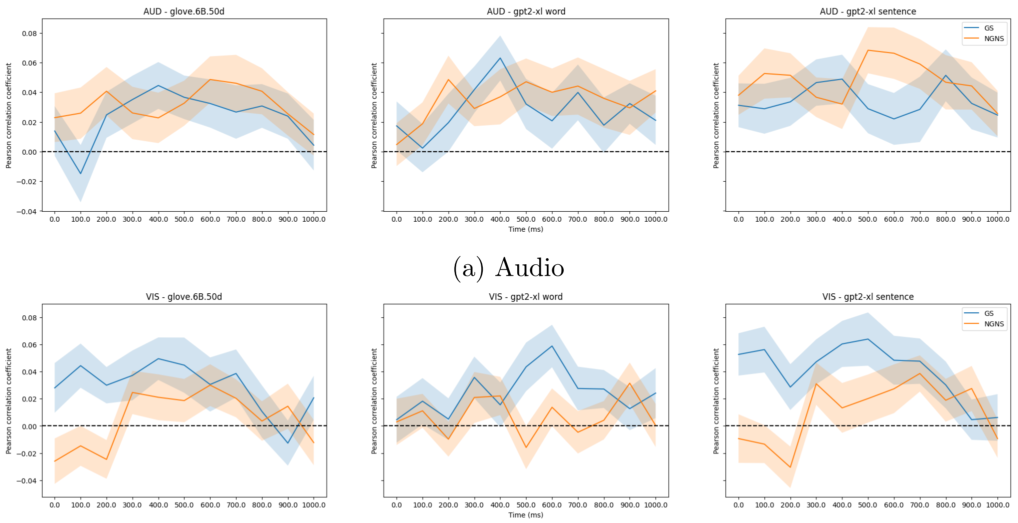
(b) Visual

Figure 4.13: Comparison of the mean Pearson correlation for modality responsive electrodes across models. The shaded area is the standard error of the mean.

spikes around the 0-200ms window and the 400-700ms window. This difference in timing hints toward potential variations in temporal processing and integration of linguistical content in the case of syntactic violation.

Given that W4 is identical in the two conditions, this potential difference can only be attributed to the impact of the syntactic violation from the inversion of noun and verb at W2 and W3, resulting in a lasting variation in linguistic processing at W4 that can be observed from the language model comparison. Using an average across electrodes makes assessing the temporal and local interactions at play challenging. Further analysis is required to confirm the significance of these observations and offer additional insights into the spatial and temporal dynamics leading to the observed alignment scores.

The visual stimuli, in comparison, show a higher alignment score in the GS



(a) Audio

(b) Visual

Figure 4.14: Comparison of the mean Pearson correlation for modality responsive electrodes across sentence type. The shaded area is the standard error of the mean.

condition compared to the nGnS one. The task paradigm might be at play as one could hypothesize that from W2, the subject knows he will press the red button and might be paying less attention to the visual stimuli, resulting in poorer decoding of the neural activity. In contrast, audio stimuli might force the subject into a more attentive state. However, the significance of these results, the impact of electrode location, and differences in the audio and visual processing of linguistic information should also be investigated further to offer better insights into the observed trends.

Chapter 5

Conclusion

5.1 Summary of Findings

In this study, we have explored the neural correlates of language processing using advanced neural network models to predict brain activity from stereo encephalography (sEEG) recordings of linguistic inputs under multiple modalities and sentence structures. Our findings confirm the presence of robust alignment between the models' predictions and neural responses across audio and visual stimuli, as well as correct and grammatically incorrect sentences. This highlights the models' ability to capture intricate dynamics of language processing in the human brain across modalities and grammatical variations.

By systematically investigating factors driving alignment between the models and neural activity, we have observed the impact of hidden variables, such as word position, that can lead to spurious conclusions about the models' performance. In some cases, these variables can even lead to faulty interpretations about the responsiveness of a brain area to specific modalities. Our results underscore the importance of controlling for these variables to ensure accurate and reliable assessments of model performance in predicting neural responses.

Furthermore, through statistical analyses of the models' alignment score, we have successfully identified language-responsive electrodes across visual and audio modalities through statistical analyses of the models' alignment score. Averaging the scores of these electrodes allowed us to compare performance across models and sentence structures. Our results align with previous studies on the role of context in improving the models' predictions in language processing tasks, suggesting that the complexity of the linguistic input influences the models' performance.

Finally, our results suggest that models' predictions are sensitive to syntactic violations in a sentence. This finding highlights the exciting potential of large language models to provide novel insights into the neural mechanisms and temporal dynamics underlying language processing in the human brain across a broad range of linguistic contexts, in this case, grammatical variations.

5.2 Discussion

The results presented in this study underscore the potential of comparing artificial neural networks with neural data to enhance our understanding of the brain's language networks. The observed correlations affirm the models' efficacy in predicting neural responses. However, challenges such as the variability in data and the complexity of neural processes call for cautious interpretation of these correlations.

Following our findings on the impact of hidden factors on alignment scores, such as word position, further research should be cautious in interpreting the models' performance as evidence of their ability to capture language processing in the brain. For example, electrodes involved in audio processing may exhibit patterns of neural activity correlated with word and sentence information without being directly involved in language processing. Future studies should consider these factors and the complexity of their interactions to ensure accurate and reliable conclusions about the models' performance.

The small data size and high variability in alignment scores presented challenges in assessing the statistical significance of the observed results. Future research would benefit from more robust methods in evaluating the significance of alignment scores. An approach we are exploring involves computing confidence intervals under a null hypothesis by measuring the alignment scores on permuted labels before assessing where the true alignment score falls within this distribution. We expect this method to offer more reliable estimates for the significance of a model's performance and contribute to a better identification of brain areas involved in language processing. Increasing the sample size and number of electrodes is also expected to improve the generalizability of our results.

Finally, our investigation of single electrode activity parallel to the models' predictions motivates a deeper look into the processes underlying alignment in specific brain regions. A fine-grained analysis of the neural activity and alignment scores across electrodes would provide valuable additional insights into the mechanisms underlying language processing in the human brain. In the case of syntactic processing, the identification of local neural correlates of syntactic violations would be of particular interest. Future research should aim to identify neural signatures of syntactic violations that large language models can explicitly capture at a spatiotemporal level.

Bibliography

[1] A. D. Friederici, "The brain basis of language processing: From structure to function," vol. 91, no. 4, pp. 1357–1392, 1371 citations (Semantic Scholar/DOI) [2024-04-12] 1176 citations (Crossref) [2024-04-12]. [Online]. Available: https://www.physiology.org/doi/10.1152/physrev.00006.2011

[2] A. D. Friederici, "Towards a neural basis of auditory sentence processing," vol. 6, no. 2, pp. 78–84, 1805 citations (Semantic Scholar/DOI) [2024-04-12] 1348 citations (Crossref) [2024-04-12]. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364661300018398

[3] A. D. Friederici, N. Chomsky, R. C. Berwick, A. Moro, and J. J. Bolhuis, "Language, mind and brain," vol. 1, no. 10, pp. 713–722, 234 citations (Semantic Scholar/DOI) [2024-04-12] 159 citations (Crossref) [2024-04-12]. [Online]. Available: https://www.nature.com/articles/s41562-017-0184-4

[4] L. Frazier and J. D. Fodor, "The sausage machine: A new two-stage parsing model," vol. 6, no. 4, pp. 291–325. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0010027778900021

[5] W. D. Marslen-Wilson, "Sentence perception as an interactive parallel process," vol. 189, no. 4198, pp. 226–228, publisher: American Association for the Advancement of Science. [Online]. Available: https://www.science.org/doi/10.1126/science.189.4198.226

[6] E. Fedorenko, A. Nieto-Castañon, and N. Kanwisher, "Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses," vol. 50, no. 4, pp. 499–513, 134 citations (Crossref) [2024-04-12]. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0028393211004210

[7] F. Artoni, P. d'Orio, E. Catricalà, F. Conca, F. Bottoni, V. Pelliccia, I. Sartori, G. L. Russo, S. F. Cappa, S. Micera, and A. Moro, "High gamma response tracks different syntactic structures in homophonous phrases," vol. 10, no. 1, p. 7537, 0 citations (Semantic Scholar/DOI) [2024-04-12] 14 citations (Crossref) [2024-04-12]. [Online]. Available: https://www.nature.com/articles/s41598-020-64375-9

[8] A. J. Reddy and L. Wehbe, "Can fMRI reveal the representation of syntactic structure in the brain?" 24 citations (Semantic Scholar/DOI)

[2024-04-12] 7 citations (Crossref) [2024-04-12]. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2020.06.16.155499

[9] A. D. Friederici and A. Mecklinger, "Syntactic parsing as revealed by brain responses: First-pass and second-pass parsing processes," vol. 25, no. 1, pp. 157–176. [Online]. Available: https://doi.org/10.1007/BF01708424

[10] E. Kaan, A. Harris, E. Gibson, and P. Holcomb, "The p600 as an index of syntactic integration dif??culty," vol. 15.

[11] D. Rumelhart, G. Hinton, and R. Williams, "Learning internal representations by error propagation." Elsevier, pp. 399–421, book Title: Readings in Cognitive Science. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9781483214467500352

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," vol. 9, pp. 1735–80.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. [Online]. Available: https://arxiv.org/abs/1706.03762v7

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space." [Online]. Available: http://arxiv.org/abs/1301.3781

[15] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Association for Computational Linguistics, pp. 1532–1543, 9997 citations (Semantic Scholar/DOI) [2024-04-12] 13173 citations (Crossref) [2024-04-12]. [Online]. Available: https://aclanthology.org/D14-1162

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding." [Online]. Available: http://arxiv.org/abs/1810.04805

[17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners."

[18] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners." [Online]. Available: http://arxiv.org/abs/2005.14165

[19] E. Pavlick, “Symbols and grounding in large language models,” vol. 381, no. 2251, p. 20220041, 25 citations (Semantic Scholar/DOI) [2024-04-12] 14 citations (Crossref) [2024-04-12]. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsta.2022.0041

[20] W. Gurnee and M. Tegmark, “Language models represent space and time.” [Online]. Available: http://arxiv.org/abs/2310.02207

[21] K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko, “Dissociating language and thought in large language models.” [Online]. Available: http://arxiv.org/abs/2301.06627

[22] I. Tenney, D. Das, and E. Pavlick, “BERT rediscovers the classical NLP pipeline,” 1196 citations (Semantic Scholar/arXiv) [2024-04-12]. [Online]. Available: http://arxiv.org/abs/1905.05950

[23] A. Rogers, O. Kovaleva, and A. Rumshisky, “A primer in BERTology: What we know about how BERT works,” vol. 8, pp. 842–866, 1154 citations (Semantic Scholar/DOI) [2024-04-12] 361 citations (Crossref) [2024-04-12]. [Online]. Available: https://direct.mit.edu/tacl/article/96482

[24] J. Hu, J. Gauthier, P. Qian, E. Wilcox, and R. P. Levy, “A systematic assessment of syntactic generalization in neural language models,” 160 citations (Semantic Scholar/arXiv) [2024-04-12]. [Online]. Available: http://arxiv.org/abs/2005.03692

[25] T. Linzen and M. Baroni, “Syntactic structure from deep learning,” vol. 7, pp. 195–212, publisher: Annual Reviews. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-linguistics-032020-051035

[26] A. Chen, R. Shwartz-Ziv, K. Cho, M. L. Leavitt, and N. Saphra, “Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs,” 10 citations (Semantic Scholar/arXiv) [2024-04-12]. [Online]. Available: http://arxiv.org/abs/2309.07311

[27] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” vol. 5, no. 4, pp. 115–133. [Online]. Available: https://doi.org/10.1007/BF02478259

[28] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain,” vol. 65, no. 6, pp. 386–408, place: US Publisher: American Psychological Association.

[29] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” vol. 323, no. 6088, pp. 533–536, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/323533a0

[30] D. O. Hebb, *The organization of behavior; a neuropsychological theory*, ser. The organization of behavior; a neuropsychological theory. Wiley, pages: xix, 335.

[31] T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton, “Backpropagation and the brain,” vol. 21, no. 6, pp. 335–346, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41583-020-0277-3

[32] J. Sacramento, R. P. Costa, Y. Bengio, and W. Senn, “Dendritic cortical microcircuits approximate the backpropagation algorithm.” [Online]. Available: http://arxiv.org/abs/1810.11393

[33] V. Francioni, V. D. Tang, N. J. Brown, E. H. Toloza, and M. Harnett, “Vectorized instructive signals in cortical dendrites during a brain-computer interface task,” p. 2023.11.03.565534. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10635122/

[34] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” vol. 521, no. 7553, pp. 436–444, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nature14539

[35] D. H. Hubel and T. N. Wiesel, “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex,” vol. 160, no. 1, pp. 106–154.2. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1359523/

[36] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate.” [Online]. Available: http://arxiv.org/abs/1409.0473

[37] Sparse distributed memory. [Online]. Available: https://mitpress.mit.edu/9780262514699/sparse-distributed-memory/

[38] T. Bricken and C. Pehlevan, “Attention approximates sparse distributed memory.” [Online]. Available: http://arxiv.org/abs/2111.05498

[39] M. Kawato, S. Ohmae, H. Hoang, and T. Sanger, “50 years since the marr, ito, and albus models of the cerebellum,” vol. 462, pp. 151–174.

[40] L. Kozachkov, K. V. Kastanenka, and D. Krotov, “Building transformers from neurons and astrocytes,” pages: 2022.10.12.511910 Section: New Results. [Online]. Available: https://www.biorxiv.org/content/10.1101/2022.10.12.511910v1

[41] N. Kriegeskorte, M. Mur, and P. A. Bandettini, “Representational similarity analysis - connecting the branches of systems neuroscience,” vol. 2, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/articles/10.3389/neuro.06.004.2008

[42] A. S. Morcos, M. Raghu, and S. Bengio, “Insights on representational similarity in neural networks with canonical correlation.” [Online]. Available: http://arxiv.org/abs/1806.05759

[43] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, “Similarity of neural network representations revisited,” in *Proceedings of the 36th International Conference on Machine Learning.* PMLR, pp. 3519–3529, ISSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v97/kornblith19a.html

[44] M. Khosla and A. H. Williams, “Soft matching distance: A metric on neural representations that captures single-neuron tuning,” in *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models.* PMLR, pp. 326–341, ISSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v243/khosla24a.html

[45] J. S. Prince, C. Conwell, G. A. Alvarez, and T. Konkle, “A case for sparse positive alignment of neural systems.” [Online]. Available: https://openreview.net/forum?id=8FnN1QmR84

[46] D. D. Cox and R. L. Savoy, “Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex,” vol. 19, no. 2, pp. 261–270.

[47] Y. Kamitani and F. Tong, “Decoding the visual and subjective contents of the human brain,” vol. 8, no. 5, pp. 679–685, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nn1444

[48] A. Ettinger, A. Elgohary, and P. Resnik, “Probing for semantic evidence of composition by means of simple classification tasks,” in *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP.* Association for Computational Linguistics, pp. 134–139. [Online]. Available: https://aclanthology.org/W16-2524

[49] G. Alain and Y. Bengio, “Understanding intermediate layers using linear classifier probes.” [Online]. Available: https://openreview.net/forum?id=HJ4-rAVtl

[50] J. Hewitt and P. Liang, “Designing and interpreting probes with control tasks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, pp. 2733–2743. [Online]. Available: https://aclanthology.org/D19-1275

[51] A. A. Ivanova, J. Hewitt, and N. Zaslavsky, “Probing artificial neural networks: insights from neuroscience.” [Online]. Available: http://arxiv.org/abs/2104.08197

[52] S. Marks and M. Tegmark, “The geometry of truth: Emergent linear structure in large language model representations of true/false datasets.” [Online]. Available: http://arxiv.org/abs/2310.06824

[53] K. Liu, S. Casper, D. Hadfield-Menell, and J. Andreas, “Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness?” [Online]. Available: http://arxiv.org/abs/2312.03729

[54] P. Bashivan, K. Kar, and J. J. DiCarlo, “Neural population control via deep image synthesis,” vol. 364, no. 6439, p. eaav9436, publisher: American Association for the Advancement of Science. [Online]. Available: https://www.science.org/doi/full/10.1126/science.aav9436

[55] G. Tuckute, A. Sathe, S. Srikant, M. Taliaferro, M. Wang, M. Schrimpf, K. Kay, and E. Fedorenko, “Driving and suppressing the human language network using large language models.”

[56] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, “Performance-optimized hierarchical models predict neural responses in higher visual cortex,” vol. 111, no. 23, pp. 8619–8624. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.1403112111

[57] S.-M. Khaligh-Razavi and N. Kriegeskorte, “Deep supervised, but not unsupervised, models may explain IT cortical representation,” vol. 10, no. 11, p. e1003915, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003915

[58] C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, “Deep neural networks rival the representation of primate IT cortex for core visual object recognition,” vol. 10, no. 12, p. e1003963, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003963

[59] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva, “Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence,” vol. 6, no. 1, p. 27755, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/srep27755

[60] U. Güçlü and M. A. J. v. Gerven, “Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream,” vol. 35, no. 27, pp. 10 005–10 014, publisher: Society for Neuroscience Section: Articles. [Online]. Available: https://www.jneurosci.org/content/35/27/10005

[61] C. Zhuang, S. Yan, A. Nayebi, M. Schrimpf, M. C. Frank, J. J. DiCarlo, and D. L. K. Yamins, "Unsupervised neural network models of the ventral visual stream," vol. 118, no. 3, p. e2014196118, publisher: Proceedings of the National Academy of Sciences. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2014196118

[62] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, F. Geiger, K. Schmidt, D. L. K. Yamins, and J. J. DiCarlo, "Brain-score: Which artificial neural network for object recognition is most brain-like?" 442 citations (Semantic Scholar/DOI) [2024-04-12] 208 citations (Crossref) [2024-04-12]. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/407007

[63] S. Wang, J. Sun, Y. Zhang, N. Lin, M.-F. Moens, and C. Zong, "Computational models to study language processing in the human brain: A survey," 1 citations (Semantic Scholar/arXiv) [2024-04-12]. [Online]. Available: http://arxiv.org/abs/2403.13368

[64] C. Conwell, J. S. Prince, C. J. Hamblin, and G. A. Alvarez, "Controlled assessment of CLIP-style language-aligned vision models in prediction of brain & behavioral data." [Online]. Available: https://openreview.net/forum?id=T90SJkeDKm

[65] C. Conwell, J. S. Prince, G. A. Alvarez, and T. Konkle, "Language models of visual cortex: Where do they work? and why do they work so well where they do?" vol. 23, no. 9, p. 5653. [Online]. Available: https://doi.org/10.1167/jov.23.9.5653

[66] A. J. E. Kell, D. L. K. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott, "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy," vol. 98, no. 3, pp. 630–644.e16.

[67] J. Millet, C. Caucheteux, P. Orhan, Y. Boubenec, A. Gramfort, E. Dunbar, C. Pallier, and J.-R. King, "Toward a realistic model of speech processing in the brain with self-supervised learning."

[68] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," vol. 320, no. 5880, pp. 1191–1195.

[69] J. Sassenhagen and C. Fiebach, "Traces of meaning itself: Encoding distributional word vectors in brain activity," vol. 1, pp. 1–41.

[70] S. Jain and A. Huth, "Incorporating context into language encoding models for fMRI," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc. [On-

line]. Available: https://papers.neurips.cc/paper_files/paper/2018/hash/f471223d1a1614b58a7dc45c9d01df19-Abstract.html

[71] M. Schrimpf, I. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. Tenenbaum, and E. Fedorenko, "The neural architecture of language: Integrative modeling converges on predictive processing."

[72] C. Caucheteux and J.-R. King, "Brains and algorithms partially converge in natural language processing," vol. 5, no. 1, p. 134, 124 citations (Semantic Scholar/DOI) [2024-04-12] 90 citations (Crossref) [2024-04-12]. [Online]. Available: https://www.nature.com/articles/s42003-022-03036-1

[73] C. Caucheteux, A. Gramfort, and J.-R. King, "Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, pp. 3635–3644, 22 citations (Semantic Scholar/DOI) [2024-04-12] 13 citations (Crossref) [2024-04-12]. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.308

[74] C. Caucheteux, A. Gramfort, and J.-R. King, "Deep language algorithms predict semantic comprehension from brain activity," vol. 12, no. 1, p. 16327, 36 citations (Semantic Scholar/DOI) [2024-04-12] 27 citations (Crossref) [2024-04-12]. [Online]. Available: https://www.nature.com/articles/s41598-022-20460-9

[75] A. Goldstein, Z. Zada, E. Buchnik, M. Schain, A. Price, B. Aubrey, S. A. Nastase, A. Feder, D. Emanuel, A. Cohen, A. Jansen, H. Gazula, G. Choe, A. Rao, C. Kim, C. Casto, L. Fanda, W. Doyle, D. Friedman, P. Dugan, L. Melloni, R. Reichart, S. Devore, A. Flinker, L. Hasenfratz, O. Levy, A. Hassidim, M. Brenner, Y. Matias, K. A. Norman, O. Devinsky, and U. Hasson, "Shared computational principles for language processing in humans and deep language models," vol. 25, no. 3, pp. 369–380, 172 citations (Semantic Scholar/DOI) [2024-04-12] 124 citations (Crossref) [2024-04-12]. [Online]. Available: https://www.nature.com/articles/s41593-022-01026-4

[76] A. Goldstein, E. Ham, M. Schain, S. Nastase, Z. Zada, A. Dabush, B. Aubrey, H. Gazula, A. Feder, W. K. Doyle, S. Devore, P. Dugan, D. Friedman, R. Reichart, M. Brenner, A. Hassidim, O. Devinsky, A. Flinker, O. Levy, and U. Hasson, "The temporal structure of language processing in the human brain corresponds to the layered hierarchy of deep language models," 7 citations (Semantic Scholar/arXiv) [2024-04-12] 7 citations (Semantic Scholar/DOI) [2024-04-12] 6 citations (Crossref) [2024-04-12]. [Online]. Available: http://arxiv.org/abs/2310.07106

[77] Z. Zada, A. Goldstein, S. Michelmann, E. Simony, A. Price, L. Hasenfratz, E. Barham, A. Zadbood, W. Doyle, D. Friedman, P. Dugan, L. Melloni,

S. Devore, A. Flinker, O. Devinsky, S. A. Nastase, and U. Hasson, “A shared linguistic space for transmitting our thoughts from brain to brain in natural conversations,” 4 citations (Semantic Scholar/DOI) [2024-04-12] 3 citations (Crossref) [2024-04-12]. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2023.06.27.546708

[78] C. Caucheteux, A. Gramfort, and J.-R. King, “Evidence of a predictive coding hierarchy in the human brain listening to speech,” vol. 7, no. 3, pp. 430–441, 59 citations (Semantic Scholar/DOI) [2024-04-12] 47 citations (Crossref) [2024-04-12]. [Online]. Available: https://www.nature.com/articles/s41562-022-01516-2

[79] M. Kutas and S. A. Hillyard, “Reading senseless sentences: Brain potentials reflect semantic incongruity,” vol. 207, no. 4427, pp. 203–205, place: US Publisher: American Assn for the Advancement of Science.

[80] R. Antonello and A. Huth, “Predictive coding or just feature discovery? an alternative account of why language models fit brain data,” pp. 1–16, 19 citations (Semantic Scholar/DOI) [2024-04-12] 11 citations (Crossref) [2024-04-12]. [Online]. Available: https://direct.mit.edu/nol/article/doi/10.1162/nol_a_00087/113632/Predictive-Coding-or-Just-Feature-Discovery-An

[81] G. Mischler, Y. A. Li, S. Bickel, A. D. Mehta, and N. Mesgarani, “Contextual feature extraction hierarchies converge in large language models and the brain,” 1 citations (Semantic Scholar/arXiv) [2024-04-12]. [Online]. Available: http://arxiv.org/abs/2401.17671

[82] J. Hale, C. Dyer, A. Kuncoro, and J. Brennan, “Finding syntax in human encephalography with beam search,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 2727–2736, 117 citations (Semantic Scholar/DOI) [2024-04-12] 39 citations (Crossref) [2024-04-12]. [Online]. Available: http://aclweb.org/anthology/P18-1254

[83] C. Caucheteux, A. Gramfort, and J.-R. King, “Disentangling syntax and semantics in the brain with deep networks.”

[84] C. Kauf, G. Tuckute, R. Levy, J. Andreas, and E. Fedorenko, “Lexical semantic content, not syntactic structure, is the main contributor to ANN-brain similarity of fMRI responses in the language network.”

[85] A. Marin Vargas, A. Bisi, A. S. Chiappa, C. Versteeg, L. E. Miller, and A. Mathis, “Task-driven neural network models predict neural dynamics of proprioception,” vol. 187, no. 7, pp. 1745–1761.e19, 1 citations (Crossref) [2024-04-12].

[86] V. Subramaniam, C. Conwell, C. Wang, G. Kreiman, B. Katz, I. Cases, and A. Barbu, "Revealing vision-language integration in the brain with multimodal networks." [Online]. Available: https://openreview.net/forum?id=7Scc7Nl7lg

[87] Y. Han, T. Poggio, and B. Cheung, "System identification of neural systems: If we got it right, would we know?"

[88] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen, "MEG and EEG data analysis with MNE-python," vol. 7, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2013.00267/full

[89] X. Jia, S. Tanabe, and A. Kohn, "Gamma and the coordination of spiking activity in early visual cortex," vol. 77, no. 4, pp. 762–774. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0896627313000445

[90] M. Schrimpf, I. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. Tenenbaum, and E. Fedorenko, "Artificial neural networks accurately predict language processing in the brain," pages: 2020.06.26.174482 Section: New Results. [Online]. Available: https://www.biorxiv.org/content/10.1101/2020.06.26.174482v1

[91] A. Goldstein, A. Grinstein-Dabush, M. Schain, H. Wang, Z. Hong, B. Aubrey, M. Schain, S. A. Nastase, Z. Zada, E. Ham, A. Feder, H. Gazula, E. Buchnik, W. Doyle, S. Devore, P. Dugan, R. Reichart, D. Friedman, M. Brenner, A. Hassidim, O. Devinsky, A. Flinker, and U. Hasson, "Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns," vol. 15, no. 1, p. 2768, 0 citations (Semantic Scholar/DOI) [2024-04-12] 0 citations (Crossref) [2024-04-12]. [Online]. Available: https://www.nature.com/articles/s41467-024-46631-y

[92] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "HuggingFace's transformers: State-of-the-art natural language processing." [Online]. Available: http://arxiv.org/abs/1910.03771

[93] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch."

[94] E. Fedorenko, T. L. Scott, P. Brunner, W. G. Coon, B. Pritchett, G. Schalk, and N. Kanwisher, "Neural correlate of the construction of sentence meaning," vol. 113, no. 41, 142 citations (Semantic Scholar/DOI) [2024-04-12] 134 citations (Crossref) [2024-04-12]. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.1612132113

[95] M. S. Koyama, D. O'Connor, Z. Shehzad, and M. P. Milham, "Differential contributions of the middle frontal gyrus functional connectivity to literacy and numeracy," vol. 7, no. 1, p. 17548.

[96] O. Blanke, L. Spinelli, G. Thut, C. M. Michel, S. Perrig, T. Landis, and M. Seeck, "Location of the human frontal eye field as defined by electrical cortical stimulation: anatomical, functional and electrophysiological characteristics," vol. 11, no. 9, pp. 1907–1913.

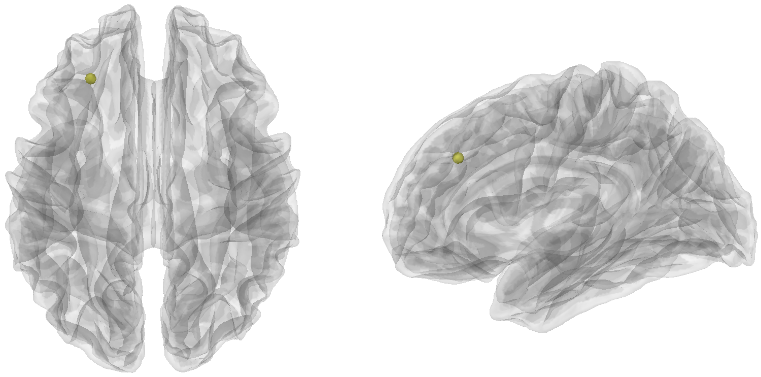
[97] E. Kaestner, T. Thesen, O. Devinsky, W. Doyle, C. Carlson, and E. Halgren, "An intracranial electrophysiology study of visual language encoding: The contribution of the precentral gyrus to silent reading," vol. 33, no. 11, pp. 2197–2214. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8497063/

[98] A. C. Gouvea, C. Phillips, N. Kazanina, and D. Poeppel, "The linguistic processes underlying the p600," vol. 25, no. 2, pp. 149–188, publisher: Routledge _eprint: https://doi.org/10.1080/01690960902965951. [Online]. Available: https://doi.org/10.1080/01690960902965951

[99] E. Fedorenko, J. Duncan, and N. Kanwisher, "Language-selective and domain-general regions lie side by side within broca's area," vol. 22, no. 21, pp. 2059–2062. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3494832/

Appendix A

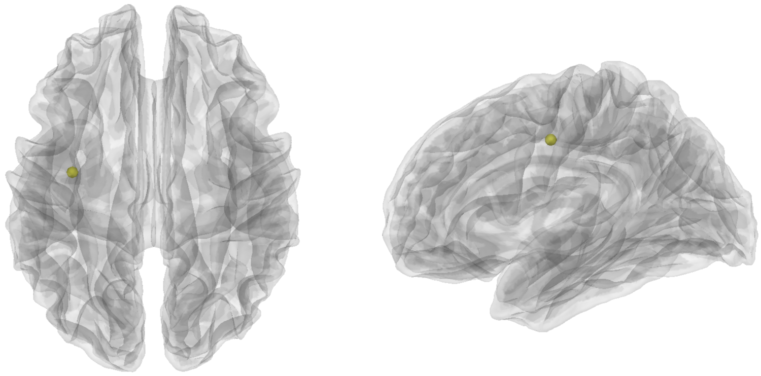
Electrode Locations



(a) Superior View

(b) Left Lateral View

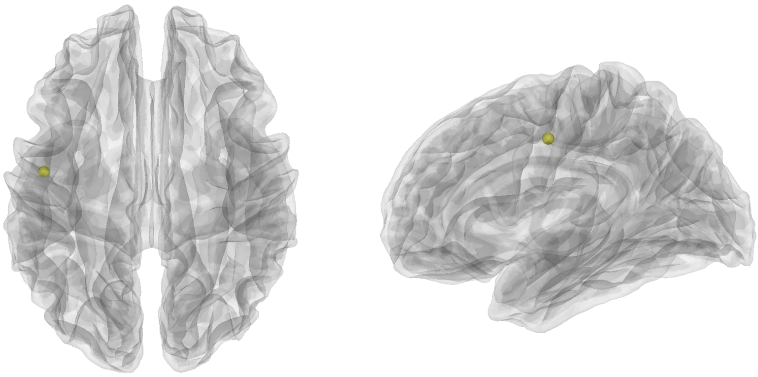
Figure A.1: Electrode 16 - Left Middle Frontal Gyrus



(a) Superior View

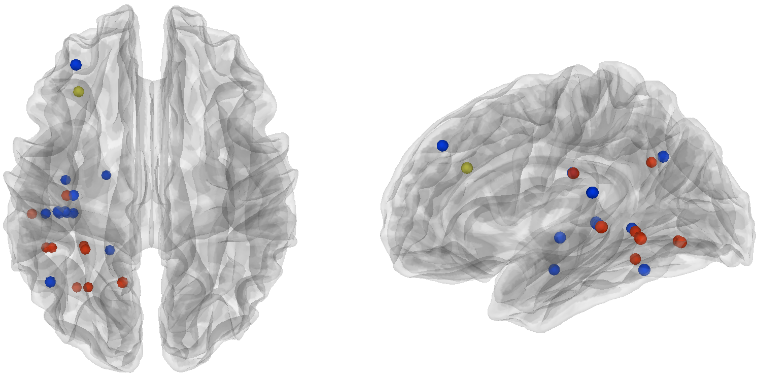
(b) Left Lateral View

Figure A.2: Electrode 127 - Left Precentral Gyrus



(a) Superior View (b) Left Lateral View

Figure A.3: Electrode 131 - Left Precentral Gyrus



(a) Superior View (b) Left Lateral View

Figure A.4: Modality responsive electrodes according to the 97.5% threshold. Blue indicates audio-only responsive electrodes, red indicates visual-only responsive electrodes, yellow indicates electrodes responsive to both modalities.

Appendix B

Neural Activities

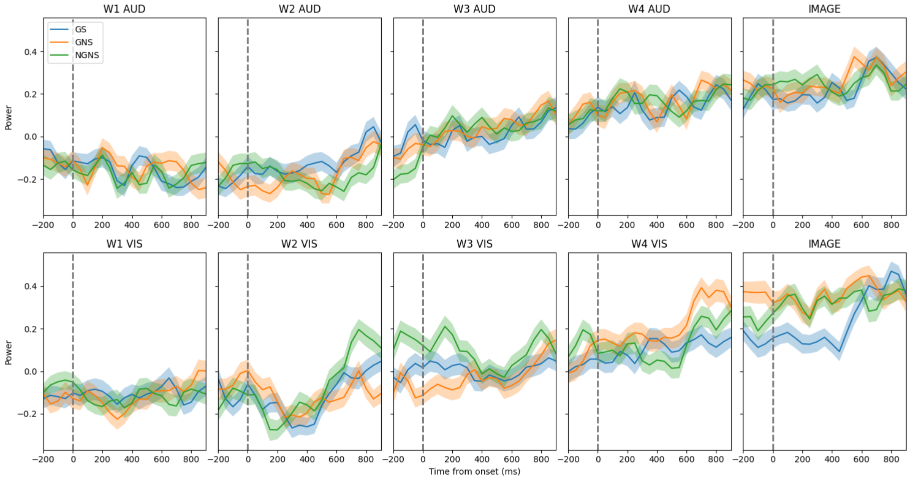
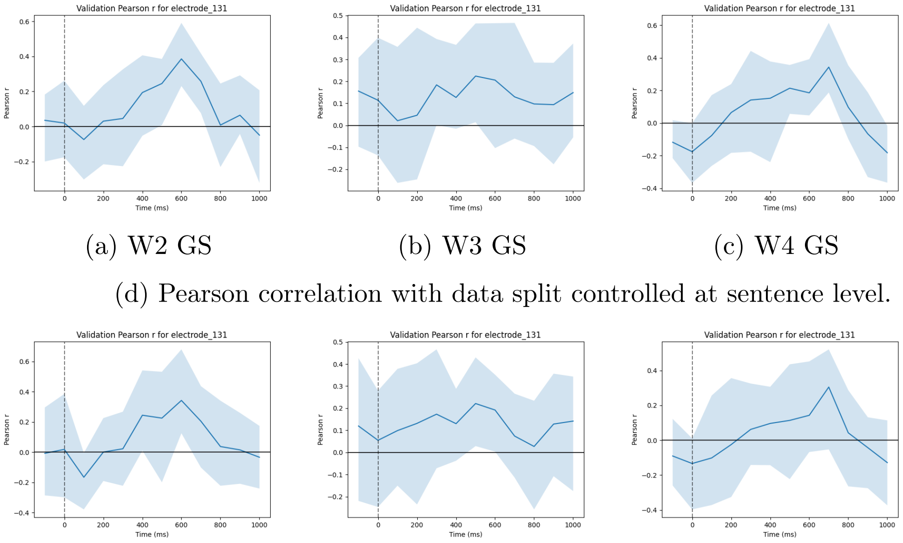


Figure B.1: Electrode 186, mean and standard estimate of the mean between words.

Appendix C
Pearson Correlations



(a) W2 GS

(b) W3 GS

(c) W4 GS

(d) Pearson correlation with data split controlled at sentence level.

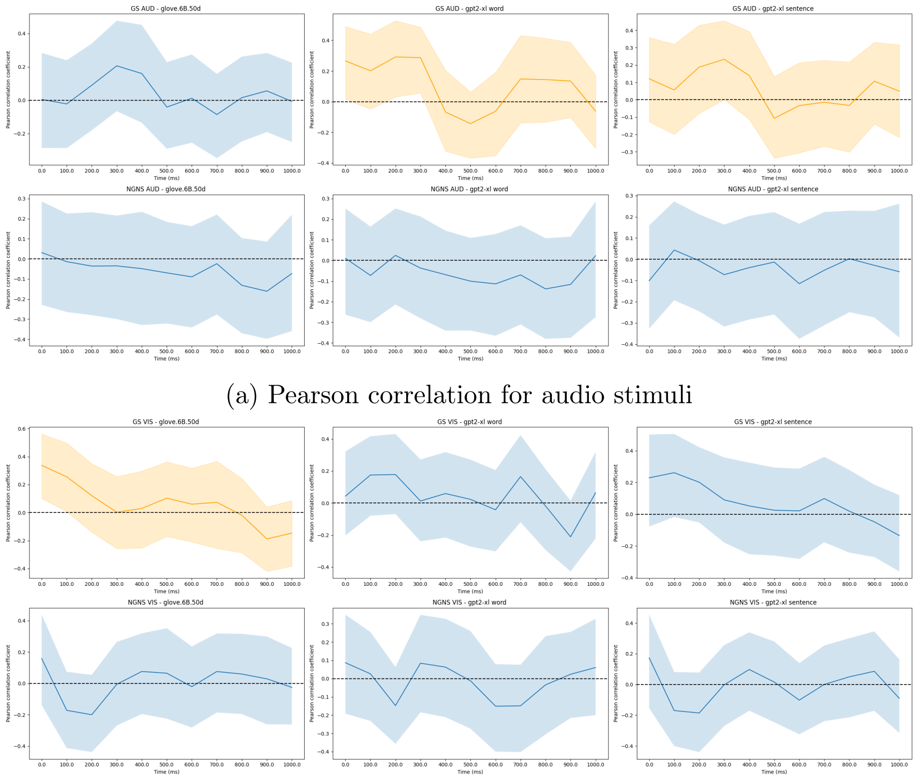
(e) W2 GS

(f) W3 GS

(g) W4 GS

(h) Pearson correlation with data split controlled at word level.

Figure C.1: Electrode 131, correlation with neural activity at single word level from Glove embeddings.



(a) Pearson correlation for audio stimuli

(b) Pearson correlation for visual stimuli

Figure C.2: Correlation for electrode 186. The shaded area represents the 90% confidence interval. Orange indicates that more than 95% of the bootstrap samples are above zero at any timestamp.