Using Large Language Models to Predict Linguistic Neurological Intracranial Signals for Multimodal Sentences

A thesis presented

by

Alliyah Steele

The Faculty of the Committee on Degrees in Computer Science
and Neuroscience in partial fulfillment of the requirements for the degree of

Bachelor of Arts

Harvard College

Cambridge, Massachusetts

March 13, 2026

The Harvard College Honor Code

Members of the Harvard College community commit themselves to producing academic work of integrity – that is, work that adheres to the scholarly and intellectual standards of accurate attribution of sources, appropriate collection and use of data, and transparent acknowledgement of the contribution of others to their ideas, discoveries, interpretations, and conclusions. Cheating on exams or problem sets, plagiarizing or misrepresenting the ideas or language of someone else as one's own, falsifying data, or any other instance of academic dishonesty violates the standards of our community, as well as the standards of the wider world of learning and affairs.

The Neuroscience Concentration AI Policy

Completing a thesis develops valuable skills in analytical thinking and scientific writing. **Students are strictly forbidden from including AI-generated content in any part of their thesis manuscript, including preliminary drafts**. Incorporating AI-produced text into your thesis constitutes an Honor Code violation at Harvard College and will be considered academic misconduct.

Certain applications of generative AI tools are acceptable during other research phases. For example, these tools may be valuable for broadening or deepening your knowledge of specific concepts, understanding methodologies, or troubleshooting analyses. However, exercise caution when using these resources: current AI models regularly produce inaccurate information. This is especially true for scientific research, as journals are often paywalled and thus under-represented in training data.

You bear full responsibility for the accuracy of all assertions in your thesis, making it essential to rigorously verify any AI contributions. You should be prepared to answer questions about any part of your thesis during the oral conference.

If you are unsure as to whether a particular use is allowed, contact your concentration advisor promptly.

By signing below, I confirm that I have adhered fully to all components of the Honor Code and AI Policy in the production of this thesis.

Digital Signature: Alliyah Steele

Acknowledgements

Working on this thesis has been a multi-year effort, and as such, I want to make sure to thank all the people who it wouldn't have been possible without. I want to thank many individuals from the Kreiman Lab, including Professor Gabriel Kreiman, Pranav Misra (my research mentor), Victor Gillioz, Morgan Talbot, Ellisa Pavarino, Dianna Hidalgo, Bastien Le Lan, and Narek Alvandian, and anyone else I may have encountered during my time in the lab. I really appreciated all of your feedback as I iterated through the many project stages. I also want to thank Kristina Penikis for all her support and for becoming my research advisor, and David Alvarez-Melis for being my CS reader. I also want to thank the entire Harvard undergraduate neuroscience and computer science teams. I've learned so much through taking on the challenge of this thesis and am highly grateful for all your help throughout the process.

List of Contributions

Student referenced the modeling work (on one participant) of Victor Gillioz on the 4-word task dataset owned by Pranav Misra as a baseline for developing an independent signal processing and model-fitting pipeline for all subjects. Student completed all aspects of the project independently over the course of 2 years, including 1) Behavioral data preprocessing, 2) Neurological signal processing, 3) Neurobehavioral trigger alignment, 4)Transformer embedding extraction, 5) Ridge Regression Modeling Pipeline, 6) Statistical analysis, 7) Plotting and Interpretation of results. This compiled work is present within the following GitHub repository owned by the student: https://github.com/HelloUniversePyC/GPT-T5-BERT-Alignment-4WTSEEGData

1. Abstract

Prior work using fMRI and EEG has demonstrated correlations between large language model (LLM) representations and neural language signals, but these approaches are limited by low temporal resolution and small sample sizes. To address these gaps, we analyzed stereoelectroencephalography (SEEG) recordings from 16 multilingual participants completing an auditory and visual sentence task with semantic and syntactic violations. Per-electrode ridge regressions mapped hidden layer embeddings from three transformer models (BERT, GPT-2-XL, and mT5-XL) onto gamma-band neural signals across cortical and subcortical regions, without anatomical priors. Permutation testing and False Discovery Rate (FDR) correction were applied to control for false positives.

Significant neural-LLM alignment was sparse but anatomically specific, clustering in right-hemisphere homologous language regions. The right inferior frontal sulcus (IFS) showed the highest significant explained variance ($r^2 = 0.27$, $p < .001$) driven primarily by responses to syntactic and semantic violations. This was consistent with previous research indicating right-hemisphere linguistic regional recruitment under increased linguistic processing demand. The insular cortex, despite not being a classical language area, showed selective alignment with BERT embeddings under grammatical conditions, supporting its proposed role in auditory-motor integration and grammar learning.

These findings suggest transformer representations may capture aspects of both classical and non-classical language processing, particularly under conditions of linguistic surprisal. This work demonstrates the potential of transformer-SEEG alignment as a tool for mapping language-relevant cortical regions and has potential downstream implications for modeling learning disabilities, brain-computer interface decoding, and cognitively-inspired LLM architectures.

Github Experimental Code Repository:

https://github.com/HelloUniversePyC/4-Word-Task-Preprocessing-and-ML-Analysis-Pipeline

2. Table of Contents

3. Introduction

As technology advances into the 21st century, interactions between machine learning and neuroscience have become more apparent. The finding that activations of convolutional neural networks (CNNs) linearly correlate with neuronal activations within the visual cortex yielded benefits to the neuroscience and computer science fields (Cadena et al., 2019; Kar et al., 2019; Kuzovkin et al., 2018, 2018; Yamins et al., 2014; Yamins & DiCarlo, 2016). Within neuroscience it led to improved descriptive models of the visual cortex to simulate experiments (Kar et al., 2019; Lindsay, 2021). It has also improved decoding model performance for vision based BCIs (Wilson et al., 2024). Within Software Engineering, ML engineers have leveraged human eye tracking data to guide CNN training (van Dyck et al., 2022), and have incorporated lateral geniculate nucleus-inspired layers as CNN front-ends to improve robustness (Dapello et al., 2020; Lonnqvist et al., 2024).

Given that there is proven alignment between human and computer vision a natural question is whether there may also be alignment between human neurolinguistic processing neural activity and the representations of large language models (LLMs). Such a finding would have implications for both cognitive neuroscience and computer science. For neuroscientists, language models could provide new tools to simulate and interpret high-dimensional linguistically evoked neural responses. For computer scientists, it could identify novel ways to improve LLM performance by mimicking the behavior and structure of linguistic neural circuits. Investigating whether this kind of alignment exists requires a robust understanding of both human linguistic processing and

computational linguistic acquisition. This analysis will begin by first discussing the most ancient form of intelligence, the human brain, and its linguistic implications.

3.1 The Human Brain: A state-of-the-art (SOTA) biological language model

The human brain contains over 100 trillion connections (Gebicke-Haerter, 2023), and is, in a sense, the first ever "large language model", affording individuals the ability to process language both visually and auditorily. This processing is possible due to the interaction of distributed and hierarchical neural networks located within frontal, parietal, and temporal regions of the brain. A single neuron is composed of dendrites (neuronal appendages that listen for signals from other neurons), a soma (main cell body that handles metabolic functions of the neurons), and an axon (a long, thin projection along with electrical signals called axon potentials that travel to allow a neuron to send messages to other cells). This simple structure is not unlike the structure of single perception within an artificial multilayer perceptron (MLP) model. The dendrites are analogous to the inputs of the perceptron. A region called the axon hillock, found at the initial segment of the axon, receives the summed inputs from the dendrites and then determines nonlinearly if the neuron has reached threshold to fire. This parallels the affine transformation and subsequent nonlinearity applied in a single perceptron. A single neuron cannot do very much on its own; in the same way, a single perceptron model would be futile. Neurons gain representative power by forming synapses with other neurons. A synapse is a small gap between neurons that allows for either electrical (via direct transduction) or chemical (via neurotransmitter molecules) sending of signals between cells (Pereda, 2014). Within the language system, these complex neural networks allow for lower-level information like photons and auditory frequencies to be gradually processed and built up into higher-level representations like grammar, syntax, and semantic meaning (Deniz et al., 2019; Hagoort, 2019; Matchin & Hickok, 2020; Misra et al., 2024; Moro et al., 2001).

Importantly, within the imaging method used in this study, SEEG, gamma band frequencies (30-150Hz) are used as a proxy of cellular firing rates within neuronal populations surrounding an electrode (Ray &

Maunsell, 2011). SEEG recording uses intracortical depth electrodes to record neuronal frequencies from distributed brain regions. It is an invasive form of neural recording often leveraged for individuals with treatment-resistant epilepsy for seizure loci monitoring (Youngerman et al., 2019). Unlike surface-level EEG (5-9cm resolution), SEEG (1 mm resolution) has improved spatial resolution and can record from deeper brain areas (Ex, Basal ganglia, striatum, etc.) (Herff et al., 2020; Parvizi & Kastner, 2018). It also has improved temporal resolution (SEEG-1ms temporal resolution) compared to functional magnetic resonance imaging (fMRI- 1 second temporal resolution). SEEG recordings are on the level of 10,000s of neurons at a time per millimeter of spatial resolution (Bernabei et al., 2021; Lachaux et al., 2012). The gamma band frequency amplitudes directly correlate with the multi-unit spiking activity (MUA) of neurons surrounding the electrode tip. Functionally, gamma oscillations are generated by the interaction of excitatory (glutamatergic) and inhibitory (GABAergic) neurons. Increased firing of these circuits (action potentials) results in higher gamma powers (Buzsáki & Wang, 2012). Classically, these gamma band frequencies correlate with higher-level processes such as focused attention, memory binding, perception, and consciousness (Hudson & Jones, 2022).

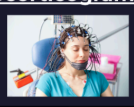
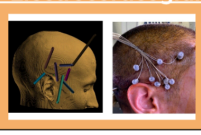
	Spatial Resolution	Temporal Resolution	Invasive?
Electrocorticogram (EEG) 	5-9cm	1 ms	No
Functional Magnetic Resonance Imaging (fMRI) 	<1mm	1 sec	No
Stereo-Electrocorticogram (SEEG) 	1 mm	1 ms	Yes

Figure 1. Spatial and Temporal Resolutions of Neural Imaging Methods. Summary of metrics from (Herff et al., 2020; Parvizi & Kastner, 2018, Bernabei et al., 2021; Lachaux et al., 2012)

3.1.1 The Neuroanatomy of Language

The path from low-level signals to high-level syntax and grammar in the brain is hierarchical and modality dependent. For instance, auditory processing of language begins in the cochlea (eardrum), which transduces sound waves into neural impulses. These impulses are then transmitted to the primary auditory cortex after a quick pit stop in the thalamus, the brain's sensory relay station. Within the primary auditory cortex (located in the temporal lobe), neurons are selective to low-level variables such as loudness and pitch of sounds. Hierarchical neuronal networks in the superior temporal sulcus (STS) tuned to specific frequencies then begin to decompose features within the raw input signals, transmitting these recognized patterns to downstream neurons. Next, Wernicke's area, located at the superior temporal gyrus (STG), is utilized to extract high-level linguistic features from these processed signals, such as specific phonemes and lexical elements. The middle temporal gyrus (MTG) also participates in this feature extraction. Notably, Wernicke's area is a known language comprehension area, given the widespread documentation of the pathology of Wernicke's aphasia: a condition that results in an inability to produce coherent sentences as a result of damage to this cortical region (Damasio & Tranel, 1993; Fridriksson et al., 2015). Finally, right hemisphere-based networks in temporo-parietal regions process the rhythm and syncopation of sentences (Deniz et al., 2019; Forseth et al., 2020; Hagoort, 2019; Matchin & Hickok, 2020; Misra et al., 2024; Moro et al., 2001).

Visual language processing (when reading text) has a slightly different pathway. First, the photoreceptors in the retina respond to photons entering the eye and transmit these signals to synapsed retinal ganglion cells (RGCs). RGCs, although early in the cascade, respond to low-level features such as contrast in the input signals. The retinal ganglion cells then synapse onto the retina, where signals are then transduced by the optic nerve (connected at the back of the eye) and sent to the lateral geniculate nucleus (the specialized region of the thalamic sensory relay station for vision). These signals are then transmitted to the primary visual cortex, where simple and complex cells respond to basic features such as spatial location, orientation, and color of stimulus elements. The stimulus information then proliferates through the remaining three layers of the visual cortex

(there are 4 layers, V1-4), where receptive fields of neurons are gradually stacked to go from detecting simple bars to shapes to compound shapes. The receptive field refers to the range of input a neuron responds to. For instance, a simple cell might respond the most to a bar at a 60° angle in the input. If we were to combine three of these simple cells, we could construct a neural network that fires in response to equilateral triangles, which is similar to the kind of hierarchical processing that occurs in later layers of the visual cortex (Antolik & Bednar, 2011).

These processed symbols then go to the ventral stream (known colloquially as the “what” stream) for visual recognition to occur. This ventral stream is a distributed neuronal network located within distributed temporal and occipital regions. Importantly, the visual word form area (VWFA) in the left fusiform gyrus is responsible for recognizing the letters and symbols in text. Additionally, the Left Inferior Frontal Gyrus (IFG), also known as Broca’s area, plays a role in high-level syntax and parsing sentence structure of text. Wernicke's area also contributes to semantic comprehension (Deniz et al., 2019; Forseth et al., 2020; Hagoort, 2019; Kuzovkin et al., 2017, 2018; Matchin & Hickok, 2020; Misra et al., 2024; Moro et al., 2001).

Crucially, the architecture of CNNs is a good analogy for the functioning of the visual cortex. In fact, as mentioned previously, previous studies have found high levels of representational alignment between CNN and visual cortex representations (Cadena et al., 2019; Kar et al., 2019; Kuzovkin et al., 2017, 2018; Yamins et al., 2014). CNNs, like the visual cortex, learn representations gradually with earlier layers learning simple features that can be stacked and reused in later layers to detect compound features (Ex, going from detecting curves and dots to detecting a smiley face). Networks that contact skip or residual connections have an additional similarity with biological networks, given that connections in the visual cortex are both feedforward and recurrent. Within CNNs, residual connections allow the model to use previously learned representations more efficiently instead of re-learning the same features redundantly (Alzubaidi et al., 2021). The mathematical operation of convolution itself to find and learn features with CNNs also has some biological correspondence when

compared to the mechanism of receptive fields and neuroplasticity within visual cortex networks (Kar et al., 2019; Kuzovkin et al., 2017, 2018).

Independent of auditory or visual modalities, frontal lobe regions such as the Angular Gyrus and Inferior Frontal Lobe are responsible for the interpretation of words across modalities (visual and auditory), allowing for dynamic understanding of language. Past Studies indicate that the high-level representations of linguistic content at high-level language processing sites (Ex, IFG, ATL, etc.) are not modality dependent. Hence, the fully pre-processed auditory or visually delivered information is represented using similar neural networks in the brain. This hints at there being a possibility of some universal representation of syntax or grammar within the brain, independent of low-level dynamics (Deniz et al., 2019; Misra et al., 2024).

Furthermore, Linguistic processing is strongly lateralized to the left hemisphere (for right-handed individuals) (Caucheteux et al., 2021; Chen et al., 2021; Matchin & Hickok, 2020; Moro et al., 2001) but there is burgeoning evidence for collaboration between the left and right hemispheres to produce language function. For instance, Wernicke's and Broca's areas have been shown to demonstrate co-activation of left and right hemisphere networks when processing language. Additionally, the right IFG and right angular gyrus have been linked to attentional control during long speech exchanges (Hagoort, 2019). These observations have led to the hypothesis that right and left hemisphere regions co-activate in response to linguistic task difficulty. Specifically, in the presence of increased linguistic cognitive load (possibly due to grammar or syntax violations, surprisal, etc.) homologous language areas in the right hemisphere often co-activate to help process the load. Hence, the L-IFG and other classical language areas may recruit their right hemisphere counterpart to aid in processing the challenging input. The left hemisphere still takes the lead in language and is more specialized, but this adaptive behavior may aid the human brain to be more flexible and adaptive to linguistic task difficulty (Lindell, 2006;

Moro et al., 2001; Muller & Meyer, 2014; Qi et al., 2019; van Ettinger-Veenstra et al., 2010).

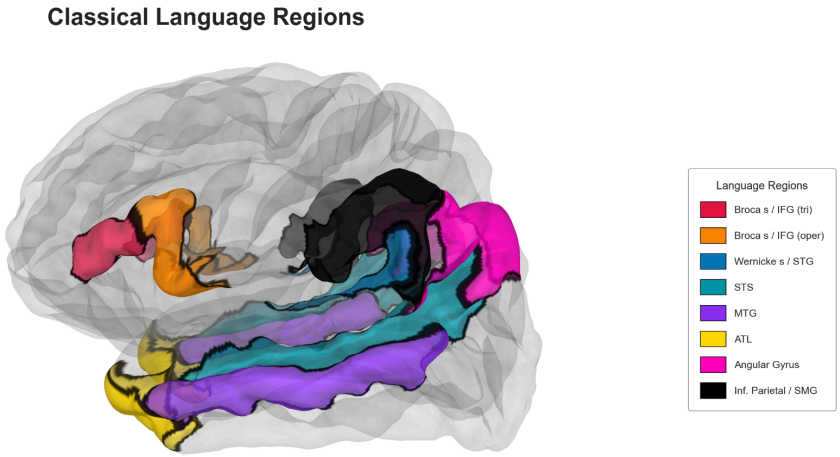


Figure 1. Classical left hemisphere language regions. Color-coded. Abbreviations: Inferior Temporal Gyrus Pars Triangularis (IFG tri), Inferior Temporal Gyrus Pars Opercularis (IFG oper, Superior Temporal Sulcus (STS), Medial Temporal Gyrus (MTG), Anterior Temporal lobe (ATL), Inferior parietal lobe, Supramarginal Gyrus (SMG). Figure created using MNE FsAverage brain model and Matplotlib.

3.1.2 Neurolinguistics: Syntax vs Grammar and Nouns vs Verbs

Linguistics is the study of the structure of language and can provide a vital entry point by which to make descriptive claims about neurological language processing. This paper will focus on two axes of language: Syntax and Grammar. Grammar is defined as the general rules that make up a language, while syntax and semantic meaning are related subsets of this more general property. Syntax refers to rules regarding proper word orderings to create coherent sentences. However, Semantics describes the high-level meaning behind utterances and describes how otherwise arbitrary sounds become significant (Kracht, 2007).

Past studies have investigated syntax and grammar within the brain. Matchin & Hickok (2020) proposed a dual-region model, challenging the “Broca-only” view of linguistic processing. The model proposed a consideration of the distinct roles of the posterior-MTG (pMTG) and posterior-IFG (pIFG) in syntactic processing. The framework argues that the pMTG supports hierarchical semantic-syntactic integration, while the pIFG is primarily responsible for syntactic sequencing. This study introduces a non-classical view of

language regions and partially informs the present study's choice to use all electrodes when searching for potential "high-alignment regions" later in this study. Such an approach avoids forming assumptions regarding the neurological linguistic system, which is far more distributed than previously thought.

In Friederici & Meyer (2004), it was found using EEG event-related potentials (ERPs) that the brain could distinguish between local phrase structure violations and complex verb-argument structure violations. Their analysis revealed that the brain employs distinct temporal processing strategies in response to different violation types. Phrase structure violations involve an immediate breakdown in the expected word order or category (e.g., "The small in on" or "The child will in the garden"). These violations occur early (100–300ms) and can elicit an Early Left Anterior Negativity (ELAN) via the frontal operculum. This presents functionally as a large downward current within the recording graph. Verb-argument violations, however, involve a mismatch between the verb and the required participants. Some examples are a verb lacking its object or a verb having too many arguments. For instance, sentences like "The cleaner cleaned the floor, the dirty" or "The boy sneezed the napkin off the table". These kinds of sentences activated Broca's area and the left posterior-STG (pSTG). These violations occurred later and produced N400-P600 patterns. These results show evidence for the hierarchical processing of linguistic content, given that syntactic errors were present on the EEG trace before semantic ones, indicating hierarchical activity. It also provides a framework by which to hypothesize about the timescale of grammatical and syntactical violations explored later in this paper.

Vigliocco et al. (2011) proposed that nouns and verbs are processed by distinct, distributed, and left-lateralized neural networks. The review highlights that while nouns activate left inferior frontal and temporal regions, verbs show greater activation in motor and premotor areas (The action-oriented language of the verb can trigger action-related areas). Within EEG recording data, it was shown that these processes diverge within 200ms of linguistic stimulus onset. These findings support alternative neural organization schemas where knowledge is categorized by semantic properties (object/action). This finding is directly relevant to the

word-level analysis that occurs later in this study and the hypothesis regarding the delivery of nouns vs verbs to participants.

In Deniz et al. (2019) researchers found that the human brain utilizes a shared, modality-independent system to process the meaning of language (semantics). Words appeared to be organized in "semantic maps" that were independent of information delivery methods (auditory or visual). The study used fMRI scans while N= 9 individuals listened to or read a textual story. The researcher then trained an encoding model (ridge regression) where the continuous vector space representation of the sentence was the X variable, and the y variable was the per voxel BOLD signal. The researchers found high levels of representational similarity between the linguistic embedding vectors and the FMRI BOLD signals. The high-correlation semantic network found was located across the temporal, parietal, and prefrontal cortices. The similarity between representations between modalities was so robust that models trained on auditory data could accurately predict brain activity during reading (visual data) and vice versa. The methodological structure of this experiment is a strong source of motivation for the study design presented in this paper. If fixed word embeddings lead to such a strong alignment with neural data, it's possible that richer transformer features may lead to even more insights regarding descriptive modeling of neural data.

Overall, neurological approaches to studying language have been fruitful, but with the advent of artificial intelligence (AI) models, namely, large language models (LLMs). It's possible that utilizing richer, multi-dimensional representations from these models may allow for improved latent feature discovery of language regions in the brain. The classical method of descriptively modeling language in the brain consists of training an encoding model on some representation of grammar and syntax and then computing correlations between the model and the brain. Importantly, using LLMs as that representation (given the improved complexity and comparability to the human brain) over static word embeddings (GLoVE, Word2Vec, etc.) may allow for improved findings. However, before one can make such a conclusion, we must discuss in detail the second and newer form of intelligence: Artificial Intelligence.

3.2 Artificial Models of Language: SOTA Neural Networks

3.2.1 From the Neuron to the Perceptron

Notably, the field of neural networks (NNs) first began by studying and emulating human neural networks (Al-Mahasneh et al., 2017; Raiaan et al., 2024, 2024). LLMs are deep neural networks (DNNs) specialized as generalists to produce textual output to fit a variety of user needs, from conversational tasks to programming to logic. LLMs are specialized models that live under the umbrella term NN.

The simplest unit of a NN is called a perceptron. A perceptron models all inputs to a single node (the smallest unit of a NN), which conceptually is similar to the behavior of a neuron in the brain. Each perceptron unit consists of all the inputs to a given node, an affine transformation that transforms the data, an activation function to introduce nonlinearity, and finally the outputs. Each node learns its own set of weights and biases in order to satisfy an objective function. The affine transformation, where weights and biases are applied to the inputs, determines whether the node "fires" or not. The activation function applied subsequently (Some examples being GeLu, Relu, Sigmoid, and Tanh) allows the model to learn nonlinear relationships. An individual neuron cannot trigger (Transmit information to other nodes in the network) until the total input exceeds a specific threshold. This is similar to how neurons biologically do not fire until the membrane potential exceeds a specific value. (Al-Mahasneh et al., 2017; Nagda et al., 2020; Naveed et al., 2025; Widrow & Lehr, 1990).

GPT-2-XL and Bidirectional encoder representations from transformers (BERT) multilingual-base-cased (two of the models that will be explored in this paper) both use the activation function Gaussian Error Linear Unit (GELU) based on the rectified linear unit (ReLU) function to introduce nonlinearity into the model, allowing it to extrapolate more complex trends beyond simple linear relationships (Gardazi et al., 2025; Rogers et al., 2020; Topal et al., 2021; Wu et al., 2023; Yenduri et al., 2024). The third model that will be explored in this paper, mT5-XL, utilizes GeGlu, a modified form of the Gelu activation function (model choices justified later in Section 2.2.2). Neural net activation functions mirror how biological synapses have varying neuronal weights

instead of all inputs to an individual neuron having the same importance (allowing for increased plasticity and functionality of circuits). Neurologically, cellular connections are upweighted or downweighted via long-term plasticity effects (e.g., long-term potentiation and long-term depression) and short-term plasticity effects (Ex, facilitation and depression) to impact both synaptic and structural plasticity (Daoudal & Debanne, 2003; Hiratani & Fukai, 2014; Vitureira & Goda, 2013). Traditionally, computational neural networks use backpropagation and gradient descent during training to adjust their weights and modify how inputs are modeled across layers to enhance performance (Al-Mahasneh et al., 2017; Nagda et al., 2020; Naveed et al., 2025; Widrow & Lehr, 1990).

Combining layers of perceptrons creates a multilayer perceptron or NN. Individual perceptron units are stacked into three-layer types: input, output, and hidden layers (located between input and output). NNs that exceed one hidden layer are called deep neural networks (DNNs) and can perform even more complex functions than simpler approximators. According to the universal function approximation theorem of neural networks, even a 1-layer network can approximate any algebraic function given enough nodes (Yun et al., 2020). Although this proof has not been formally approached with multilayer neural networks in practice, generally, more nodes and layers yield greater network accuracy and flexibility. Just like in human brains, complex functions do not arise from single units or neurons but instead from the interactions between networks of nodes.

2.2.2 Model Selection Justification

GPT-2-XL, BERT-base-multilingual-cased, and mT5 are the three neural networks that will be explored in this study. They are special kinds of neural networks called transformer neural networks (TNNs). TNNs combine the standard structure of a neural network with per-token (where a token is a small unit of input text) attention mechanisms to process textual inputs. Given the differences between the architectures of these three models (GPT-2-XL is decoder-only, BERT-base-multilingual-cased is encoder-only, and mT5 contains both encoder and decoder) (Gardazi et al., 2025; Rogers et al., 2020; Topal et al., 2021; Wu et al., 2023; Yenduri et al., 2024),

comparing them provides a substantial amount of coverage over types of LLM architectures and their potential alignment with neural data. It may also allow for some speculation regarding whether humans and language models employ similar methods for comprehending language (e.g., next word prediction- decoder algorithm vs. summarizing from surrounding words- encoder algorithm or a combination of the two). High-level, an encoder is an architectural block that transforms the input text, while the decoder produces the output text from that transformed input.

Also related to architecture, the models have varying numbers of hidden layers, with GPT-2-XL being the deepest and BERT being the shallowest. This also allows for some reasoning regarding the number of trainable parameters and their correlation with representational similarity. Importantly, all three models are cased as well (meaning their attention mechanism differentiates between capital and lowercase letters), which aligns with the corpus of textual data used in this experiment. Given the precedent of XL models having higher representational alignment with the human brain (Hong et al., 2024), each model chosen is amongst the largest in its respective family of models (T5, GPT, or BERT). These models are also vanilla, non-instruction-tuned raw models, preventing confounding factors like reinforcement learning from human feedback (RLHF) artifacts in the data.

Additionally, BERT and T5 have performed well on multilingual benchmarks, which is crucial for the dataset of this study (The dataset is majority English speakers but also contained participants who completed the tasks in Hindi and Mandarin Chinese, labeled as Taiwanese in study documents to reflect the nature of the dialect). See Figure 3 for more information about study demographics. Although GPT-2 has not performed extremely well on multilingual benchmarks (Yoo et al., 2025), it is included given its precedent of high alignment with neurolinguistic tasks and as a potential control model (Blank et al., 2014; Fedorenko et al., 2016; Gillioz, 2024; Goldstein et al., 2022; Hong et al., 2024; Pereira et al., 2018; Schrimpf et al., 2021; Tuckute et al., 2024). Crucially, all three models are also open-source (Through the Hugging Face transformers library) , allowing

their embeddings to be inspected and utilized for training in this experiment. See Figure 2 for model summary information.

Summary of Chosen Language Models

Model	Company (Year)	Architecture	Hidden Layers	Parameters (M)	Activation Function	Training Corpus (# Languages)
BERT-Base Multilingual-Cased	Google (2018)	Encoder	12	179M	GELU	Wikipedia (104 langs)
GPT-2-XL	OpenAI (2019)	Decoder	48	1,500M	GELU	WebText (1 lang)
mT5-XL	Google (2021)	Encoder-Decoder	24	3,700M	GeGLU	mC4 Corpus (101 langs)

Figure 2. Chosen Language Model Parameters. Summary of Chosen Hugging Face language models. Metrics obtained from Hugging Face Documentation.

See more model information on the Hugging Face website: https://huggingface.co/google-bert/bert-base-multilingual-cased

https://huggingface.co/openai-community/gpt2-xl

https://huggingface.co/google/mt5-xl

3.2.3 Transformer Neural Networks: Pay Attention!

The self-attention algorithm is key to TNN's understanding of human language. Attention is utilized during training and inference to determine the importance of individual tokens in input sequences. This allows the model to develop a conceptual understanding of both the semantic and syntactic meaning and rules behind human language to better understand user intent. To calculate attention weights first, the text is input into a tokenizer that decomposes large sentences into individual words, sub-words (Ex, a compound word like Superhero may be broken into super and hero as individual tokens), and characters.

The self-attention algorithm is then run, which is a token-wise method of computing weighted interactions with every other token in the sequence using learned query, key, and value vectors. The Query, key, and value vectors are learned mathematical projections of the input text embeddings (representations of text in vector form in the latent model word dimension): Queries ask “what am I looking for?”, Keys represent “what do I contain?”, and Values are “what information do I return if selected?”. For instance, consider the prompt: “Explain the basics of quantum computing to me.” Each token in this sentence, Explain, basics, quantum,

computing, and so on, produces its own query, key, and value vectors. Suppose the model is processing the token “basics”. The query vector derived from basics represents what that token is “looking for” in the rest of the sentence. It is compared against the key vectors of all other tokens. If the keys corresponding to “quantum” and “computing” are highly compatible with the query from “basics”, they will receive higher attention scores. The value vectors associated with “quantum” and “computing “are then weighted according to those attention scores and combined to form a context-sensitive representation of basics. Importantly, the value vectors for “quantum” do not literally store related words and phrases like “qubit” or “particle physics”. Instead, it contains a learned distributed representation that captures statistical and semantic patterns from training. Through this weighted combination of values, the model builds an internal representation that integrates the relationships among tokens and ultimately supports the eventual generation of an appropriate explanation. While Biological networks rely on regions like Wernicke's Area, Broca’s Area, the Arcuate Fasciculus, and the Inferior Temporal lobe for semantic processing, TNNs utilize this vector-based attention schema. Notably, GPT, BERT, and T5 all use Scaled Dot-Product Attention to compute attention weight via softmax (Gardazi et al., 2025, 2025; Nagda et al., 2020; Rogers et al., 2020; Topal et al., 2021; Wu et al., 2023) See the formula below:

$$\alpha = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

Where alpha is the multivariable probability distribution of attention weights, Q is the query vector, K is the key, and d_k is a normalization constant to avoid large gradients when dimensionality is high.

The resulting attention weights are a probability distribution over the tokens, indicating how much each unit contributes to the overall semantics of the fragment. A Positional Encoding vector is then added to input embeddings to inject sequence order, since attention itself is permutation-invariant and will not respect the initial syntax of the fragment (without this positional encoding, the model has no idea where tokens originally

were in the sequence). Finally, each value vector is multiplied by its corresponding attention weight, and these are summed up. This produces the final output, a context-aware embedding for the original query token (Soydaner, 2022; Vaswani et al., 2023).

Embeddings in Transformers are high-dimensional vectors that convert discrete input tokens (words, subwords, or pixels) into continuous numerical representations within a learned latent model word embedding space. These representations capture semantic meaning and relationships, allowing models to process language. The latent embedding space is learned during training, enabling similar words (Ex, Dog and Cat) to have closer vector representations than words that are more distant in meaning (Ex, Submarine and key). While humans understand language through complex neuronal firing patterns, transformers understand language via transforming features of language (Ex, syntax, grammar, sentiment, etc.) into different axes of a learned continuous vector space.

The idea of modeling words in continuous embedding spaces started with smaller static language-based models like global word vectors (GLoVe) and word to vector (Word2Vec). However, transformers revolutionized computational linguistic processing by not just computing static word embeddings but adding the dynamic attention-based mechanisms to understand words in context (Goldstein et al., 2022; Mahajan et al., 2025; Reddy & Wehbe, 2021; Soydaner, 2022; Vaswani et al., 2023). It's somewhat similar to how the executive processing of the Inferior Frontal Lobe in the brain allows humans to understand language better in context instead of just relying on static, non-environmentally influenced representations. Frontal lobe networks also compute selective attention in humans, similarly to attention networks (Merkx & Frank, 2021). Attention mechanisms in a TNN are the difference between the model knowing a language's rules superficially and actually becoming a practitioner.

Unlike their predecessor models, recurrent neural networks (RNNs), transformers compute all token representations simultaneously, enabling full graphical processing unit (GPU) batching (Mienye et al., 2024; Soydaner, 2022; Vaswani et al., 2023). Each one of these parallel processes that compute attention at different

positions within the input sequence is called an attention head. GPT, BERT, and T5 all use multi-headed attention in which multiple attention heads operate in parallel, allowing the model to learn different relational patterns within the text (e.g., local syntax vs long-range dependencies) (Gardazi et al., 2025; Rogers et al., 2020; Topal et al., 2021; Wu et al., 2023; Yenduri et al., 2024). This general framework applies to most transformer neural networks; however, there are some key differences between the attention-based mechanisms of GPT-2-XL, BERT-Base-multilingual-cased, and mT5-XL.

3.2.4 Neural Network Training

Neural networks differ greatly from biological intelligence in how they learn. Neural networks generally utilize a large corpus of data subdivided into separate training, test, and validation datasets to achieve optimal performance. From a full dataset of features, ordinarily, 80% is used for training, while the remaining 20% is split between the validation and the training set.

Firstly, the training set is used to teach the neural network its parameters (weights and biases)(Deuschle, 2019; Jiang et al., 2020). Many neural networks leverage self-supervised learning during training instead of unsupervised learning. Self-supervised learning is a subset of unsupervised learning in which the model uses already-provided data (in this case, the input text) to generate its own labels and predictions. Unlike supervised learning, no human annotations are required for training to commence, but the internal generation of objectives can promote better performance than unsupervised learning alone (Gardazi et al., 2025; Rogers et al., 2020; Topal et al., 2021; Wu et al., 2023; Yenduri et al., 2024). It also allows for the model to still generate more varied and unstructured responses, as is intended with traditional unsupervised learning, without relinquishing all structure (making it a nice medium between unsupervised and supervised learning). In the case of a TNN, the model is learning the optimal latent word embedding space that best minimizes the task loss function.

Next, the Validation Set (Usually formed from 10-15% of the train data after splitting the data again after the initial 80/20 or 70/30 test train split) helps tune hyperparameters (values that need to set before training such

as learning rate, layer depth, and width) and monitor performance during training (e.g., for early stopping), preventing overfitting to the training data. Cross-validation is also often employed to fit multiple training and validation set combinations by iterating over a fixed number of shuffled partitions of the data (called folds) and then averaging hyperparameter estimates over the n folds. This prevents the model from being overfit to a single randomly chosen validation set and makes the hyperparameter estimates less biased.

Finally, the test set serves as an untouched dataset used only after training to get an unbiased measure of how well the final model performs on new data. Once a model is successfully trained and reaches the desired performance benchmarks, it can then be used for inference tasks. Inference refers to using a pretrained model to make predictions on new, unseen data (Deuschle, 2019; Jiang et al., 2020).

"Learning" in a neural network means tuning the weights and biases of each node within the network. This manipulation is achieved via the process of backpropagation. Backpropagation (Backprop for short) is an algorithm that allows the neural network to calculate error and then make minute changes spread backwards across its nodes to correct it. The algorithm begins with a forward pass, where the input is passed through all layers of the model to calculate the output and a snapshot of all the current node weights. Next, the gradient descent algorithm is applied at each node to calculate how much the weights should change to reduce error. Conceptually, calculating the gradient of the loss at that node tells us which direction to go to make the error increase. Therefore, we take steps in the opposite direction of the gradient to reduce error by negating the gradient vector. In practice, most training loops utilize stochastic gradient descent (SGD) or mini-batch stochastic gradient descent (MGSD), in which instead of using all the node weights to calculate the gradients, we instead take a random sample of nodes, either 1 or a set batch size that is usually a power of 2, to optimize graphical processing unit (GPU) performance. Computing gradients in small batches instead of the entire dataset at once functions as a built-in regularization method and can help prevent overfitting within the final model. Finally, the backward pass applies the calculated optimal weight update changes in reverse (From the

output layer all the way back to the input layer) throughout the network using the calculus chain rule (Deuschle, 2019). Mathematically, the formula for calculating the gradients in backpropagation is as follows:

$$\frac{\partial L}{\partial W} = \left(\frac{\partial L}{\partial y}\right)\left(\frac{\partial y}{\partial W}\right)$$

1. Calculate the gradient of the loss with respect to the weights starting from the output layer

$$\frac{\partial L}{\partial W_i} = \left(\frac{\partial L}{\partial a_L}\right)\left(\frac{\partial a_L}{\partial a_{L-1}}\right)\left(\frac{\partial a_{L-1}}{\partial a_{L-2}}\right)\cdots\left(\frac{\partial a_{i+1}}{\partial a_i}\right)\left(\frac{\partial a_i}{\partial W_i}\right)$$

2. Use the chain rule to propagate the error backwards through all layers of the network from output to input

$$W_{\text{new}} = W_{\text{old}} - \eta\left(\frac{\partial L}{\partial W}\right)$$

3. Apply the weight update rule to each node to move each weight opposite the gradient to decrease overall loss

Neural networks are generally trained over a fixed number of trials (called epochs, where an epoch is defined as the model iterating through all the training data once) until the model reaches the desired benchmarks. These two general classes of problem models are used to solve either a regression (predicting a number) or classification (predicting a label a sample belongs to) task. The network architecture is specialized to fit the given task. In a regression task, we simply end in a linear affine layer to output the desired metric. In a classification network the final layer must have as many nodes as the number of classes we are predicting and the either the sigmoid (in binary class prediction) or softmax (in multi-class prediction) are used to normalize our class scores to a probability distribution so for each observation we have the percent probability it is in each of the k classes outputted (The largest probability out of the k classes is chosen as the assigned class for the observation).(Al-Mahasneh et al., 2017; Nagda et al., 2020; Naveed et al., 2025; Widrow & Lehr, 1990). Now that general transformer training and attention have been reviewed, one has the necessary context to review more of the specific implementation-level details of the three chosen language models: GPT-2-XL, BERT-Base-Multilingual-cased, and mT5-XL.

3.2.5 All about GPT-2-XL

GPT-2-XL is an open-source transformer neural network that was developed by OpenAI in 2019. It contains 48 layers, where each hidden layer has a dimension of 1600 nodes. The model was trained on 406 billion bytes of web-based text sources. GPT is an autoregressive language model and is decoder-only, meaning it utilizes the previous n-1 tokens to predict the nth token in its learning paradigm. It specifically uses masked self-attention instead of relying on contextual embeddings when processing input.

This masked self-attention works by ensuring that each token predicts the next token by only looking at past tokens, never future ones, enforcing causality for text generation. This is achieved by adding a "look-ahead" mask (often negative infinity) to the attention scores before the softmax, effectively zeroing out connections to future tokens, forcing the model to predict text word-by-word. For instance, in the case of casual masking, a triangular mask is applied, blocking attention to subsequent positions. If the sequence is "The cat sat", when processing "cat", the mask prevents GPT from seeing "sat". Implementation-wise, the mask is added to the scaled dot-product scores.

Scaled dot product attention is augmented with mask M:

$$\alpha = \text{softmax}\left(\frac{QK^T + M}{\sqrt{d_k}}\right)$$

Where M is defined as:

$$M_{ij} = 0 \text{ if } j \leq i; \quad M_{ij} = -\infty \text{ if } j > i$$

In this formulation, positions to be masked get a negative infinity allocation, turning into near-zero weights after softmax, while valid positions get 0. This effectively prevents the model from “cheating” during training or inference by preventing peaking ahead at the next token in the sequence (Al-Mahasneh et al., 2017; Topal et al., 2021; Wu et al., 2023; Yenduri et al., 2024).

3.2.6 All about BERT-Base-Multilingual-cased

BERT-base-multilingual-cased, unlike GPT-2-XL, is an Encoder-only model. This means that its output is not autoregressively computed, but instead, BERT can use both the left and right input tokens in every layer. BERT was developed by Google in 2018 and was trained on the BookCorpus (800 M words) and English Wikipedia (2,500 M words). BERT-base-multilingual-cased has 24 layers, where each hidden layer has a dimension of 1024 nodes, and 16 parallel attention heads. The term "cased" means that outputs are not invariant to casing of letters (Ex, A vs a). GPT-2-XL is also a cased language model, making BERT-base-multilingual-cased a good model to compare it to. Additionally, within the neural dataset (explained later in more detail), there are multilingual Taiwanese sentences, which necessitate the use of models that were trained on multiple languages, justifying the choice of multilingual models like mT5.

Computationally, BERT relies on contextual embeddings to produce valid linguistic outputs. Specifically, BERT makes use of a "masked language model" (MLM) pre-training objective to do away with the unidirectional constraint of traditional language models. MLM randomly masks or hides some of the tokens in the input and predicts the masked word based only on its surrounding context. The BERT framework includes two steps, namely pre-training and fine-tuning. In the pre-training stage, BERT is trained on unlabeled data over a large number of different tasks. Pre-training of BERT occurs in two steps: masked LM and next sentence prediction (NSP). During the MLM step, a certain percentage of input tokens are masked at random, and then the original vocabulary of these masked tokens is predicted. During Next sentence prediction, it is confirmed that the model understands the relationships between sentences. To accomplish this, two sentences, A and B, are chosen for each pretraining example. For instance, 50% of the time, B is the actual sentence after A, and 50% of the time, it is a random sentence from the corpus. For the fine-tuning step, the task-specific inputs and outputs are plugged into the model, and all parameters are fine-tuned end-to-end to obtain the fine-tuned model. Fine-tuning refers to adapting a pre-trained model to a new, specific task or dataset by continuing its training

with smaller adjustments, often freezing early layers (which learn general features) and training later layers (which learn task-specific features) (Gardazi et al., 2025; Rogers et al., 2020; Topal et al., 2021).

3.2.7 All about mT5-XL

Multilingual text-to-text transformer five extra large (mT5-xl) is a fine-tuned version of Google's text-to-text transformer five (T5), specialized to 101 languages from the Multilingual Common Crawl Corpus (1 trillion tokens). It contains 3.7 billion trainable parameters. Given that this experiment will include data from individuals who spoke Taiwanese, using transformer neural networks that perform well on multilingual benchmarks is essential to prevent bias. mT5 is called "text to text" since the model architecture is specialized both for input and output text fragments. This structure allows the model to perform well on natural language processing (NLP) tasks such as language translation, document summaries, and question answering. mT5 is an encoder-decoder model, meaning it computes attention mechanisms autoregressively like GPT-2-XL in the decoder portion, but bi-directionally, like BERT in the encoder portion. mT5 also uses cross-attention to allow for the decoder module to take in the unmasked output of the encoder, in order to allow for the model to focus on the entire input sample more effectively.

3.2.8 Evidence of Alignment between DNN representations and the human brain

There have been several studies that have attempted to establish a link between DNN and TNN representations and human neurolinguistic processing. Establishing such a connection would benefit both the fields of computer science and neuroscience. In terms of direct applications, in computer science, novel TNN architectures based on human neurological circuits could be developed. Within neuroscience, novel methods to simulate experiments with transformer networks to determine experimental feasibility and viability could be developed. More indirect applications to neuroscience include new computational models of dyslexia, improved BCI linguistic decoding models, and contributions to neurolinguistics that could impact educational policy.

Goldstein et al. (2024) sought to determine whether the brain also utilizes continuous embedding spaces to represent language similarly to DNNs. To test this hypothesis, they recorded electroencephalogram (EEG) recordings from the IFG in N = 3 participants during a continuous podcast listening task (With the pre-central gyrus and postcentral gyri acting as controls). They then trained a DLM (deep neural network) to represent words as positional geometries and evaluated whether the model could predict corresponding geometries of neural frequencies evoked by phrases in the podcast.

They found that the DLM could predict geometric embedding patterns of a left-out word even in zero-shot learning paradigms (The model predicts the answer given only a natural language description of the task. No gradient updates are performed). They found that contextual embeddings that contained information about the semantic meaning of phrases captured the geometry of the IFG neural embeddings significantly better than static (syntax-only) word embedding models (GloVe, Word2Vec, etc.). This analysis is compelling as it suggests that even though neural frequencies and LLM embeddings don't exist in the same vector space, there may be some basis for comparing them due to similarities in geometric alignment when placed in linear subspaces. It suggests a need for future research with a higher sample size to determine the specifics of how properties of language (syntax and grammar) are represented in the brain and in LLMs across modalities (visual and auditory, as in this study, language was only presented auditorily.

Other studies have specifically examined more widely available large-language models, such as BERT, GPT-2-XL, and T5, to determine whether their embeddings can be leveraged to predict neural activity in humans. These analyses have been conducted largely in two recording methods: functional magnetic resonance imaging (fMRI) and Stereoelectrocorticogram (SEEG).

For instance, Reddy & Wehbe (2021) aimed to determine whether fMRI could detect neural activity correlated with syntactic processing within the brain. To answer this question, the researchers leveraged fMRI data from N=9 participants while reading a chapter of a novel. They trained an encoding model on voxel-level brain activity and then assessed the explained variance scores between various constructed embedding types and

the voxel BOLD signals. Embeddings were constructed to represent various components of sentence syntax tree representations from the largest completed subtree of each word, to predict embeddings of the next word, and to predict embeddings of incremental top-down parses of the sentence. They chose to add BERT embeddings to their vectors in experimental trials as a variable to determine whether the model was actually learning noise vs actual rich syntactic and semantic information.

Overall, they found that embeddings representing predictive modeling (next word prediction) had the highest correlation with voxel-level activity. The model specifically had significant performance in areas linked to linguistic processing, such as the anterior temporal lobe, posterior temporal lobe, and angular gyrus. Additionally, the model had a higher correlation with the rich BERT embeddings compared to the static GloVe embeddings. Their results provide additional support for the benefit of using transformer neural networks over static word embedding models when developing predictive models of brain activity. This analysis would be improved further by also directly including decoder-only models (ex, GPT-2-XL) and exploring a recording metric that has higher temporal resolution (Ex, SEEG).

Tuckute et al. (2024) also explored fMRI data by assessing whether GPT2-xl's last hidden layer embeddings could be used to generate sentences that either suppress or excite neural responses in human subjects. Experimentally, they first showed that GPT2-xl embeddings predict fMRI BOLD signal activations with high significance ($p < .001$, $r = 0.48$). The researchers then conducted a 3-phase experiment: Train (N = 5,1000 sentences), Evaluation (N = 3,1500 sentences), Blocked fMRI (N = 4,720 sentences). Sentences presented to each novel group of subjects were separated into three categories: baseline, drive, and suppress.

It was found that drive and suppress sentences engineered by GPT2-xl could increase the BOLD signal by 12.9% (in the case of drive sentences) in subjects or decrease it by 56.6% (in the case of suppress sentences). This result is promising as it shows that GPT-2-xl embeddings contain some relation to the surprise level or emotional content of sentences. It also indicates that even with a low temporal resolution imaging method like

fMRI, the signals extracted are sufficient to extrapolate linguistic trends. This study should be repeated with a more temporally significant imaging method and also explore encoder-only models (Ex, BERT)

Schrimpf et al. (2021) conducted a comprehensive evaluation of 43 artificial neural network (ANN) language models, including embeddings (e.g., GloVe), recurrent networks (e.g., LM1B), and transformers (e.g., GPT, BERT), by comparing their internal activations to human brain activity and behavior during language processing. Neural data were collected from three datasets: fMRI recordings during sentence-by-sentence reading (Pereira et al., 2018), ECoG data from word-by-word sentence reading (Fedorenko et al., 2016), and fMRI during naturalistic story listening (Blank et al., 2014). A linear regression mapped model activations to neural responses, and performance was quantified as a normalized Pearson correlation ("brain score") relative to dataset-specific noise ceilings.

It was found that transformer models, particularly GPT2-xl, consistently outperformed other architectures, achieving near-ceiling predictivity for (Pereira et al., 2018) and (Fedorenko et al., 2016), and ~32% for (Blank et al., 2014). Notably, untrained transformer models still achieved substantial brain predictivity (~51%), indicating that neural network architecture alone explains a significant portion of neural variance. Additionally, model success in next-word prediction (measured on WikiText-2) was strongly correlated with reading-time predictivity ($r = 0.67$, $p < 0.0001$****), further supporting the link between language model performance and human linguistic processing.

Notably, there are studies that have utilized T5 representations. (Zhang et al., 2025) investigated the relationship between deep language model (DLM) representational similarity and human brain activity during an abstractive summarization (ABS) task. Specifically, they constructed representational dissimilarity matrices (RDMs) to assess the degree of alignment (correlation) between artificial activations and human neural responses to the summarization task. They specifically compared the representations of BART, PEGASUS, and T5 LLM model layers (multiple hidden layers were tested) against human EEG data at multiple timescales.

Results demonstrated that increased hidden layer depth led to greater alignment between artificial and neural responses. Additionally, altering the weights of deeper layers significantly impaired brain-model similarity and model summarization capability. Out of the three models tested, T5 showed the highest similarity to the brain in the N400 component window. This specific ERP, 400ms after stimulus onset within an EEG trace, has been linked to the brain's processing of semantic content. Summatively, this study hints that encoder-decoder models like T5 may also be a good candidate for alignment with linguistic neural data. However, notably, there are far fewer studies using T5 representations than BERT and GPT representations, which suggests a gap in research that should be explored.

There is also a body of literature that leverages SEEG, which has higher temporal resolution compared to FMRI and fair spatial resolution to manage some of the trade-offs seen in past studies with fMRI analysis. Specifically, SEEG offers millimeter-level spatial resolution (thousands to millions of neurons at a time) and millisecond-level temporal resolution compared to fMRI, which has submillimeter spatial resolution but only second-level temporal resolution (Bernabei et al., 2021). Despite having less spatial resolution than fMRI, SEEG may have the potential to reveal more of the small time-scale neuronal fluctuations involved in neurolinguistic processing.

Goldstein et al. (2022) conducted a study in which n=9 individuals with treatment-resistant epilepsy listened to a 30-minute podcast. N=50 behavioral participants were used to provide data on the human ability to predict the next word within the podcast (28% accuracy). The goal was to perform a large-scale analysis of the gamma power electrocorticography (ECoG) activations of the left hemisphere cortices of the brain and their correlation to machine learning activations based on specific features of language (next-word prediction, entropy/uncertainty, surprisal, etc.).

They found that GPT-2 embeddings predicted human neural responses with a higher correlation ($r = 0.79$) than static embedding methods like word2vec and GloVe. This false discovery rate, q, was less than 0.01, indicating there may be strong potential to use deep language models (DLMs) to simulate neural responses

(given linguistic input data) and learn more about neuro-linguistic processing. This extremely high explained variance suggests that working with cleaner neurological signals (in this case, surgically implanted electrodes in epilepsy patients) has the potential to reveal an even clearer correlation with language model embeddings.

Following the promising results of (Goldstein et al., 2022) other researchers have attempted to replicate and further these results. One specific dataset that has been used within the Kreiman lab is (Misra, 2024) 4-word-task dataset

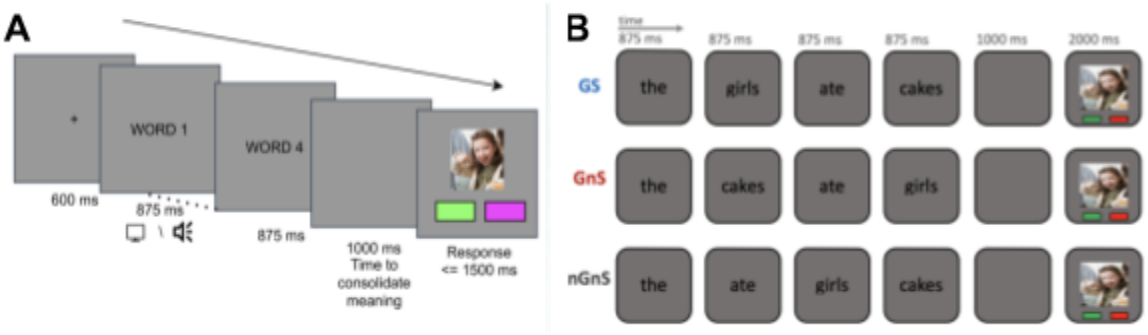


Figure 3. (Misra, 2024) **4-Word-Task. A) Trial Structure.** Each trial began with a 600 ms fixation point, followed by a sequence of four words presented for 875 ms each, delivered either visually (centered, ~3° visual angle, black font) or auditorily (blank screen). Following the stimulus, a 1,000 ms gray screen appeared, followed by a 2,000-ms image. Participants determined if the sentence's meaning matched the image, using semantically correct (GS, blue), semantically implausible but grammatically correct (GNS, red), or grammatically and semantically incorrect (NGNS, black) sentences, randomized across trials. **Figure B. Grammar Conditions.** Sentences were categorized into NGNS, GNS, and GS trials based on the type of context violation (grammatical violation vs syntax violation).

Misra (2024) Conducted the 4-word language task with N=17 participants with treatment-resistant epilepsy who had individualized SEEG microelectrode arrays. The trials consisted of sentences that fit one of three categories: GS- Grammatical and Syntactical (Ex, "The girls ate cakes"), GNS- Grammar but non-syntactical (Ex, "The cakes ate girls"), and NGNS- Non-grammatical and non-syntactical ("The ate girls cakes"). Participants experienced trials that consisted of a 600 ms fixation period, 875 ms presentation of each of the 4 words, a 1000 ms pause for consolidation of the semantic meaning, and a 1500 ms response period. During the response period, participants were either presented with an image or an audio recording of a sentence being spoken. The task was to determine whether the stimuli presented matched the 4 words they previously saw (Ex, presented with a sentence "The girls ate cakes" and given an image of a girl eating cake, the response should be yes). Overall, it was found that the left lateral orbitofrontal cortex showed some selection for

nouns vs adjectives. Even though participants spoke Mandarin, Hindi, and English, results were consistent across participants regardless of language spoken, indicating the potential universality of linguistic processing

Based on this dataset, (Gillioz, 2024) attempted to identify whether GPT-2-xl's last hidden output layer embeddings were predictive of human linguistic gamma power responses per electrode. The study aimed to use all n=10 English study participants, but was only able to run analyses for one subject (Subject 8) due to time constraints. The analysis consisted of training a ridge regression per electrode to predict SEEG gamma power in NGNS, GS, and GNS settings. Pearson explained variances on 1000 bootstrapped test train splits were used to calculate confidence intervals, and assess significance. The study found that many features, such as word position (Especially from W2 to W4) and context violations (From NGNS and GNS sentences), significantly affected the model alignment with neural activity. Known language regions, such as the middle frontal gyrus and pre-central gyrus, had more stable predictions from the model. From this preliminary analysis, the experimenter was also able to identify electrodes that were selective for different stimulus types (audio, visual, syntax, grammar, etc.). These preliminary results are compelling, but could be further improved by incorporating all N=17 participants and also including an analysis for the alignment of an encoder-only model and an encoder-decoder model (like BERT-base-multilingual-cased and mT5-XL) to better understand the scope of transformer architectures that potentially align with neurological data.

4. Gap

Despite the wealth of past studies utilizing encoding models to predict neural activity from transformer embeddings, few studies had a large enough sample size or examined higher-level linguistic properties like semantic meaning and syntax (Blank et al., 2014; Fedorenko et al., 2016; Gillioz, 2024; Goldstein et al., 2022, 2024; Pereira et al., 2018; Schrimpf et al., 2021; Tuckute et al., 2024). Additionally, there is a need for more evidence demonstrating alignment between LLM contextual embeddings and human neural responses using data that has higher temporal and spatial resolution and less irreducible noise. This analysis would also benefit from

including a wider and more representative sample of sufficiently deep transformer architectures (encoder-only, decoder-only, and encoder/decoder models) to provide a better survey of alignment potential. Additionally, many previous studies make use of naturalistic stimuli such as long-form podcasts as the linguistic stimulus (Deniz et al., 2019; Goldstein et al., 2022, 2024; Hong et al., 2024). However, considering simpler and shorter phrases may allow for greater experimental interpretability and more atomic conclusions regarding biological and artificial representations of syntax and grammar.

5. Justification

To address this gap in the literature, this study will leverage the SEEG 4-word task dataset from (Misra, 2024) to examine whether transformer-based language models (decoder-only GPT-2-XL, encoder-only BERT-base-multilingual-cased, and multilingual mT5-encoder/decoder) show differential alignment between their hidden-layer embeddings and linguistically evoked neural activity. Specifically, we will test which set of layer embeddings (from early, middle, late, or final hidden layers) exhibits stronger correspondence with gamma-band responses embeddings drawn from intermediate layers of each model (e.g., layer 48 of GPT-2-XL and layer 12 of BERT-base-multilingual-cased). Such an analysis would allow for reasoning about the brain's internal model of linguistic processing, especially if high alignment centers cluster in known linguistic-ROIs (Ex, IFG, ATL, STS, etc.). In the same way, using GloVE and Word2Vec embeddings was commonly used to learn about underlying semantic and syntactic structure; swapping in transformer embeddings could boost neurological understanding (Goldstein et al., 2022; Mahajan et al., 2025; Reddy & Wehbe, 2021; Soydaner, 2022; Vaswani et al., 2023). Additionally, reasoning about the impact of experimental time window (consolidation vs word-processing) as well as the impact of context violations (NGNS and GNS) could be used as a tool to understand the time resolution of the human brain in response to violations and subsequently if there is computational alignment.

More specifically, the analysis will also examine how model–brain alignment varies as a function of temporal window (W2, W3, W4, and consolidation phases), transformer layer depth (early, middle, late, and last), and grammar condition: NGNS (non-grammatical, non-syntactic), GNS (grammatical, non-syntactic), or GS (grammatical, syntactic). Neural responses will be evaluated within key linguistic ROIs, including the orbitofrontal cortex (OFC), inferior frontal gyrus, medial temporal lobe, Wernicke's area, and Broca's area. Together, this design enables a more comprehensive assessment of how architectural type, layer hierarchy, temporal dynamics, and grammatical structure jointly modulate the alignment between transformer embeddings and human neural activity. The results of this experiment will allow for novel reasoning regarding the processing of both transformers and the human brain in response to contextual violations.

This project demonstrates an integration of computer science and neuroscience through the use of transformer-based language models (like GPT-2-XL, BERT-base-multilingual-cased, and mT5-XL) to study human language processing. By applying neural encoding techniques, the proposed study will map linguistic features from these models to neural activity, enabling comparisons between artificial and biological language representations. The development of efficient tools for neuroanalysis also showcases how software engineering can potentially advance neuroscience research. Ultimately, this work would enhance understanding of brain-based language processing and suggest potential improvements to AI models by aligning them more closely with neurobiological principles. It would also produce results that could be relevant to creating new tools to allow neuroscientists to explore and simulate neurological data and further our knowledge of neurolinguistic processing.

6. Goals

To achieve this comparison between biological and artificial syntax and grammar representations, the following experimental goals will be set:

1) **Per-Electrode Encoding Models**

a) Fit an encoding ridge regression model per electrode to quantify the impact of combinations of the following variables on model performance between preprocessed sliding-window SEEG gamma power signals and LLM embedding weights:
    i) Modality: Auditory vs. visual presentation during the 1500 ms response period
    ii) Sentence Condition: NGNS (non-grammatical, non-syntactic), GNS (grammatical, non-syntactic), GS (grammatical, syntactic)
    iii) Transformer Layer: Early (layer 2), Middle, Late (third-to-last layer), Last (final hidden output layer)
    iv) Time Window: W2 (875 ms), W3 (875 ms), W4 (875 ms), Consolidation (1000 ms)

b) Notably, a per-electrode model is chosen to allow for each cortical region to be considered independently and to support a study design without anatomical priors. All electrodes will be independently considered as potentially aligning with transformer embeddings and only compared and filtered later to allow for unbiased results. Such a model also allows for greater interpretability of results since only the neural activity of one electrode and time window at a time is ever considered within each model.

2) Architectural Comparisons

a) Compare encoding performance across GPT-2-XL, BERT-base-multilingual-cased, and mT5, evaluating the potential impacts of model differences (architecture, number of parameters, number of hidden layers, etc.) on neural alignment.

3) Regional Parcellation Analysis

a) Examine whether highly aligned electrodes cluster within canonical language-related regions (e.g., Inferior Frontal Gyrus (IFG), Wernicke's area, Anterior Temporal Lobe-ATL, and Posterior Temporal Lobe-PTL) relative to non-linguistic regions (e.g.,

precentral gyrus). If the model predicted classical language regions better than surrounding regions, it would present increased evidence for linguistic model-brain alignment compared to a model that is invariant to region.

4) Interaction Effects

a) Quantify higher-order interactions amongst time window × grammar condition × modality to determine whether model–brain alignment depends on specific representational stages or linguistic constraints.

7. Hypotheses

1) Architectural Hypothesis

a) mT5-XL is predicted to show the strongest neural alignment among the three models, given evidence that brain encoding performance scales consistently with parameter count (Antonello et al., 2023; Hong et al., 2024). BERT-base-multilingual-cased, with the fewest parameters, is predicted to show the weakest alignment. Whether mT5's encoder-decoder architecture provides any additional advantage over parameter count alone remains an open question (as there are very few studies to our knowledge that include mT5 embeddings)

2) Layer Hierarchy Hypothesis

a) The last hidden output layer and late transformer layers will show stronger alignment with neural responses during GS (grammatical + syntactic) conditions, whereas earlier layers will show relatively stronger alignment during NGNS conditions, reflecting lower-level lexical or surface-form processing. This would result as a direct result of hierarchical processing and gradual building of complex linguistic features (ex, grammar)

across transformer hidden layers (Alleman et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022).

3) **Temporal Dynamics Hypothesis**
    a) The consolidation and W4 window will exhibit the highest concentration of significant encoding electrodes for GS sentences, reflecting integration and higher-level compositional processing. Earlier windows (W2–W3) may show stronger effects for NGNS or GNS conditions, reflecting incremental processing. This would be a result of gradual, time-resolved sentence processing in the human brain from basic syntax to grammar and semantics (Friederici, 2011).
4) **Regional Specificity Hypothesis**
    a) Language-selective regions (Ex, IFG, Wernicke's area, anterior temporal lobe, and posterior temporal lobe) will exhibit higher encoding performance (Pearson's r and R^2) compared to non-linguistic control regions (e.g., precentral gyrus), consistent with prior findings (Blank et al., 2014; Caucheteux & King, 2022; Fedorenko et al., 2016; Misra et al., 2024; Schrimpf et al., 2021)

8. Materials and Methods

8.1 Participants

Data from this experiment came from N =17 participants (7 male, ages 13-50 years old, 3 left-handed) individuals with pharmacologically treatment-resistant epilepsy while undergoing seizure foci monitoring. See full participant demographics in Figure 3. Experiments were conducted at Children's Hospital Boston (CHB), the Cleveland Clinic (Ohio), or Taipei Veterans General Hospital (TVGH). Experimental procedures passed hospital institutional review boards (IRB), and informed consent was obtained from all subjects and or legal

guardians. Electrode implantation was solely for medical purposes, as this experimental procedure did not interfere with the patient's standard medical care. These participants underwent the 4-word task experiment detailed in (Misra et al., 2024). Both the SEEG neural recordings and sentence data used in this experiment were drawn from this dataset.

8.1.1 Electrode Location and Neural Recordings

Participants had intracranial depth electrodes (Ad-Tech, Racine, WI, USA) as a component of their preexisting healthcare regimen. Intracranial field potentials were recorded using XLTEK (Oakville, ON, Canada), Bio-Logic (Knoxville, TN, USA), Nihon Kohden (Tokyo, Japan), or Natus (Pleasanton, CA, USA) systems. Sampling rates were 2048 Hz at BCH and TVGH, and either 1024 Hz or 512 Hz at BWH. Bipolar montage was utilized to reduce noise within neural signals. No seizures occurred during this analysis.

Electrode location information was obtained by taking computed tomography (CT) scans of each patient after surgery and digitally aligning them with high-resolution T1-weighted MRI scans of the brain taken before surgery. This image alignment was done using the iElvis Software package, and the neuronal surfaces were reconstructed utilizing FreeSurfer. This analysis allowed every electrode to be localized to specific, customized coordinates along each participant's cortical surface, defined by the Desikan-Killianay Atlas. Electrodes were excluded from analysis based on signal quality and whether they were located at epileptic foci or diseased tissue (as determined by clinical teams). After this filtration, 1,801 electrodes across participants remained, 844 of which were located in gray matter, and 719 in white matter.

Subject	Age	Gender	Language	Handedness	#Trials	%Correct	#Electrodes
1	13	M	HI	R	457	50%	134
2	15	F	TW	R	942	79%	35
3	19	M	EN	R	901	97%	174
4	37	F	EN	R	913	97%	78
5	40	M	EN	R	906	99%	73
6	21	M	TW	L	904	80%	64
7	30	M	TW	R	952	70%	78
8	42	F	EN	R	909	78%	154
9	27	F	EN	R	882	94%	62
10	32	F	EN	R	900	84%	141
11	30	F	EN	L	906	97%	72
12	25	M	EN	L	900	84%	92
13	33	F	EN	R	900	97%	65
14	50	F	EN	R	972	96%	117
15	20	M	EN	R	862	86%	65
16	20	F	EN	R	909	78%	81
17	20	F	EN	R	907	95%	78
						TOTAL	**1563**

Figure 3. Participant Demographics. Abbreviations: Hindi (HI), Taiwanese (TW)- referring to a specific dialect of Mandarin Chinese, English (EN), left-handed (L), right-handed (R)

8.2 Neurological Features

8.2.1 SEEG Signal Pre-processing

Importantly, prior to fitting any regressions, neural and behavioral data had to be extensively preprocessed to extract the gamma signal band, remove artifacts, and synchronize behavioral and neurological features.

To achieve this, a comprehensive signal processing pipeline was developed in Python to allow for later use of machine learning libraries such as Scikit learn and Transformers to be integrated into analysis later in the experiment (See the link). To facilitate this analysis, first, within MATLAB, the experimental data for each participant (Downloaded from a private, Kreiman lab proprietary Dropbox folder) had to be converted to a struct datatype to allow for mat files to be loaded in by scipy and then converted into pandas dataframes. For each participant, the following dataframes were loaded:

- Trig_list: Contains behavioral data, including sentences and timestamps of audio and visual experimental triggers
- Hdr_table: Contains neurological data (dimensions: timestamps x electrodes) and neurological metadata (sampling rate, recording frequency, etc.)

To facilitate the loading of and pre-processing of behavioral and neurological data, a Subject class was implemented in Python to allow for the instantiation of subject objects that each contained unique IDs and neurological and behavioral records (Github link https://github.com/HelloUniversePyC/GPT-T5-BERT-Alignment-4WTSEEGData). All pipeline steps were run within a Python 3 environment instantiated within the root level directory. After each subject object was initialized with a subject number and a directory, the preprocessing consisted of the following steps for every participant:

1) Loading and initial pre-processing of behavioral data

The trigg_list is loaded as an attribute of the object (Field of the curr_state table) and converted into a pandas dataframe to allow easy lookup of behavioral data.

2) Loading in of neurological data

The HDR table was loaded in using the scipy.io.loadmat function while h5py.File was utilized to load in the substantially larger record data matrix (timestamps x electrodes). A chunk size of 100000 bytes at a time is employed to avoid excessive memory usage when loading in subject files. Memory mapping (via the numpy library) to alleviate the load on RAM while loading in 10-20 GB neural data files. During the initialization loop for all subjects, the Python gc library was utilized to manually employ garbage collection of this raw record frame in memory for all participants (as the filtered record data will be prioritized) to also conserve memory. Both the record data and the HDR table were converted to pandas dataframes to simplify later computation.

7.2.2 Gamma Signal Band Feature Extraction

The `pre_process_gamma` function extracted signals within the gamma frequency range (30-100 Hz), which has been shown in prior studies to correlate with linguistic consolidation activity. `pre_process_gamma` function extracted signals within the gamma frequency range (30-150 Hz), which has been shown in prior studies to correlate with linguistic consolidation activity. `pre_process_gamma` function extracted signals within the gamma frequency range (30-150 Hz), which has been shown in prior studies to correlate with linguistic consolidation activity (Hudson & Jones, 2022).

1) Bipolar Montage: Raw voltage signals were bipolar referenced by computing the difference between spatially adjacent electrode pairs to reduce common-mode noise:

$$s_{\text{bipolar}}^{(i,j)}(t) = s_i(t) - s_j(t)$$

where $s_i(t)$ and $s_j(t)$ are the raw voltage recordings from electrodes i and j at time t.

2) Bandpass Filtering: A 4th-order Butterworth bandpass filter was applied using the Scipy library to isolate gamma frequencies (30-150 Hz). The filter was implemented using scipy's `butter` function with parameters `btype='band'` and `output='sos'` (second-order sections) for numerical stability. The squared magnitude response is:

$$|H(\omega)|^2 = \frac{1}{1 + \left(\frac{\omega^2 - \omega_0^2}{\omega \cdot BW}\right)^{2n}}$$

where $\omega_0 = 2\pi\sqrt{f_{\text{low}} \cdot f_{\text{high}}}$ is the geometric center frequency, $BW = 2\pi(f_{\text{high}} - f_{\text{low}})$ is the bandwidth ($f_{\text{low}} = 30$ Hz, $f_{\text{high}} = 100$ Hz), and $n = 4$ is the filter order.

$$s_{\text{filtered}}(t) = \text{sosfiltfilt}(H_{\text{sos}}, s_{\text{bipolar}}(t))$$

3) Notch filtering: Line noise at 90 Hz and its harmonics was removed using MNE's `notch_filter` function, which applies notches in the frequency domain:`notch_filter` function, which applies notches in the frequency domain:`notch_filter` function, which applies notches in the frequency domain:

$$S_{\text{notched}}(f) = S_{\text{filtered}}(f) \cdot \prod_{k=1}^{K} (1 - G_k(f))$$

where $G_k(f)$ represents a notch centered at 90 Hz.

For all of these steps, processing was parallelized across batches of 20 electrodes using the joblib library with forced garbage collection (`gc.collect()`) after each batch to prevent memory overflow. The output was a matrix $\mathbf{S} \in \mathbb{R}^{T \times E}$ containing filtered gamma-band signals for E electrodes over T time points.

7.2.3 Behavioral and Neurological Trigger Alignment

To accurately align behavioral events with neural activity, trigger timestamps from the trig_frame were matched to the `record_filter` dataframe timestamps. After standardizing trigger keys and correcting errors, the `align_triggers_with_word_timing` function computed trial boundaries using the experimental timing structure: 600ms fixation, 875ms per word presentation (4 words), and 1000ms consolidation period (the 1500ms answer period was omitted from this analysis).`record_filter` dataframe timestamps. After standardizing trigger keys and correcting errors, the `align_triggers_with_word_timing` function computed trial boundaries using the experimental timing structure: 600ms fixation, 875ms per word presentation (4 words), and 1000ms consolidation period (the 1500ms answer period was omitted from this analysis).

For each trial, the temporal boundaries were defined as:

$$t_{\text{start}} = t_{\text{WORD1}}^{\text{preOnset}}$$

$$t_{\text{end}} = t_{\text{WORD4}}^{\text{postTrigger}} + 1000 \text{ ms}$$

where $t_{\text{WORD1}}^{\text{preOnset}}$ and $t_{\text{WORD4}}^{\text{postTrigger}}$ correspond to the `system_timePreOnset` and `system_timePostTrigger` fields from the trigger list. When available, the experiment start time from the `hdr` table was used to adjust timestamp offsets:

$$t_{\text{adjusted}} = t_{\text{system}} - t_{\text{exp\_start}}$$

The `Var1` columns in the trigger frame provided alternative timestamps for validation. The output `trial_frame` contained: trial index, t_{start}, t_{end}, trial duration, sentence type (GS, GNS, NGNS), and modality (auditory or visual). `Var1` columns in the trigger frame provided alternative timestamps for validation.

8.2.4 Sliding Window Neurological Features

1) Epoch Extraction: Continuous gamma-filtered signals were segmented into trial-locked epochs using the `extract_epochs` function. Each epoch was time-locked to WORD1 onset (Word 1 was treated as timepoint 0) and spanned from t_{pre} ms before to t_{post} ms after the trigger, yielding a 3D array:

$$\mathbf{E} \in \mathbb{R}^{N_{\text{trials}} \times E \times T_{\text{epoch}}}$$

2) Multi-Taper Spectrogram: Time-frequency decomposition was performed using multi-taper spectral analysis (`mtspecgramc`) with the following parameters: `mtspecgramc`) with the following parameters: `mtspecgramc`) with the following parameters:

- Window size: $W = 200$ ms
- Step size: $\Delta t = 100$ ms (larger step size is adequate for sentence-level decoding)

$f \in [30, 150]$ Hz

- Time-bandwidth product: $NW = 3$
- Number of tapers: $K = 5$
- Sampling rate: $f_s = 1000$ Hz

For each time window w centered at time t_w, the multi-taper power spectrum was computed as: w centered at time t_w, the multi-taper power spectrum was computed as: w centered at time t_w, the multi-taper power spectrum was computed as:

$$P_w(f) = \frac{1}{K} \sum_{k=1}^{K} \left| \int_{t_w - W/2}^{t_w + W/2} s(t) h_k(t - t_w) e^{-i2\pi f t} dt \right|^2$$

where $h_k(t)$ are the discrete prolate spheroidal sequences (Slepian tapers) satisfying: $h_k(t)$ are the discrete prolate spheroidal sequences (Slepian tapers) satisfying: $h_k(t)$ are the discrete prolate spheroidal sequences (Slepian tapers) satisfying:

$$\int_{-W/2}^{W/2} h_k(t) h_j(t) dt = \delta_{kj}$$

3) Baseline Normalization: Gamma power was z-score normalized relative to a baseline period (−200 to 0 ms pre-stimulus) on a per-trial-type basis:

$$P_{\text{norm}}(t, f) = \frac{P(t, f) - \mu_{\text{baseline}}(f)}{\sigma_{\text{baseline}}(f)}$$

where:

$$\mu_{\text{baseline}}(f) = \frac{1}{N_{\text{baseline}}} \sum_{t \in [-200,0]} P(t, f)$$

$$\sigma_{\text{baseline}}(f) = \sqrt{\frac{1}{N_{\text{baseline}}} \sum_{t \in [-200,0]} \left(P(t, f) - \mu_{\text{baseline}}(f)\right)^2}$$

4) Feature Extraction The `extract_sliding_window_features` function returned a DataFrame $\mathbf{F} \in \mathbb{R}^{M \times (E+D)}$ with M time windows, E electrode power features, and D metadata columns (trial index, modality, sentence type). For regression analysis, gamma power was averaged across the entire stimulus buffer W1- consolidation window:

$$y_{\text{trial}}^{(e)} = \frac{1}{N_w} \sum_{w : t_w \in [0,4500]} P_w^{(e)}$$

where e indexes electrodes and N_w is the number of windows in the consolidation period. This yielded the target vector $\mathbf{y} \in \mathbb{R}^{N_{\text{trials}}}$ for each electrode used in ridge regression against sentence embeddings $\mathbf{X}$.e indexes electrodes and N_w is the number of windows in the consolidation period.

7.3 Artificial Intelligence Features

8.3.1 Models

The models utilized in this study were pretrained vanilla, non-instruction-tuned GPT-2-XL (48 total hidden layers, output layer dimension 4800), BERT-base-multilingual-cased (24 hidden layers, output layer dimension 1024), and MT5-XL (24 hidden layers, output layer dimension 2048) obtained from the Hugging Face Transformers library. See more detailed model information in Figure 2 above.

8.3.2 Transformer Embedding Extraction

Given that neurological data was recorded from electrodes throughout the brain, in order to compare artificial and natural intelligence, it is natural to also compare embeddings from multiple layers within each transformer. Embeddings were extracted from 4 checkpoints: 1) early hidden layer (3rd hidden layer), 2) middle hidden layer, 3) late hidden layer (3rd to last hidden layer), and 4) Last hidden layer. Past studies have indicated high correlations between the last hidden layer and linguistic consolidation. However, given the lack of positional encodings in the last hidden layer, high correlations with syntax may be found in earlier layers (Alleman et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022). Justifying the choice of also including middle and early layers. To extract embeddings, the PyTorch library was employed. Sentences were extracted from the trial_df to feed into auto-trained tokenizers specialized by model attributes. Before extracting embeddings, for English participants, given that the word “The” was always the first word in any sentence (GS, GNS, or NGNS), this word was removed to prevent the model from overfitting to this repeated feature instead of higher-level syntactical and grammatical patterns. To allow for comparisons across subjects, only W2-W4 were considered in the final analysis.

Regarding embedding extraction, all models output hidden states as a tuple where index 0 contains the embedding layer and subsequent indices contain transformer layer outputs, so the

extract_multilayer_embeddings() function adds 1 to the layer index to access the correct layer (e.g., hidden_states[7] for the 6th transformer layer). Once the hidden states are extracted from the specified layer, the models differ in how they convert token-level representations into sentence-level embeddings. GPT-2, being autoregressive where each token encodes information about previous tokens, takes the first N-1 tokens from the hidden states and flattens them into a single vector, resulting in embeddings of dimension (n_words-1) × hidden_dim (e.g., 2,304 dimensions for 3 tokens with 768-dimensional hidden states). BERT uses only the first token's embedding (the special [CLS] token), which is specifically designed to aggregate sentence-level information during pretraining, maintaining the original hidden dimension of 768. T5, lacking a dedicated sentence token, employs mean pooling across all non-padding tokens by multiplying the hidden states with the attention mask, summing across the sequence dimension, and dividing by the number of non-padding tokens, also resulting in a 768-dimensional embedding. This embedding extraction logic was optimized to run on gpu to improve runtime performance. Embeddings were vertically stacked into a matrix (sentences x embedding dimension), which serves as the X design matrix in the ridge regression. The bias trick was employed to roll the scalar bias term into this X design matrix. See Figure 4 below for the usage of these embedding vectors in the regression pipeline.

8.4 Ridge Regression Model

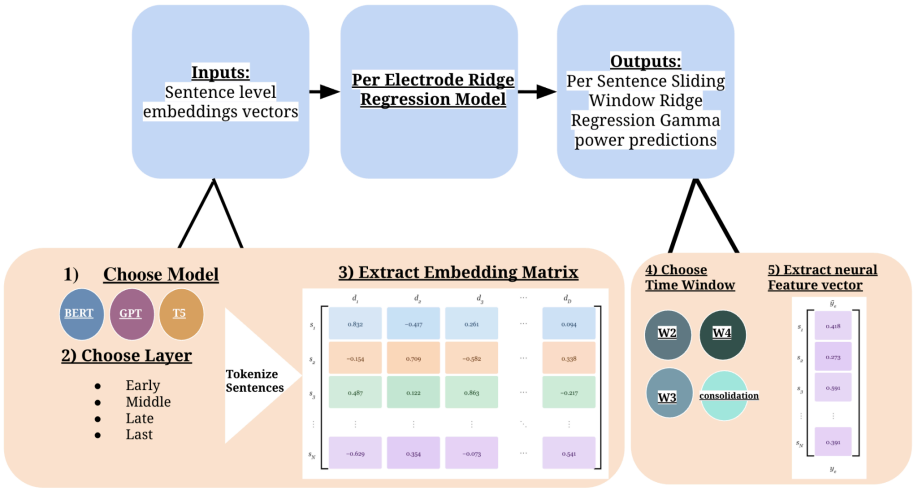


Figure 4. Model Training Visualization. Feature and target generation for per-electrode ridge regression models

A Ridge regression class utilizing RidgeCV and Ridge from Scikit-learn was used to predict within-subject per-electrode neural activity from language model embeddings:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}$$

which has the closed-form solution:

$$\boldsymbol{\beta}^* = \left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^T\mathbf{y}$$

where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the design matrix of sentence embeddings (with augmented bias column) extracted from one of three transformer models (BERT, GPT-2-XL, or mT5-XL) at one of four layer depths (early: layer 3, middle, late: 3rd-to-last, or last layer), filtered by one of six experimental conditions (overall [unfiltered], grammatical sentences [GS], grammatical non-sentences [GNS], or non-grammatical non-sentences [NGNS]), and neural data from 1 of 4 time windows (W2, W3, W4, or Consolidation) This resulted in:

$$3 \text{ models} \times 4 \text{ layers} \times 3 \text{ conditions} \times 4 \text{Time Windows} = 144$$

unique regression analyses per electrode. $\mathbf{y} \in \mathbb{R}^n$ contains trial-averaged gamma power (30-150 Hz) in thein one of 4 time windows (875 ms W2/W3/W4 or 1000 ms consolidation), and $\lambda \in [10^1, 10^6]$ is the

regularization hyperparameter selected (8 values considered) via 5-fold group cross-validation (grouping by sentence identity). Originally, 10 folds were utilized, but this was reduced to 5 for computational efficiency. The loss function minimized is:

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

Model performance was evaluated on held-out test data (80/20 train test split) using r^2 as a performance metric:

$$R^2 = 1 - \frac{\sum_{i=1}^{n_{\text{test}}}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{\text{test}}}(y_i - \bar{y})^2}$$

All model fitting was run on CPU with the implementation of batching, multithreading, and forking as runtime optimizations (due to limited GPU access)

8.5 Permutation Testing

To determine statistical significance, a permutation test was conducted for each electrode. The null distribution was constructed by randomly permuting the neural activity vector $\mathbf{y}$ (gamma power values) 100 times while keeping the embedding matrix $\mathbf{X}$ fixed, refitting the ridge regression within-subject for each permutation:

$$R^2_{\text{null}}{}^{(k)} = 1 - \frac{\sum_{i=1}^{n_{\text{test}}}\left(y_{\pi^{(k)}(i)} - \bar{y}_i\right)^2}{\sum_{i=1}^{n_{\text{test}}}\left(y_{\pi^{(k)}(i)} - \bar{y}\right)^2},\ k = 1, \ldots, 100$$

where $\pi^{(k)}$ represents the k-th random permutation. The empirical p-value for each electrode was computed as:

$\pi^{(k)}$ represents the k-th random permutation. The empirical p-value for each electrode was computed as:

$$p = \frac{1 + \sum_{k=1}^{100} \mathbb{1}(R^2_{\text{null}}{}^{(k)} \geq R^2_{\text{observed}})}{101}$$

Notably, initially 1000 shuffles per model were planned, but this number was reduced to 100 due to computational limitations (the training loop was run fully on CPU due to high-performance computing cluster access limitations).

8.6 False Discovery Rate (FDR) P-Value Correction

Family-wise FDR correction was used to correct for false positives within the p-values, given the large number of tests per electrode (144 tests per electrode). A family was defined as all the regressions with a unique combination of the experimental variables. Hence:

$$\mathcal{F} = \{\text{model}, \text{layer}, \text{window}, \text{condition}\}$$

where model $\in \{\text{GPT}, \text{BERT}, \text{T5}\}$

layer $\in \{\text{early}, \text{middle}, \text{late}, \text{last}\}$,

window $\in \{W_2, W_3, W_4, W_{\text{cons}}\}$,

and condition $\in \{\text{GNS}, \text{NGNS}, \text{GS}\}$.

For each family $\mathcal{F}_k$ containing m such electrodes, the Benjamini-Hochberg procedure first orders the m uncorrected p-values from smallest to largest:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$$

FDR correction then recalculates the significance threshold based on the maximum p-value. The null hypothesis is rejected for all $H_{(i)}$ up to:

$$i^* = \max \left\{ i : p_{(i)} \leq \frac{i}{m} \cdot \alpha \right\}$$

where $\alpha = 0.05$.

The adjusted p-values are then computed as:

$$\tilde{p}_{(i)} = \min\left(\min_{j \geq i}\left\{\frac{m}{j} \cdot p_{(j)}\right\},\ 1\right)$$

Hence, all p-values that do not meet the stricter threshold are rejected, reducing the number of false positives. Finally, a filter of $r^2 > 0.1$ was applied to these FDR-corrected p-values (post correction) to control for effect size.

9. Results

In this section, the terms "model" and "electrode" are used interchangeably since each model is fitted on only the neurological data from a single electrode. Additionally, all results plotted here include all high-alignment models defined as having explained variance greater than 0.1. This filter was only applied after the full distribution of models was used for FDR correction to control for effect size. Additionally, mT5-XL will be referred to as T5, GPT-2-XL as GPT, and BERT-base-multilingual-cased as BERT. Due to the relatively small number of significant models in Figure 6, both uncorrected (prior to FDR correction) and corrected (after FDR correction) significant model counts are visualized. See the appendix Figure 13 for the full, unfiltered model performance distribution.

Firstly, it is surprising that almost all FDR corrected significant results come from alignment of right hemisphere regions of interest, except the Inferior Temporal Sulcus (L-ITS). There is also a cluster of significant results (n = 6) within the inferior frontal sulcus (IFS), all being in the NGNS condition except for one within GNS (Figure 5A). The majority of significant results come from participant 3, who is the subject with the most electrodes and the fewest rejected electrodes (due to low signal quality or artifacts). Within the uncorrected p-value population (n=66), significant results with high effect size are more distributed across participants, but participant 3 still dominates. Notably, although the FDR correction method was employed, the number of significant results here was far less than that expected (Within 297,112 tests and a 5% false positive rate, one would expect about 14,855 false significant positives), which speaks well to the controls present in the student

design. The largest significant effective size is $r^2 = 0.275$ within the IFS NGNS condition. Although some models achieved higher performance ($r^2 = 0.414$) more moderate correlations were more likely to survive the permutation testing and FDR correction. Notably the FDR-significant maximum explained variance achieved of 0.27 correlates to a Pearson-r value of $\approx |0.51|$. While the maximum non-significant explained variance of 0.414 was found amongst the models test explained variance is $\approx |0.64|$ in person-r. Although the neural pipeline also recorded correlation coefficient, due to computational limitations, only the explained variance was formally recorded and shuffle tested. Hence, it will be used as the main performance measure across this section.

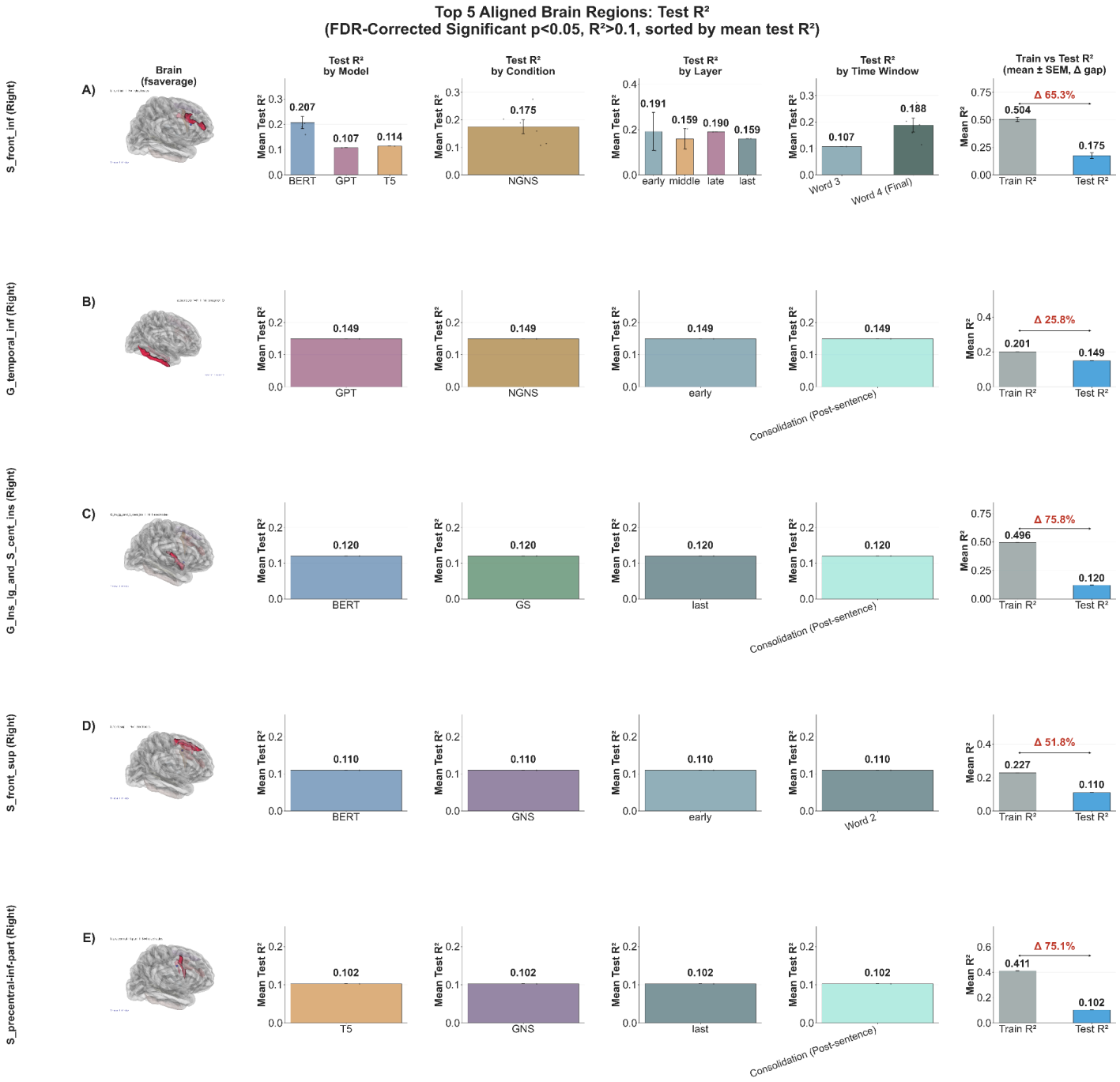


Figure 5. Top-aligning frontotemporal brain regions correlates with transformer embeddings fitted with syntactical violations (NGNS and GNS) in the context of widespread high train-test gap percentages. Significant encoding results were identified from **395 electrode–model combinations**, of which **10 combinations met the significance threshold** ($p_f dr < .05$ and test $R^2 > 0.10$. These effects were localized to **five right-hemisphere cortical regions**, shown in descending order of mean test R^2. All regions displayed **FDR-corrected significance (*p* < .001)** across the reported electrodes.**(A) Right inferior frontal sulcus**. This region showed the **highest mean predictive performance** (mean test R^2=0.175, mean train R^2=0.504, N = 2 electrodes). Significant models included **GPT** early-layer embeddings in the **NGNS** condition with **word window 3 (W3)**, and **BERT** early-layer embeddings in the **NGNS** condition with **word window 4 (W4)**.**(B) Right inferior temporal gyrus**. One electrode showed significant encoding performance (mean test R^2=0.149, mean train R^2=0.201). Significant models were obtained using **GPT** embeddings in the **NGNS** condition during **full consolidation**, including both **early** (*p* < .001) and **middle** (*p* < .05) layer representations.**(C) Right insular cortex /**

central insular sulcus. One significant model was observed (mean test R^2=0.120, mean train R^2=0.496), using **BERT** embeddings in the **GS** condition from the **final layer** during **full consolidation**.**(D) Right superior frontal sulcus.** One electrode showed significant predictive performance (mean test R^2=0.110, mean train R^2 =0.227), associated with **BERT early-layer** embeddings in the **GNS** condition with **word window 2 (W2)**. **(E) Right inferior precentral sulcus**. One significant model was identified (mean test R^2=0.10, mean train R^2=0.411), corresponding to **T5 final-layer** embeddings in the **GNS** condition during **full consolidation**. Panels display the anatomical locations of each region on the cortical surface. Regions are ordered by **mean test** R^2 across electrodes. Train–test performance gaps are reported to illustrate model generalization across electrodes.

Given that it is aimed to find both fdr-corrected significant and large enough effective size, Figure 5 focuses on showing the top 5 significant explained variance effect sizes per brain region. Interestingly, the majority of high effect sizes occurred within the right hemisphere. Many occur in classical frontotemporal language regions, but there are also correlations within the nonclassical regions, like the insula, that warrant further analysis within the discussion. Importantly, there is a high degree of overfitting within these top-aligned regions, with gaps as high as 75.1% in the pre-central (Which was initially cited as a linguistic non-corrected potential control, making its presence in the highest effect sizes unexpected). Interestingly, the models fitted for these regions did not choose the maximum alpha value within the allotted range during cross validation (10^1 to 10^6) which suggests the issue of overfitting could not be solved by improved hyperparameter tuning. However, regions like the Inferior temporal gyrus (ITG) and superior frontal sulcus (SFS) that show lower degrees of overfitting may be more suitable candidates to consider for artificial and neural alignment hypotheses.

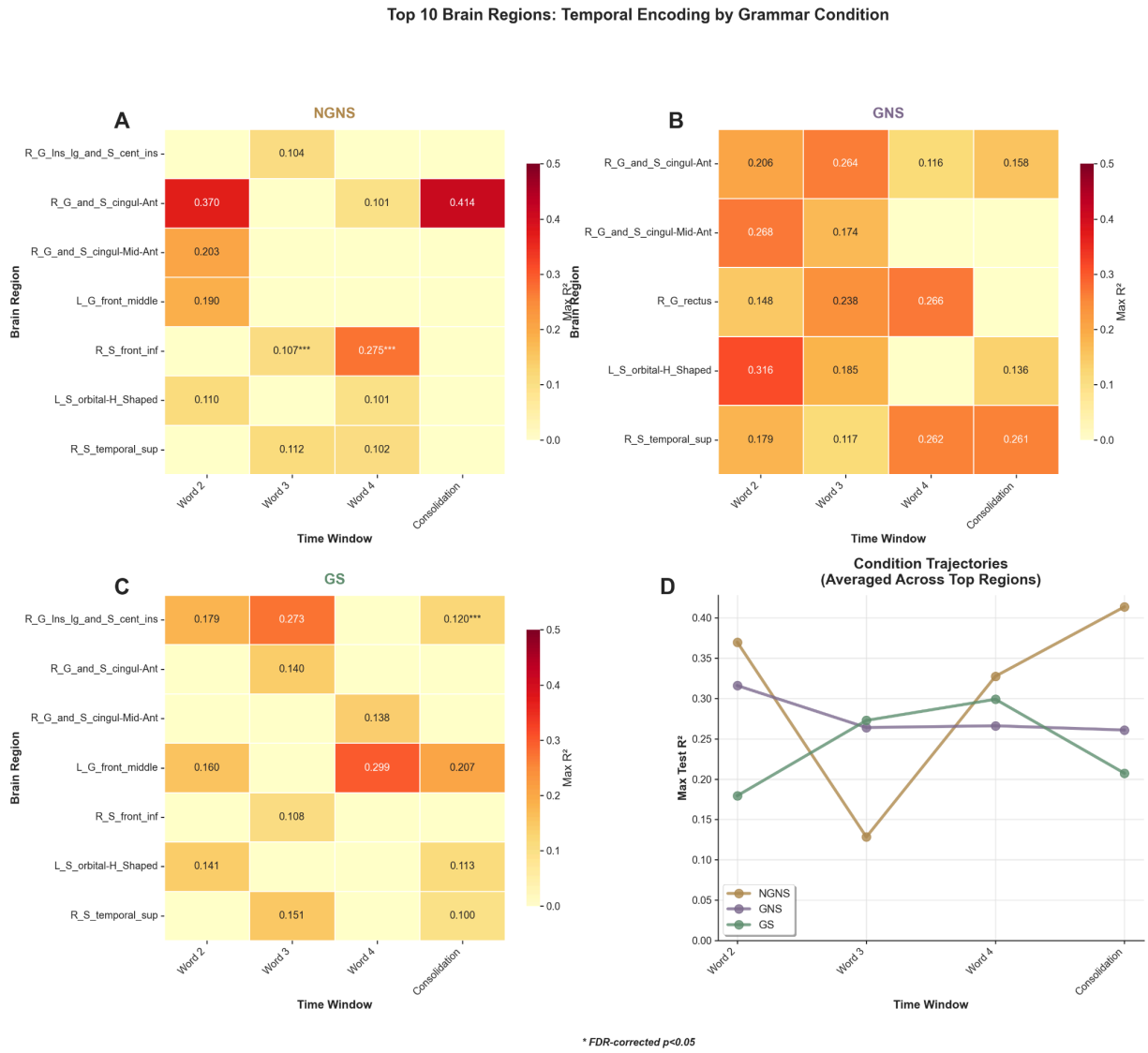


Figure 6. Significant model performance occurs in the Cingulate Cortex during Word 2 and Frontotemporal regions across W3-4 and the Insula during Consolidation. Model performance over ROIs per time window and grammar condition. Labels: Anterior Cingulate Cortex (G_and_S_cingul-Ant), Mid-Anterior Cingulate Cortex (G_and_S_cingul-Mid-Ant), Long Insular Gyrus and Central Insular Sulcus (G_Ins_lg_and_S_cent_ins), Inferior Frontal Sulcus (S_front_inf), Superior Frontal Gyrus (G_front_sup), Middle Frontal Gyrus (G_front_middle), Gyrus Rectus / Orbital Frontal Cortex (G_rectus), H-Shaped Orbital Sulcus (S_orbital-H_Shaped), Inferior Temporal Gyrus (G_temporal_inf), Superior Temporal Sulcus (S_temporal_sup), Inferior Temporal Sulcus (S_temporal_inf) **Panel A) ROI model performance (Max r^2) over the time window for models fitted with NGNS sentence embeddings.** The highest significant model explained variances occur during W4 within the inferior frontal sulcus ($p < 0.001$***). **Panel B) ROI model performance (Max r^2) over the time window for models fitted with GNS sentence embeddings.** Notably, there were significant models within GNS (see Figure 3), but they were too small in magnitude to make it into the top effect sizes within this Figure. **Panel C) ROI model performance (Max R^2) over the time window for models fitted with GS sentence embeddings.** The highest significant performance occurred during consolidation within the Long Insular Gyrus (Ig) and the Central Sulcus of the Insula ($p < .001$***). **Panel D. line graph of trajectories of model performance over time windows colored by grammar condition.**

Within Figure 6A, models fitted with NGNS embeddings showed maximally significant performance during W4 within the Inferior Frontal Sulcus ($p < 0.001$***), indicating strong alignment with frontotemporal processing at this stage. Panel B shows that GNS embeddings produced significant but smaller-magnitude effects across several ROIs, which did not reach the top effect sizes in this Figure. Panel C demonstrates that GS embeddings yielded peak model performance during the consolidation phase in the Long Insular Gyrus and Central Insular Sulcus ($p < 0.001$***), surprisingly suggesting the potential for insular involvement in post-lexical integration (will be developed more in the discussion).

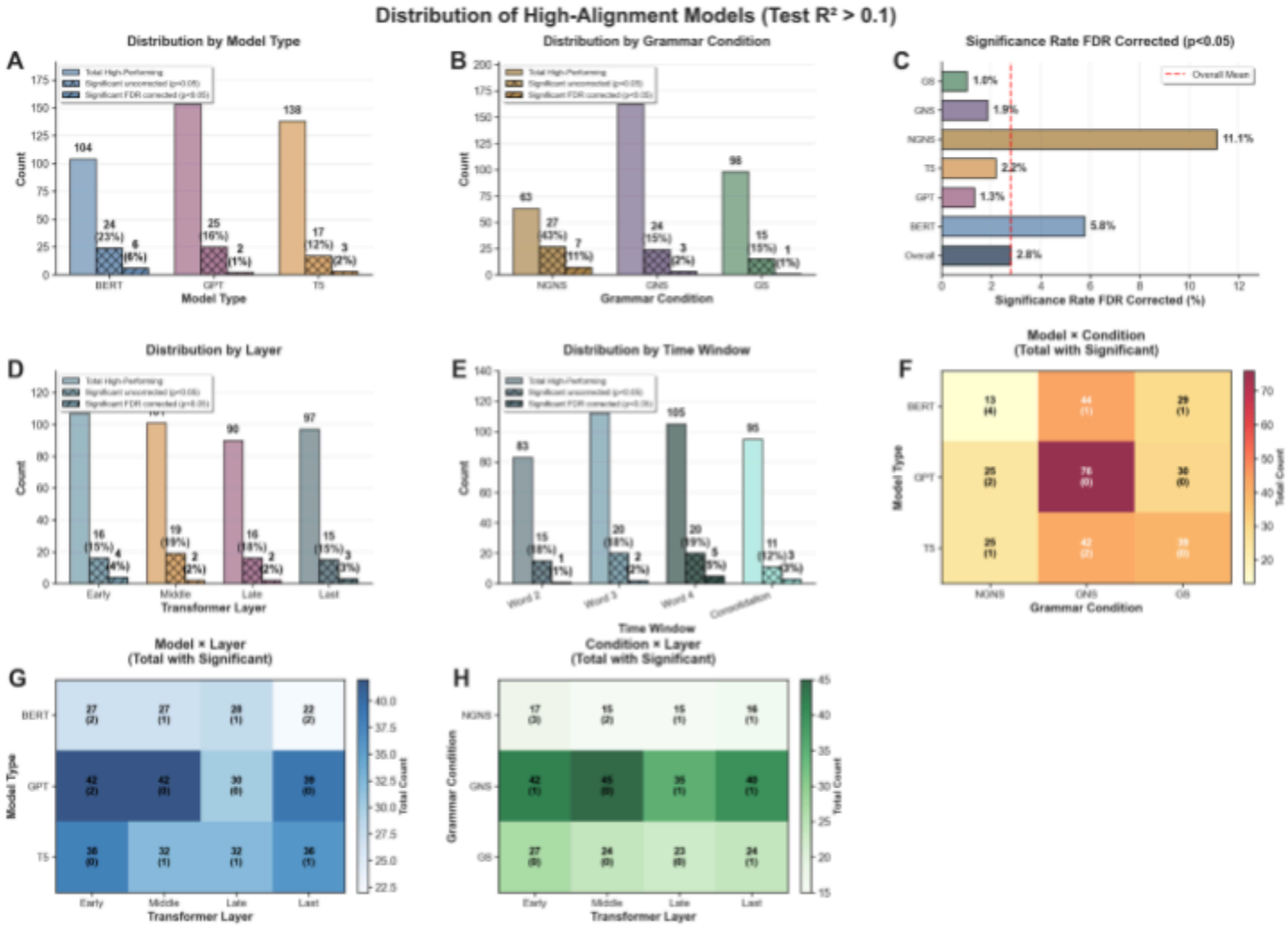


Figure 7. Significance summary (FDR Corrected and Uncorrected) High Alignment Models (R^2 > 0.1). The highest Model Significance rate was found within the BERT and NGNS Condition. Out of the total performing models, all models were filtered on the criteria $r^2 > 0.1$, which represent n= 395 total models**Panel A) Number of significant vs non-significant electrodes per model. Colored by the transformer model.** GPT has the highest number of uncorrected significant models, but BERT has the highest number of corrected significant models. **Panel B) Count of significant models by sentence type (NGNS, GNS, GS)**. **Colored by grammar conditions.** The most significant models come from regressions fitted with NGNS sentence embeddings for both uncorrected and FDR-corrected significant model counts. **Panel C) Visualization of the percentage of significant electrodes by Grammar Condition and Model type. Colored by grammar condition and transformer model. Panel D) Barplot of significant electrodes by transformer layer. Colored by layer.** Middle-layer embeddings are included in most of the uncorrected significant models. However, the majority of FDR-corrected significant models were fitted with early-layer embeddings. **Panel E) Distribution of significant models by time window.** The majority of significant electrodes (uncorrected and corrected) came from regressions with Word 4 neural signals as the target variable. **Panel F) Heat map representation of the interaction between Model type, sentence type, and number of significant electrodes. Colored by significance rate.** Notably, the highest significant electrodes occur with models fitted with the NGNS sentence embeddings type for all transformer architectures. **Panel G) Heat map of interaction between layer embeddings and model FDR-Corrected significant electrodes. Colored by FDR-corrected significance rate. Panel H) Heat map representation of the interaction between sentence type and the transformer embedding layer. Colored by FDR-Corrected significance rate.** The most significant models are found when regressions are fitted with early-layer NGNS sentence embeddings.

Within Figure 7 we see a representation of the distribution of FDR-corrected significant models compared to uncorrected significant models as they pertain to the main experimental variables (layer, time window, model, grammar condition). The FDR correction greatly reduced the number of significant models from n=66 down to n = 11. Within these 11 significant models, the majority were fit with BERT embeddings (Figure 7A). Additionally, most fdr-corrected significant models had NGNS sentence embeddings within the design matrix (Figure 2B). Figure 7C demonstrates how much NGNS outnumbers models fit using other grammar conditions (GNS and GS), with n = 395 high alignment models with explained variance over 0.1.

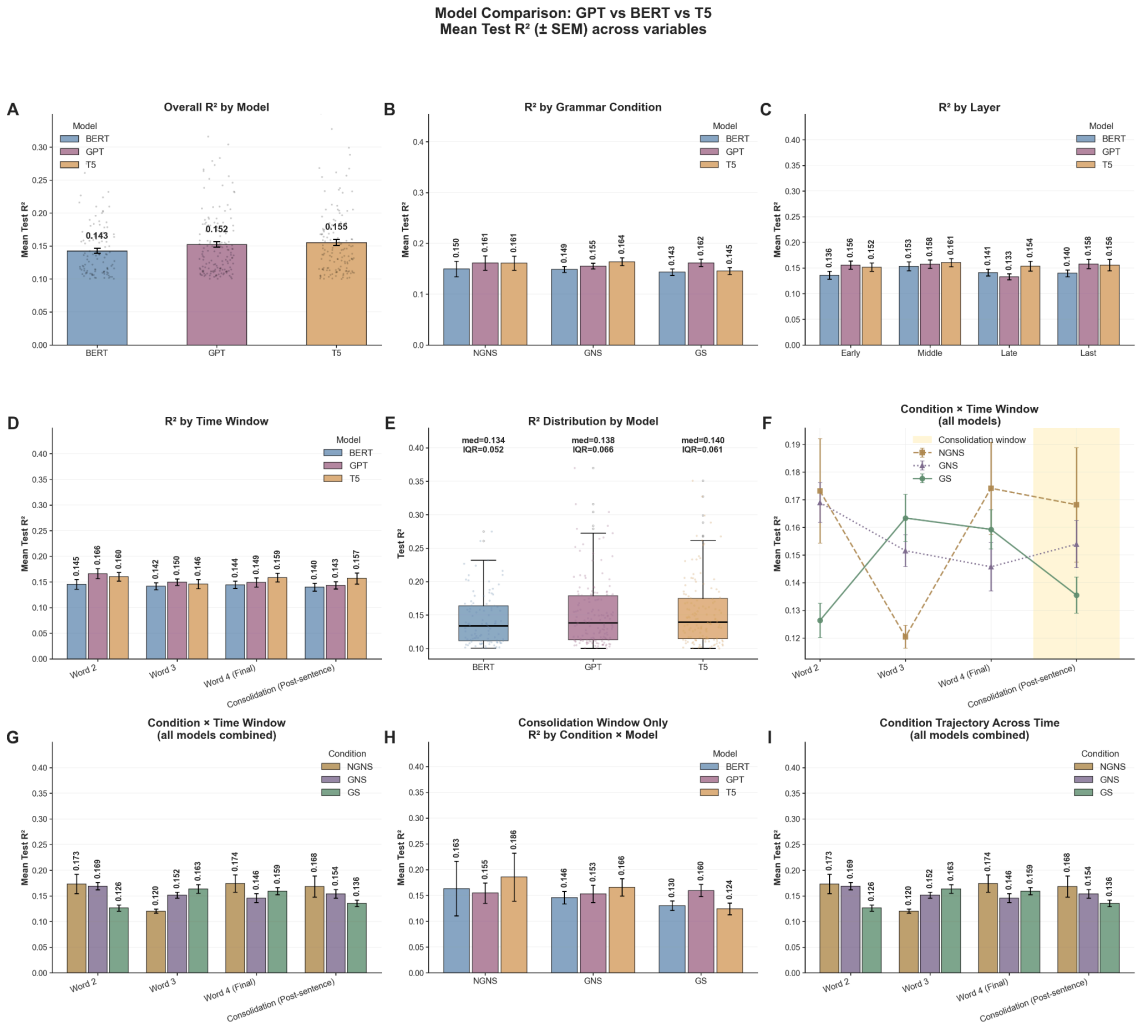


Figure 8. The T5 model has a slightly higher average explained variance than GPT and BERT within the high alignment models subset

Panel A.) Mean test r^2 per model (colored by transformer model). T5 has slightly higher average performance than BERT and GPT. Panel B) **Average model r^2 performance across grammar conditions (colored by transformer model).** Notably, GPT has slightly higher performance in the GS setting than other models C) **Model performance (r^2) split by transformer layer (colored by transformer model)**. T5 late-layer embeddings were slightly more performative than late-layer embeddings from other models. Panel D) **Model performance split r^2 , but the time window of neural data (y-vector) (colored by the transformer model).** T5 performance peaks during Word-4 and Consolidation (as does BERT). GPT performance peaks during Word 2. Panel E) **R^2 distribution per model (colored by transformer model).** The maximum explained variance (r^2 = 0.41) occurs within the GPT model, which has the highest spread of explained variances across models. **Panel F) Line plot of the interaction of the time window and grammar condition, and explained variance. Lines colored by grammar conditions.** During the consolidation time window (highlighted in yellow), models fitted with NGNS sentence embeddings have the highest performance. **Panel G) bar plot of the interaction of time window and grammar condition, and explained variance. Bars colored by grammar conditions. Panel H) Bar plot of grammar condition performance only during the consolidation window by model. Colored by the transformer model.** T5 has the highest predictivity of NGNS during consolidation, and also GNS. GPT has the highest predictivity of GS brain activity during consolidation. **Panel I) Aggregate (across models) time window x condition explained variance bar plot. Colored by grammar conditions. r^2 per model (colored by transformer model).**

Figure 8A shows that T5 exhibits slightly higher average explained variance than BERT and GPT across test sets (Although the margin is small). Panel B breaks down performance by grammar condition, revealing that GPT slightly outperforms other models in the GS condition, while T5 maintains generally higher performance across NGNS and GNS conditions. Panel C indicates that T5 late-layer embeddings are marginally more predictive than late-layer embeddings from GPT or BERT. Panel D shows temporal dynamics of model alignment, with T5 and BERT peaking during Word 4 and the Consolidation window, whereas GPT peaks earlier at Word 2. Panel E highlights that GPT has the highest maximum explained variance ($r^2 = 0.41$) and the widest spread across models. Collectively, these results suggest subtle but consistent advantages for T5 embeddings in capturing neural dynamics across sentence processing stages, with model-specific peaks depending on grammar condition and temporal window.

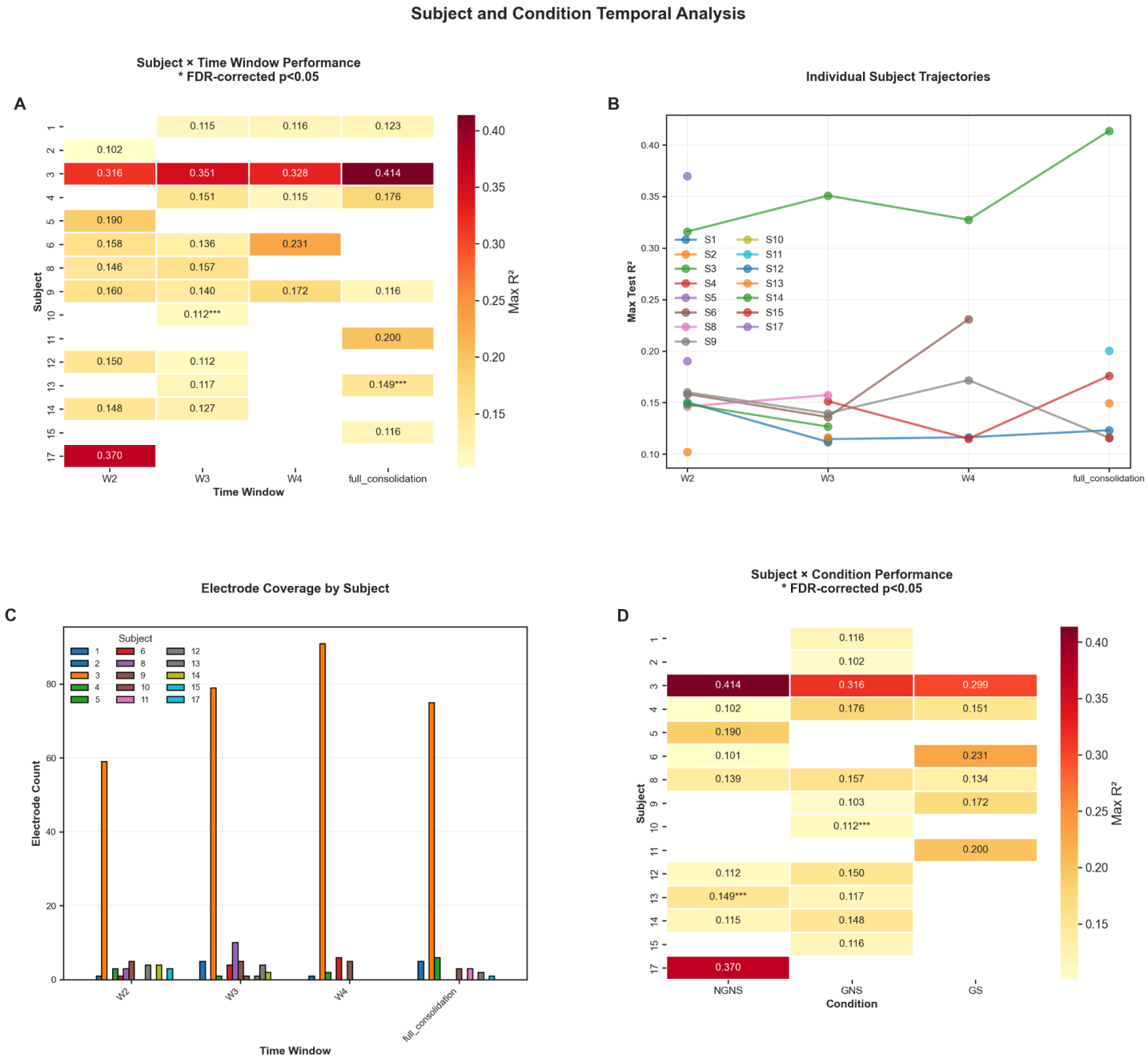


Figure 9. The highest explained variance effect sizes come from subject 3 neural data, but none are significant after 100 random shuffles and FDR Correction. Participant-level top 10 magnitude r^2 performance trends by time window and grammar condition. r^2 performance trends by time window and grammar condition. p_fdr<.001***

Panel A) Heatmap of maximum model holdout explained variance per participant per time window. The highest model performances occur for subject 3 across time windows, with consolidation having the highest $r^2 = 0.414$. However, significance occurs within models for other participants with more moderate explained variances (Ex: r^2 = 0.149 within subject 13 during consolidation). Panel B) Line plots of Subject max model r_2 **performances per subject. Note: Not all subjects had a high alignment model in every time window (hence the gaps in the heatmap) Panel C) Count of high-performing models per subject. The overwhelming majority of high performing models come from subject 3 but none of these models are significant after 100 shuffles. A moderate number of electrodes were extracted from other subjects. Panel D) Max model r^2 by grammar. The highest significant explained variance occurs within** Panel B) Line plots of Subject max

model r_2 performances per subject. Note: Not all subjects had a high-performing model in every time window Panel C) Count of high-performing models per subject. The majority of high-performing models come from subject 3 but none of these models are significant after 100 shuffles. A far more moderate number of Electrodes showed high alignment within other subjects. Panel D) Max model r^2 by grammar condition.

Figure 7 compares the high alignment models with the largest effective sizes across participants. Interestingly, the vast majority of these highly aligning significant models come from subject 3. This is a bit of a surprising result, but subject 3 was the participant with the most electrodes (See Figure 11 in the appendix for electrode counts). These electrodes were also amongst the highest signal quality and had the fewest number of sensors rejected due to data quality issues and artifacts. Importantly, the uncorrected distribution of models included a wider range of participants than the corrected distribution (See Appendix Figure 12).

10. Discussion

10.1 Main Takeaways

Overall, this study aimed to determine if alignment existed between LLM contextual embeddings and linguistically evoked SEEG gamma band signals. Per electrode (n= 1378 electrodes) regressions were fit within combinations of 4 axes of conditions (grammar condition, transformer model, transformer layer, and experimental time window). It was hypothesized that mT5 embeddings would be the most aligned with SEEG gamma power signals and that there would be evidence of hierarchical alignment in features across models. Specifically, NGNS sentence embeddings would show stronger alignment to early words in the experimental period (Ex: W) while GNS and GS sentence embeddings would align closer to later words (W3-4, consolidation). This hierarchical alignment was also expected in regard to the transformer layer, with earlier transformer layers (early and middle) correlating to NGNS evoked neurological activity, and later layers (late and last layers) aligning more with GNS and GS neurological activity. It was also expected that significant results would cluster in classical linguistic regions (Ex, IFG, ATL, etc.). After shuffle testing, FDR correction and $r^2 > 0.1$ effect size filtering the regression results, a sparse set of n = 11 significant models were found (p

< .001). Notably, FDR correction was conducted using all electrodes (even ones outside of known linguistic ROIs) as a built-in control to reduce bias and false positives in the experimental design. Hence, it is possible that the study design may have resulted in some false negatives (Which is preferable to false positives).

Despite the sparsity of the results, there was a clear pattern of results clustering in right-hemisphere homologs of classical language regions (Ex: right inferior temporal sulcus- Figure 5A) and non-classical language regions (Insular cortex- Figure 5C). Notably, a wider distribution of left hemisphere significant models were present within the uncorrected significant models, indicating that the model did find salience within left hemisphere features but that the strongest correlations were found on the right. The data overall suggests that right-hemisphere activation may increase in response to the processing of linguistic violations (NGNS and GNS conditions). Specifically, the right inferior frontal sulcus showed the strongest predictive power via early-layer embeddings (Figure 5A). While some frontal regions overfit, the inferior temporal and superior frontal gyri has more moderate test-train gaps and may be more reliable sites for brain-LLM mapping (Figure B & D). Notably, the insular cortex (typically not considered a primary language hub) aligned specifically with BERT's final layers during grammatical processing (Figure 5C). However, interestingly, generalization gaps seemed to be wider in these more “unconventional” regions compared to more classical language regions.

10.2 Non-Classical Language Regions May Activate in Response to Context Violations

Our region specificity hypothesis was only partially corroborated. It was expected that language-selective ROIs such as the inferior frontal gyrus (IFG), Wernicke’s area, anterior temporal lobe, and other frontotemporal regions would exhibit higher encoding performance than nonlinguistic regions of interest. Consistent with this prediction, robust effects were observed in classical temporal language regions, particularly within the inferior frontal sulcus (IFS), which showed the greatest concentration of significant models across grammar conditions, transformer architectures, and time windows (Figure 7C). Notably, 6 significant right hemisphere models were identified in the IFS, spanning early, middle, late, and last-layer embeddings across

models (BERT, GPT-2-XL, and T5-XL). Importantly, this ROI is linguistically salient as it serves as a key cortical locus for syntactic and semantic processing of language (W3–W4 and consolidation) (Matchin & Hickok, 2020). This convergence across architectures and representational depths suggests that there may be some degree of alignment between the representations of the IFS and transformer models.

However, there is a wide test-train performance gap within this region (approximately 65% lower on test). This is interesting, as within the train data, the model is able to explain a large amount of variance (over 50%), but within the holdout set, this explained variance reduced significantly. Interestingly, none of the significant models found within this region chose the top of the alpha range (10^6) indicating that more regularization would not be a remedy to improve overfitting (See appendix Figure 10 for the alpha distribution across model conditions). Additionally, the pipeline did not have any train-test leakage due to the usage of independent GroupShuffleSplit splitters (from the scikit-learn model selection module) to ensure clear separation and shuffling of train and test data. Given that the regularization parameter and leakage are not at fault, it's possibly a result of the intrinsic model design. Due to the limited number of sentences in the corpus compared to the larger dimensions of the neural data and transformer embeddings, the number of predictors is larger than the number of samples for almost all models. This scenario can lead to increased overfitting due to the high dimensionality of the feature space.

In an attempt to remedy this issue, PCA of the transformer embeddings was built into the pipeline as an additional tool and was tested, but actually led to lower model performance (leading to its omission from the final pipeline). This is intuitive, especially for later layers (late and last hidden output layer), contextual embeddings are already highly summarized, and cutting off dimensions can lead to information loss. Additionally, implementing an unsupervised method like PCA would decrease the interpretability of the final models by projecting results into a new coordinate space independent of the original labels. Notably, this p >> n overfitting problem while retaining model explainability is common in neuroAI studies where the number of neural samples is often outnumbered by model features (Crosse et al., 2021; Kohoutová et al., 2020; Yin et al.,

2025). Overcoming this problem while maintaining model explainability is the subject of ongoing research (Antonello et al., 2023; Hadidi et al., 2025). In terms of this dataset, another possible approach for future research could involve attempting PCA (Antonello et al., 2023) on the neural space-time bins or using a larger sentence corpus to reduce overfitting (Hong et al., 2024).

In addition to applying regularization to the current study design, a separate time-resolved model could be considered. Specifically, subsequent studies could consider a model that instead uses per-electrode dynamics to predict all brain regions and time windows at once using transformer embeddings from each model (BERT, GPT, or T5) and layer type (early, middle, late, or last). Such a model would significantly increase the amount of data the model would be able to train on (potentially improving the $p > n$ problem) and would allow the regression to more clearly learn how representations change over the entire experiment, which is not possible within the current static model. In fact, this kind of modeling approach is already feasible with the structure of the Github pipeline as is, with a few added functions.

Furthermore, region specificity was less clear-cut than hypothesized. Significant models also emerged outside canonical language areas. For instance, reliable encoding performance was detected in the Inferior Frontal Gyrus (IFG) and Inferior Temporal Sulcus (ITS), but strong effects were also observed in the insular cortex during consolidation, which was unexpected (Figures 5 & 6). However, there is a body of research linking the insula to language features such as grammar learning and auditory syntax (Ardila et al., 2014; Moro et al., 2001). Within this study, the insula was the only fdr-corrected significant region that was aligned with signals during the consolidation period from GS sentences. Taken together, these findings suggest that although transformer representations align strongly with activity in established temporal language regions, especially the STS, the mapping between model embeddings and neural signals is not strictly confined to classical frontotemporal language networks. Instead, encoding performance appears distributed across a wider frontotemporal–parietal–cingulate network(Tremblay & Dick, 2016), potentially reflecting the interaction

between linguistic computation and domain-general control systems. The pipeline itself in its unbiased approach may be a valid tool to model nonclassical latent language involved regions like the Insula.

Importantly, the majority of regions showing reliable encoding performance were right-hemispheric, which is consistent with an emerging literature on interhemispheric collaboration in language function. Wernicke's and Broca's areas have been shown to demonstrate co-activation of left and right hemisphere homologues when processing language, and right hemisphere regions are known to be recruited as linguistic cognitive load increases (Ma et al., 1996). Specifically, the right IFG and right angular gyrus have been linked to attentional control during extended speech processing (Prat et al., 2007; van Ettinger-Veenstra et al., 2010) and greater right IFG activation has been associated with processing novel or demanding linguistic input (Qi et al., 2019). This has led to the hypothesis that in the presence of increased linguistic cognitive load (such as that induced by grammar or syntax violations, surprisal or other contextual anomalies central to the present study design) homologous language areas in the right hemisphere co-activate to aid processing. The left hemisphere remains more specialized for language, but this adaptive bilateral recruitment may reflect a flexible neural strategy for managing difficult linguistic input (Muller & Meyer, 2014; Nasios et al., 2019). Notably, the insula and STS both exhibited a large test-train gap (Figure 5C), which, as discussed above, may be an artifact of the algorithmic design rather than a true reflection of encoding reliability. This unbiased neuroAI alignment pipeline may therefore be a valid tool for locating not just classical language areas but also more latent subregions involved in language processing. However, it is important to mention that there were 124 more right hemisphere electrodes than left hemisphere within the full dataset, leading to the right hemisphere having slightly more statistical power (See appendix Figure 11). Future research with more robust statistical verification could better characterize the salience of these unexpected activations and clarify whether they reflect genuine linguistic computation or domain-general processes co-opted during challenging syntactic input. Regardless of this limitation, the model generalization was moderate both within the inferior temporal gyrus (ITG) and superior frontal sulcus (SFS) which (Figure 5 B & D) may make these regions more reliable results to suggest for

potential NeuroAI alignment. The ITG is part of the ventral visual stream and plays a role in semantic word meaning while the superior frontal sulcus is involved in top-down cognitive control of language and is involved in syntactic and semantic processing via connections with Broca's area (Deniz et al., 2019; Forseth et al., 2020; Hagoort, 2019; Kuzovkin et al., 2017, 2018; Matchin & Hickok, 2020; Misra et al., 2024; Moro et al., 2001).

Crucially, there were some additional limitations. Subject 7 was omitted from the analysis due to widespread artifacts within the neurological data. Figure 9 also demonstrates a somewhat concerning trend. Figure 9C shows that the majority of high performing models ($r^2 > 0.1$) come from subject 3 across time windows and conditions. Additionally, Figure 9A indicates that the highest test r^2 value of 0.414 during consolidation occurs within subject 3. However, when subject to the 100 shuffle permutation test and FDR correction, none of these models were significant. This indicates that many of these higher explained variance models occurred due to random chance. However, our highest effect significant model still comes from subject 3 with $r^2 = 0.27$ within the IFS. This indicates our spread of explained variances within the significant range is just 0.1-0.27 despite models breaking the 0.3 ceiling (and reaching up to 0.41). Hence, although the highest effect sizes within subject 3 were not all significant, they still retain some interpretable results. Importantly, the number of electrodes, signal quality, and montage used per subject varied across subjects, with subject 3 having the most un-rejected electrodes. See the appendix Figure 11 for electrode count per subject.

10.3 T5 slightly edges out other encoding models, but the between-model performance margin is small

Firstly, the hypothesis regarding model architecture was somewhat corroborated. Overall, all sets of model embeddings performed roughly equivalently with a few minor deviations. Within Figure 7A, regressions fit to GPT-2-XL embeddings actually had the highest significance rate. However, in Figure 8A, T5 slightly edges out GPT in average explained variance over all models with $r^2 > 0.1$. This is an interesting result, as

even the differences between GPT and T5's average test r^2 values across models are quite small (only an approximately 0.003 margin). BERT is also only slightly less performative than the other two models by a small margin of about 0.01 in test explained variance. Relatedly though, past research has indicated in many cases it's simply a factor of increasing embedding dimensions that leads to increased predictivity of neural signals (Hong et al., 2024). Hence, given that all these models were amongst the largest in their representative model classes, which model was used to fit embeddings may have been irrelevant.

The margin of significant model counts is a bit more compelling to examine (compared to average r^2 performance alone) as BERT has the most FDR-corrected significant models (n =6) despite having the lowest average performance of the three models. This indicates that despite the reduced explained variance, regressions fit with BERT embeddings may have had more stability over the 100 shuffles compared to the other models. Notably, all three models have different embedding dimensions (1600-GPT-2-XL, 768<-BERT-Base-Multilingual-cased, mT5-xl-2048 dimensions) and also completely different architectures (See Figure 2 for model information). So one may wonder if the encoder-only structure of BERT may have any bearing on this result. Past studies do indicate some improved neuro-alignment for encoder-only models compared to decoder-only (Antonello et al., 2023). However, given that only 100 shuffles of the target data were conducted, the relatively sparse number of FDR-corrected significant models, and the correlational nature of the experiment, it isn't possible to directly determine if architecture is responsible for this result. Future research with a more controlled study design would be required to determine the impact of transformer architecture on alignment.

10.4 Surprisal Modulates both Artificial and Neurological Networks

The hypothesis that regressions with consolidation as the target variable would have the highest performance when GS sentence embeddings were the input was not corroborated. Surprisingly, we see in Figure 9F that the highest average r^2 during the consolidation period across models was observed for models fit with

NGNS embeddings not GS. However, when we look at Figure 8G and look at this trend per model within the consolidation window, T5 regressions perform better with NGNS embeddings over all other grammar types, as do BERT regressions. However, GPT-2-XL regressions have higher performance for GS embeddings during the consolidation time window which matches the initial hypothesis. Notably past studies have already established a basis for GPT-2-XL embeddings having high alignment with human linguistic neurological semantic processing (Blank et al., 2014; Fedorenko et al., 2016; Gillioz, 2024; Goldstein et al., 2022; Hong et al., 2024; Pereira et al., 2018; Schrimpf et al., 2021; Tuckute et al., 2024).

It is possible that the stand alone autoregressive decoder architecture of GPT-2-XL may be capturing more of the underlying information in the neurological activity related to grammar during consolidation than BERT or T5. The attention scheme of predicting the next word may implicitly include more information about gradually developing grammatical features (like the increasingly building semantic information as the experiment progresses from W1-W4 for GS sentences) than a bidirectional encoder that already has access to all adjacent features. Neurologically there is evidence in auditory language processing that the brain leverages next word prediction heuristics to speed up computation (Forseth et al., 2020). Additionally, some fMRI and MEG studies show that the human brain constantly predicts upcoming words to process language efficiently. Overall, brain activity in language centers strongly correlates with the degree of "surprisal”: defined as the difficulty of predicting the next word and hence how improbable the next word in the sequence is (Caucheteux et al., 2023; Forseth et al., 2020; Ryskin & Nieuwland, 2023). Studies that compare the human brain to autoregressive language models have found high alignment between decoder-only models and linguistic fMRI and MEG activity (Schrimpf et al., 2021). Importantly this suggests potential alignment regarding the impact of surprisal on both biological and neurological networks. Sentences with no grammar or syntax (NGNS) appear to elicit surprisal in both artificial and natural intelligence systems.

Mechanistically, when a human reads a nonsensical, scrambled sentence (e.g., "Apple the walk purple fast"), the brain's language processing centers often work harder to make sense of the noise. This can often

trigger stronger activation in the prefrontal cortex than simple, fluent sentences (Blank et al., 2014; Schrimpf et al., 2021; Shain et al., 2020). EEG imaging shows this struggle through distinct ERP effects. Non-syntactic words elicit an N400 (negative deflection at approximately 400ms) indicating semantic violation, while incorrect word order triggers a P600 (positive deflection at approximately 600ms), representing the brain's attempt to parse nonexistent syntax (Seyednozadi et al., 2021). Past studies have even indicated that there is some limited ability for the surprisal of large language models to be leveraged to predict this N400 deflection in human EEG data (Huber et al., 2024), indicating the potential alignment between biological and artificial systems in states of surprisal

Similarly, transformers are also deeply affected by surprisal. TNNs process language by calculating the probability of a word given the context. This means that scrambled, non-grammatical inputs like NGNS lead to high surprisal. The model interprets this out-of-place word as having a high negative log-probability. Consequently, the model's loss function spikes since its internal, training-based expectations are violated. This can lead to a failure in predicting the next or surrounding words accurately. This can result in scattered attention maps and high-entropy (low-confidence) output. While bidirectional transformers can analyze the whole sentence at once, true nonsense like NGNS also causes their ability to generate a coherent embedding to break down as also occurs in auto-regressive networks. This means even transformers like BERT or T5, which have bi-directional encoder blocks, would not be able to make sense of an NGNS sentence. This effect as a whole can even lead to increased attention weights since, when a transformer cannot find a logical relationship, the attention mechanism might produce unusual, high-weight patterns (Lin & Schuler, 2026; Ryu & Lewis, 2021).

These effects taken together may explain why many high-performing and significant models contain NGNS embeddings (Figure 7C shows that the majority of fdr-corrected significant models include NGNS regression features). In these regression states, the sliding window gamma power features (y variable of the regression) would potentially contain higher gamma power magnitudes in response to the non-standard input. Then, the transformer model embedding weights (x variable) would also show a similar pattern of high

magnitude and variance attention weights. Given that both the x and y variables are moving in the same direction in similar patterns, it could lead to the higher explained variances being seen in the NGNS setting compared to other grammar conditions. This is an especially interesting result as it allows for connections to be drawn between artificial and biological language predictive processing schemes. It suggests that internal next-word prediction schemes are not just isolated to transformers, but also possibly within the human brain.

10.5 Grammatical Information is likely present within Neural and Artificial Activations by W3

It was expected that there would be increased encoding performance for GS sentences during the consolidation window. However, this was not supported. It is possible that the transformer models were more likely to be correlated with grammatical neurological activity earlier in the stimulus window. For instance, in english GS sentences (majority of the subjects were english speaking except subjects 1,2, 6 & 7 see Figure 3 for demographic information) by the third word of the sentence (when the verb is introduced, allowing for a connection to be made between the subject and predicate of the sentence) grammatical content and activity begin to be introduced into the neural signals (this structure would also be present within the attention weights of the models). This would also be present within the confidence (log likelihood) of the transformer's next word prediction schemes.

Building on this, in Figure 9G we see that globally the highest average performance occurred for GS when fitted with W3 embeddings, not consolidation. However, when looking at Figure 8H, which only shows the consolidation window and model-specific trends, we see that this hypothesis was corroborated for GPT as a model but not BERT or T5. Although the hypothesis was not corroborated globally, one can conclude that word 3 still potentially represents the period in the experiment at which the semantic meaning of the sentence can be constructed for GS sentences. Hence, this correlation does not rule out the potential for grammatical alignment between the models and the human brain.

Relatedly to the discussion of the impact of language spoken, even in the uncorrected significant models, only 1 non-english participant (Participant 6- Mandarin Chinese) was ever represented and in a much lower frequency than other English participants. This suggests that although multilingual models were used (T5 and BERT) they fundamentally may have still had contextual embeddings more suited for English sentences. Additionally, there was heterogeneity in the electrode montages, recording quality, and number of electrodes that may have led to the non-english participants being downpowered statistically. Participant, 2 for instance, a Hindi-speaking participant had notably fewer electrodes than every other participant (Figure 11 appendix). Additionally, Subject 7 (Mandarin Chinese) was rejected due to data quality issues leaving only n = 3 non-english speaking participants. Future research should include more non-english speakers in their datasets in order to make more universal claims about neurolinguistic processing.

It was also expected that models fit with GNS and NGNS layer embeddings would be more correlated to activity during early windows other than consolidation like the W2-W3 period. This hypothesis was not corroborated for NGNS (as detailed above) but was consistent for GNS, as Figure 9G shows that regressions fit with GNS embeddings had the highest performance during Word 2.

Regarding our layer hypothesis, it was expected that regressions with GS embeddings would perform better when those embeddings were extracted from late or last hidden output layers. The data, however, indicated a more nuanced picture that aligns with findings from the transformer interpretability literature. Figure 8C shows that T5 late-layer embeddings are slightly more predictive than late layers from other models, supporting a modest late-layer advantage for some architectures. However, Figures 7D, G, and H reveal that the majority of FDR-corrected significant models are associated with early-layer embeddings, particularly for NGNS sentence embeddings, while middle layers dominate uncorrected significant models.

This pattern is interpretable in light of probing studies showing that early transformer layers preferentially encode surface-level syntactic and morphological features, while middle-to-late layers encode increasingly abstract semantic representations (Jawahar et al., 2019; Tenney et al., 2019). The dominance of

early-layer embeddings in FDR-corrected models, particularly under violation conditions, suggests that neural responses to syntactically and semantically anomalous sentences may be more strongly driven by surface-level lexical features than by the deep compositional representations encoded in late layers. This is consistent with the interpretation that violation processing recruits more form-level analysis, which early transformer layers happen to capture well. It also presents a potential connection to human processing schemes, which can also recognize syntactic violations sooner in the processing hierarchy than semantic (Seyednozadi et al., 2021).

The middle-layer dominance observed for GNS conditions further supports this interpretability framework. (Jawahar et al., 2019) demonstrated that intermediate BERT layers best capture hierarchical syntactic structure. GNS sentences (grammatical sentences without semantic content) may specifically recruit the structural and compositional processing that middle layers encode. This contrasts with NGNS, where early layers dominate, suggesting a gradient from surface-level to structural processing across conditions that maps onto the known representational hierarchy of transformer layers.

Temporal analyses (Figures 8D, 8F–H) additionally indicate that alignment peaks during Word 4 and the consolidation window for most models. The convergence of late time windows with early-layer dominance is theoretically interesting: it suggests that neural consolidation at the end of a four-word sentence may rely more on stable, surface-level transformer representations than on the highly contextualized compositional representations encoded in late layers. This diverges somewhat from prior fMRI encoding work, where intermediate layers tended to peak for brain alignment in naturalistic paradigms (Mischler et al., 2024; Schrimpf et al., 2021), and may reflect a paradigm-specific effect. Perhaps the violation-laden sentences used here may not sufficiently engage the long-range contextual integration that late transformer layers are optimized to represent.

Overall, these results suggest that the predictive power of transformer embeddings for neural signals may not be uniformly concentrated in the last layer. Early-layer representations can be critical for capturing neural dynamics, particularly under conditions of linguistic violation. This layer-brain alignment relationship is

modulated by both grammatical condition and time window in ways that are consistent with (and help extend) pre-existing literature.

11. Conclusion

This study investigated the representational alignment between transformer-based large language models and high–spatiotemporal resolution SEEG recordings during a four-word sentence comprehension task. Methodologically, per-electrode ridge regression models were fit between contextual embeddings and sliding-window gamma-band neural activity across electrodes, time windows, and grammatical conditions. The analysis aimed to test whether hierarchical features learned by transformer architectures correspond to neural representations of language processing. Although the number of statistically robust models was sparse after shuffle testing and FDR correction, several consistent patterns emerged that provide preliminary evidence for meaningful correspondences between artificial language representations and human neural activity.

First, the results suggest that contextual embeddings from modern transformer architectures can capture aspects of neural signals associated with sentence comprehension. Significant encoding performance was observed across multiple transformer models (BERT-base-multilingual-cased, mT5-XL, GPT-2-XL) with broadly comparable performance across architectures. While GPT-2-XL produced the greatest number of significant models, mT5 demonstrated slightly higher average explained variance among high-performing models. These results indicate that large contextual embedding spaces, rather than a specific transformer architecture, may primarily drive the ability to predict neural activity. At the same time, the presence of significant models across architectures suggests the possibility that transformer representations capture linguistic features that overlap with computations occurring in the human brain.

Second, the neuroanatomical distribution of encoding performance revealed that alignment was not restricted to canonical language areas. While effects were observed in classical temporal and frontal language regions, several reliable models emerged in right-hemisphere homologues and in regions not traditionally

considered primary language hubs such as the insular cortex. This pattern may reflect the recruitment of domain-general cognitive control systems when participants process linguistically anomalous or highly surprising sentences. The predominance of right-hemisphere significant models is also consistent with research suggesting that bilateral networks support language processing under conditions of increased cognitive load or contextual violation (Lindell, 2006; Moro et al., 2001; Muller & Meyer, 2014; Qi et al., 2019; van Ettinger-Veenstra et al., 2010). Importantly, these findings highlight the potential of unbiased NeuroAI alignment pipelines to identify both classical and latent cortical contributors to language processing.

Third, and most importantly the results provide evidence that prediction and surprisal may represent a shared computational principle between artificial and biological language systems. Models trained on nonsensical or syntactically violated sentences frequently produced higher encoding performance, particularly in NGNS conditions. Such inputs likely generate elevated prediction error within both transformer models and human neural circuits. In humans, violations of expected semantic or syntactic structure elicit well-known electrophysiological responses associated with prediction error during language processing (Huber et al., 2024; Seyednozadi et al., 2021). Similarly, transformer models experience high surprisal when encountering low-probability sequences, which can produce large shifts in internal attention patterns and embedding representations (Lin & Schuler, 2026; Ryu & Lewis, 2021). The convergence of these responses suggests that predictive processing may represent a key axis along which artificial and biological language systems align.

Beyond theoretical implications, these findings also point toward several important practical applications. Establishing measurable alignment between neural language processing and transformer embeddings can contribute to improved models of human language cognition. Such models may help clarify how grammatical structure, semantic prediction, and contextual integration are represented in the brain, offering a richer computational framework for studying language. These insights may also have translational relevance for clinical research on language disorders such as Dyslexia, where atypical neural processing of phonological and syntactic information disrupts reading and comprehension (Alkhurayyif & Sait, 2024). If transformer

representations can reliably predict neural responses during language tasks, they may eventually help identify and model neural markers associated with atypical language processing.

Additionally, improved understanding of neural–model alignment may support advances in Brain–Computer Interfaces. High-resolution neural signals combined with language model embeddings could enable more accurate decoding of linguistic intent from neural activity. This could potentially improve assistive communication technologies for individuals with speech or motor impairments. Encoding and decoding models that leverage transformer representations may therefore form an important component of future neural language prosthetics (Tang et al., 2023).

Despite these promising directions, several limitations remain. The dataset contained a relatively small number of sentences relative to the dimensionality of both neural signals and embedding features, which likely contributed to overfitting in several models. Additionally, electrode distribution varied across subjects, and a subset of results was driven by data from a single participant with higher electrode coverage. Future studies using larger training datasets, more balanced electrode distributions, and multivariate models that jointly predict activity across electrodes and time windows will be necessary to confirm the robustness of these findings.

In summary, this work contributes to a growing body of NeuroAI research suggesting that transformer-based language models exhibit measurable alignment with human neural activity during language comprehension. The results indicate that contextual embeddings may capture some aspects of linguistic computation that overlap with neural processes in the human brain. Continued investigation of these correspondences may help us gain a greater understanding of both forms of intelligence: Natural and Artificial.

12. Works Cited

Alkhurayyif, Y., & Sait, A. R. W. (2024). *Deep learning-driven dyslexia detection model using multi-modality data*. https://peerj.com/articles/cs-2077

Alleman, M., Mamou, J., A Del Rio, M., Tang, H., Kim, Y., & Chung, S. (2021). Syntactic Perturbations Reveal Representational Correlates of Hierarchical Phrase Structure in Pretrained Language Models. In A. Rogers, I. Calixto, I. Vulić, N. Saphra, N. Kassner, O.-M. Camburu, T. Bansal, & V. Shwartz (Eds.), *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)* (pp. 263–276). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.repl4nlp-1.27

Al-Mahasneh, A. J., Anavatti, S. G., & Garratt, M. A. (2017). The development of neural networks applications from perceptron to deep learning. *2017 International Conference on Advanced Mechatronics, Intelligent Manufacture, and Industrial Automation (ICAMIMIA)*, 1–6. https://doi.org/10.1109/ICAMIMIA.2017.8387619

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. https://doi.org/10.1186/s40537-021-00444-8

Antolik, J., & Bednar, J. A. (2011). *Frontiers | Development of Maps of Simple and Complex Cells in the Primary Visual Cortex*. https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2011.00017/full

Antonello, R. J., Vaidya, A. R., & Huth, A. G. (2023). *Scaling laws for language encoding models in fMRI*.

Ardila, A., Bernal, B., & Rosselli, M. (2014). Participation of the insula in language revisited: A meta-analytic connectivity study. *Journal of Neurolinguistics*, *29*, 31–41. https://doi.org/10.1016/j.jneuroling.2014.02.001

Bernabei, J. M., Arnold, T. C., Shah, P., Revell, A., Ong, I. Z., Kini, L. G., Stein, J. M., Shinohara, R. T., Lucas, T. H., Davis, K. A., Bassett, D. S., & Litt, B. (2021). Electrocorticography and stereo EEG provide distinct measures of brain connectivity: Implications for network models. *Brain Communications*, *3*(3), fcab156. https://doi.org/10.1093/braincomms/fcab156

Blank, I., Kanwisher, N., & Fedorenko, E. (2014). A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/24872535/

Buzsáki, G., & Wang, X.-J. (2012). Mechanisms of gamma oscillations. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/22443509/

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *2019-04-23*. https://doi.org/10.1371/journal.pcbi.1006897

Caucheteux, C., Gramfort, A., & King, J.-R. (2021). Disentangling syntax and semantics in the brain with deep networks. *Proceedings of the 38th International Conference on Machine Learning*, 1336–1348. https://proceedings.mlr.press/v139/caucheteux21a.html

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

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

Chen, L., Goucha, T., Männel, C., Friederici, A. D., & Zaccarella, E. (2021). Hierarchical syntactic processing is beyond mere associating: Functional magnetic resonance imaging evidence from a novel artificial grammar. *Human Brain Mapping*, *42*(10), 3253–3268. https://doi.org/10.1002/hbm.25432

Crosse, M. J., Zuk, N. J., Di Liberto, G. M., Nidiffer, A. R., Molholm, S., & Lalor, E. C. (2021). *Frontiers | Linear Modeling of Neurophysiological Responses to Speech and Other Continuous Stimuli: Methodological Considerations for Applied Research*. https://doi.org/10.3389/fnins.2021.705621

Damasio, A. R., & Tranel, D. (1993). Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences*, *90*(11), 4957–4960. https://doi.org/10.1073/pnas.90.11.4957

Daoudal, G., & Debanne, D. (2003). Long-Term Plasticity of Intrinsic Excitability: Learning Rules and Mechanisms. *Learning & Memory*, *10*(6), 456–465. https://doi.org/10.1101/lm.64103

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). The Representation of Semantic Information Across Human Cerebral Cortex During Listening Versus Reading Is Invariant to Stimulus Modality. *The Journal of Neuroscience*, *39*(39), 7722–7736. https://doi.org/10.1523/JNEUROSCI.0675-19.2019

Deuschle, W. J. (2019). *Undergraduate Fundamentals of Machine Learning*. https://dash.harvard.edu/entities/publication/aaeb98ec-2015-4c48-92e2-8746c8b52c9b

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, *113*(41), E6256–E6262. https://doi.org/10.1073/pnas.1612132113

Forseth, K. J., Hickok, G., Rollo, P. S., & Tandon, N. (2020). Language prediction mechanisms in human auditory cortex. *Nature Communications*, *11*(1), 5240. https://doi.org/10.1038/s41467-020-19010-6

Fridriksson, J., Fillmore, P., Guo, D., & Rorden, C. (2015). *Chronic Broca's Aphasia Is Caused by Damage to Broca's and Wernicke's Areas*. https://dx.doi.org/10.1093/cercor/bhu152

Friederici, A. D. (2011). The Brain Basis of Language Processing: From Structure to Function. *Physiological Reviews*, *91*(4), 1357–1392. https://doi.org/10.1152/physrev.00006.2011

Friederici, A. D., & Meyer, M. (2004). The brain knows the difference: Two types of grammatical violations. *Brain Research, Brain Research Volume 1000*, *1000*(1), 72–77. https://doi.org/10.1016/j.brainres.2003.10.057

Gardazi, N. M., Daud, A., Malik, M. K., Bukhari, A., Alsahfi, T., & Alshemaimri, B. (2025). BERT applications in natural language processing: A review. *Artificial Intelligence Review*, *58*(6), 166. https://doi.org/10.1007/s10462-025-11162-5

Gebicke-Haerter, P. J. (2023). The computational power of the human brain. *Frontiers in Cellular Neuroscience*, *17*, 1220030. https://doi.org/10.3389/fncel.2023.1220030

Gillioz, V. (2024). *Alignment of Large Language Models and Brain Activity: Exploring Language Processing through sEEG in a Multimodal Syntactic Task*. https://klab.tch.harvard.edu/publications/PDFs/gk8174_gillioz_thesis.pdf

Goldstein, A., Grinstein-Dabush, A., Schain, M., Wang, H., Hong, Z., Aubrey, B., Nastase, S. A., Zada, Z., Ham, E., Feder, A., Gazula, H., Buchnik, E., Doyle, W., Devore, S., Dugan, P., Reichart, R., Friedman, D., Brenner, M., Hassidim, A., … Hasson, U. (2024). Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature Communications*, *15*(1), 2768. https://doi.org/10.1038/s41467-024-46631-y

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U.

(2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Hadidi, N., Feghhi, E., Song, B. H., Blank, I. A., & Kao, J. C. (2025). *Illusions of Alignment Between Large Language Models and Brains Emerge From Fragile Methods and Overlooked Confounds*. https://doi.org/10.1101/2025.03.09.642245

Hagoort, P. (2019). The neurobiology of language beyond single-word processing. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/31604301/

Herff, C., Krusienski, D. J., & Kubben, P. (2020). *Frontiers | The Potential of Stereotactic-EEG for Brain-Computer Interfaces: Current Progress and Future Directions*. https://doi.org/10.3389/fnins.2020.00123

Hiratani, N., & Fukai, T. (2014). Interplay between Short- and Long-Term Plasticity in Cell-Assembly Formation. *PLoS ONE*, *9*(7), e101535. https://doi.org/10.1371/journal.pone.0101535

Hong, Z., Wang, H., Zada, Z., Gazula, H., Turner, D., Aubrey, B., Niekerken, L., Doyle, W., Devore, S., Dugan, P., Friedman, D., Devinsky, O., Flinker, A., Hasson, U., Nastase, S. A., & Goldstein, A. (2024). Scale matters: Large language models with billions (rather than millions) of parameters better match neural representations of natural language. *bioRxiv*, 2024.06.12.598513. https://doi.org/10.1101/2024.06.12.598513

Huber, E., Sauppe, S., Isasi-Isasmendi, A., Bornkessel-Schlesewsky, I., Merlo, P., & Bickel, B. (2024). Surprisal From Language Models Can Predict ERPs in Processing Predicate-Argument Structures Only if Enriched by an Agent Preference Principle. *Neurobiology of Language*, *5*(1), 167–200. https://doi.org/10.1162/nol_a_00121

Hudson, M. R., & Jones, N. C. (2022). Deciphering the code: Identifying true gamma neural oscillations. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/35985554/

Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of Language? In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3651–3657). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1356

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, *51*(5), 675–687. https://doi.org/10.1016/j.beth.2020.05.002

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974–983. https://doi.org/10.1038/s41593-019-0392-5

Kohoutová, L., Heo, J., Cha, S., Lee, S., Moon, T., Wager, T. D., & Woo, C.-W. (2020). Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nature Protocols*, *15*(4), 1399–1435. https://doi.org/10.1038/s41596-019-0289-5

Kracht, M. (2007). *Introduction to Linguistics*.

Kuzovkin, I., Vicente, R., Petton, M., Lachaux, J.-P., Baciu, M., Kahane, P., Rheims, S., Vidal, J. R., & Aru, J. (2017). *Frequency-Resolved Correlates of Visual Object Recognition in Human Brain Revealed by Deep Convolutional Neural Networks*.

Kuzovkin, I., Vicente, R., Petton, M., Lachaux, J.-P., Baciu, M., Kahane, P., Rheims, S., Vidal, J. R., & Aru, J. (2018). Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications Biology*, *1*(1), 107. https://doi.org/10.1038/s42003-018-0110-y

Lachaux, J.-P., Axmacher, N., Mormann, F., Halgren, E., & Crone, N. (2012). High-frequency neural activity and human cognition: Past, present and possible future of intracranial EEG research. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/22750156/

Lin, Y.-C., & Schuler, W. (2026). *Surprisal from Larger Transformer-based Language Models Predicts fMRI Data More Poorly* (arXiv:2506.11338). arXiv. https://doi.org/10.48550/arXiv.2506.11338

Lindell, A. K. (2006). In Your Right Mind: Right Hemisphere Contributions to Language Processing and Production. *Neuropsychology Review*, *16*(3), 131–148. https://doi.org/10.1007/s11065-006-9011-9

Lindsay, G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031. https://doi.org/10.1162/jocn_a_01544

Ma, J., Pa, C., Ta, K., Wf, E., & Kr, T. (1996). Brain activation modulated by sentence comprehension. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/8810246/

Mahajan, Y., Freestone, M., Bansal, N., Aakur, S. N., & Karmaker, S. (2025). Revisiting Word Embeddings in the LLM Era. In K. Inui, S. Sakti, H. Wang, D. F. Wong, P. Bhattacharyya, B. Banerjee, A. Ekbal, T. Chakraborty, & D. P.

Singh (Eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* (pp. 2686–2717). The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. https://aclanthology.org/2025.ijcnlp-long.145/

Matchin, W., & Hickok, G. (2020). The Cortical Organization of Syntax. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/31670779/

Merkx, D., & Frank, S. L. (2021). Human Sentence Processing: Recurrence or Attention? In E. Chersoni, N. Hollenstein, C. Jacobs, Y. Oseki, L. Prévot, & E. Santus (Eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 12–22). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.cmcl-1.2

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications. *Information*, *15*(9), 517. https://doi.org/10.3390/info15090517

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

Misra, P. (2024). *Robust and Multimodal Signals for Language in the Brain*. Harvard University. https://search.proquest.com/openview/c90cc413728a864e7ca309608870d501/1?pq-origsite=gscholar&cbl=18750&diss=y

Misra, P., Shih, Y.-C., Yu, H.-Y., Weisholtz, D., Madsen, J. R., Sceillig, S., & Kreiman, G. (2024). *Invariant neural representation of parts of speech in the human brain*. https://doi.org/10.1101/2024.01.15.575788

Moro, A., Tettamanti, M., Perani, D., Donati, C., Cappa, S. F., & Fazio, F. (2001). Syntax and the Brain: Disentangling Grammar by Selective Anomalies. *NeuroImage*, *13*(1), 110–118. https://doi.org/10.1006/nimg.2000.0668

Muller, A. M., & Meyer, M. (2014). Language in the brain at rest: New insights from resting state data and graph theoretical analysis. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00228

Nagda, K., Mukherjee, A., Shah, M., Mulchandani, P., & Kurup, L. (2020). *Ascent of Pre-trained State-of-the-Art Language Models*. https://doi.org/10.1007/978-981-15-3242-9_26

Nasios, G., Dardiotis, E., & Messinis, L. (2019). From Broca and Wernicke to the Neuromodulation Era: Insights of Brain Language Networks for Neurorehabilitation. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/31428210/

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2025). A Comprehensive Overview of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, *16*(5), 106:1-106:72. https://doi.org/10.1145/3744746

Parvizi, J., & Kastner, S. (2018). Promises and limitations of human intracranial electroencephalography. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/29507407/

Pereda, A. E. (2014). Electrical synapses and their functional interactions with chemical synapses. *Nature Reviews. Neuroscience*, *15*(4), 250–263. https://doi.org/10.1038/nrn3708

Pereira, F., Lou, B., Pritchett B, Ritter S, Gershman, S., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/29511192/

Prat, C., Keller, T., & Just, M. A. (2007). Individual differences in sentence comprehension: A functional magnetic resonance imaging investigation of syntactic and lexical processing demands. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/17892384/

Qi, Z., Han, M., Wang, Y., de Los Angeles, C., Liu, Q., Garel, K., Chen, E. san, Whitfield-Gabrieli, S., Gabrieli, J. D. E., & Perrachione, T. K. (2019). Speech processing and plasticity in the right hemisphere predict variation in adult foreign language learning. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/30853566/

Raiaan, M. A. K., Mukta, Md. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024). A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access*, *12*, 26839–26874. https://doi.org/10.1109/ACCESS.2024.3365742

Ray, S., & Maunsell, J. H. R. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/21532743/

Reddy, A. J., & Wehbe, L. (2021). Can fMRI reveal the representation of syntactic structure in the brain? *Advances in Neural Information Processing Systems*, *34*, 9843–9856. https://proceedings.neurips.cc/paper_files/paper/2021/hash/51a472c08e21aef54ed749806e3e6490-Abstract.html

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. https://doi.org/10.1162/tacl_a_00349

Ryskin, R., & Nieuwland, M. S. (2023). Prediction during language comprehension: What is next? *Trends in Cognitive Sciences*, *27*(11), 1032–1052. https://doi.org/10.1016/j.tics.2023.08.003

Ryu, S. H., & Lewis, R. (2021). Accounting for Agreement Phenomena in Sentence Comprehension with Transformer Language Models: Effects of Similarity-based Interference on Surprisal and Attention. In E. Chersoni, N. Hollenstein, C. Jacobs, Y. Oseki, L. Prévot, & E. Santus (Eds.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 61–71). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.cmcl-1.6

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

Seyednozadi, Z., Pishghadam, R., & Pishghadam, M. (2021). Functional Role of the N400 and P600 in Language-Related ERP Studies with Respect to Semantic Anomalies: An Overview. *Archives of Neuropsychiatry*, *58*(3), 249–252. https://doi.org/10.29399/npa.27422

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307

Soydaner, D. (2022). Attention Mechanism in Neural Networks: Where it Comes and Where it Goes. *Neural Computing and Applications*, *34*(16), 13371–13385. https://doi.org/10.1007/s00521-022-07366-3

Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, *26*(5), 858–866. https://doi.org/10.1038/s41593-023-01304-9

Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the Classical NLP Pipeline. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4593–4601). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1452

Topal, M. O., Bas, A., & Heerden, I. van. (2021). *Exploring Transformers in Natural Language Generation: GPT, BERT, and XLNet* (arXiv:2102.08036). arXiv. https://doi.org/10.48550/arXiv.2102.08036

Tremblay, P., & Dick, A. S. (2016). Broca and Wernicke are dead, or moving past the classic model of language neurobiology. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/27584714/

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., Kay, K., & Fedorenko, E. (2024). Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, *8*(3), 544–561. https://doi.org/10.1038/s41562-023-01783-7

van Ettinger-Veenstra, H., Ragnehed, M., Hällgren, M., Karlsson, T., Landtblom, A., Lundberg, P., & Engström, M. (2010). Right-hemispheric brain activation correlates to language performance. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/19853040/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., & Cappa, S. F. (2011). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience & Biobehavioral Reviews*, *35*(3), 407–426. https://doi.org/10.1016/j.neubiorev.2010.04.007

Vitureira, N., & Goda, Y. (2013). Cell biology in neuroscience: The interplay between Hebbian and homeostatic synaptic plasticity. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/24165934/

Widrow, B., & Lehr, M. A. (1990). 30 years of adaptive neural networks: Perceptron, Madaline, and backpropagation. *Proceedings of the IEEE*, *78*(9), 1415–1442. https://doi.org/10.1109/5.58323

Wilson, H., Chen, X., Golbabaee, M., Proulx, M. J., & O'Neill, E. (2024). Feasibility of decoding visual information from EEG. *Brain-Computer Interfaces*, *11*(1–2), 33–60. https://doi.org/10.1080/2326263X.2023.2287719

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/26906502/

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., Raj, G. D., Jhaveri, R. H., Prabadevi, B., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2024). GPT (Generative Pre-Trained Transformer)—A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *IEEE Access*, *12*, 54608–54649. https://doi.org/10.1109/ACCESS.2024.3389497

Yin, C., Yu, Q., Fang, Z., Peng, C., & Li, P. (2025). Rethinking Cross-Subject Data Splitting for Brain-to-Text Decoding. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 5675–5689). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.289

Yoo, M. H., Kim, J., & Song, S. (2025). Multilingual capabilities of GPT: A study of structural ambiguity. *PLOS One*, *20*(7), e0326943. https://doi.org/10.1371/journal.pone.0326943

Youngerman, B. E., Khan, F. A., & McKhann, G. M. (2019). Stereoelectroencephalography in epilepsy, cognitive neurophysiology, and psychiatric disease: Safety, efficacy, and place in therapy. *Neuropsychiatric Disease and Treatment*, *15*, 1701–1716. https://doi.org/10.2147/NDT.S177804

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S. J., & Kumar, S. (2020). *Are Transformers universal approximators of sequence-to-sequence functions?* (arXiv:1912.10077). arXiv. https://doi.org/10.48550/arXiv.1912.10077

13. Appendix

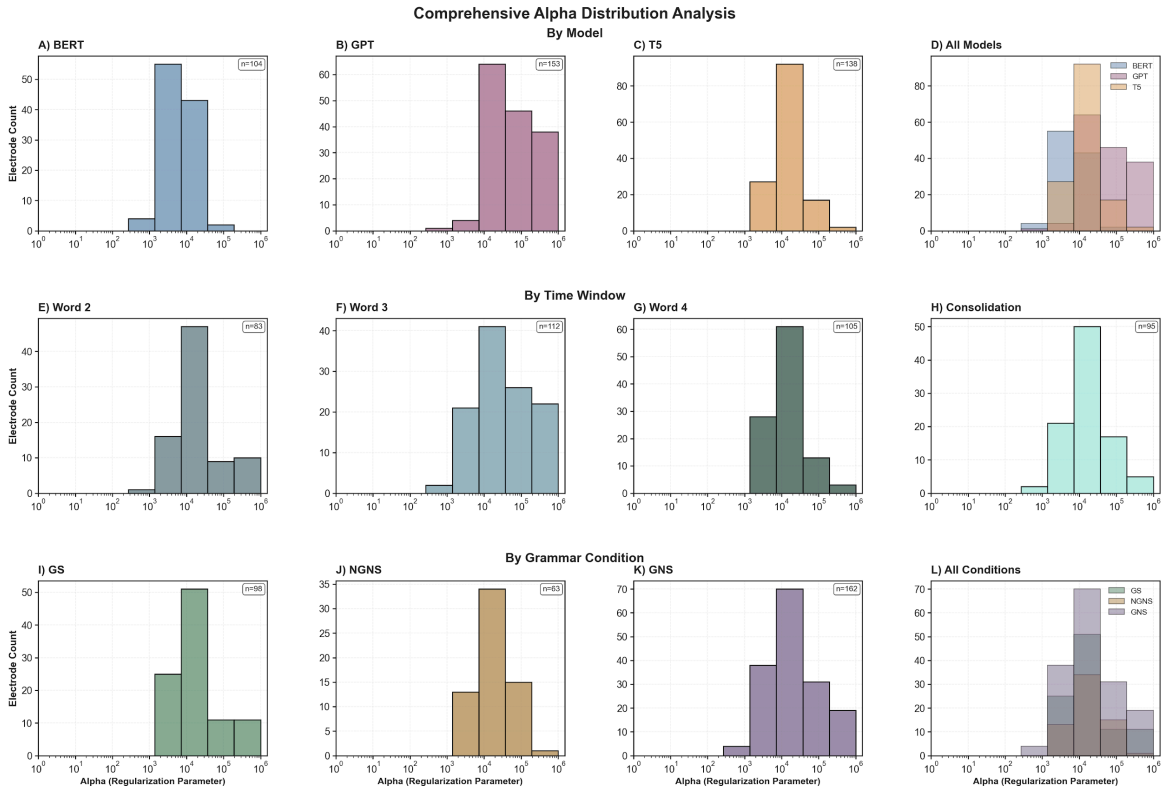


Figure 10. Model Regularization Paramter (Alpha Distribution) by Model (Row 1), Time Window (Row 2) and Grammar Condition (Row 3)

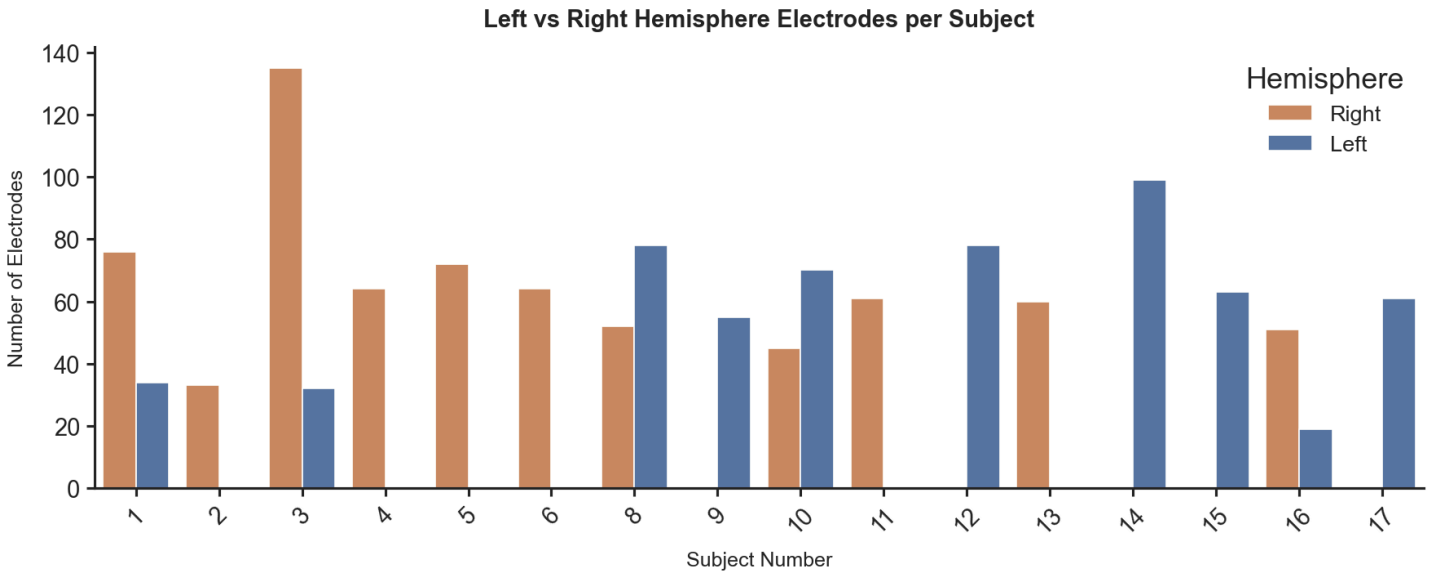


Figure 11. Left and right hemisphere electrode counts per subject (n = 589 left hemisphere electrodes and n = 713 right hemisphere electrodes)

Results Summary

Overall	
Total tests (post-FDR)	**395**
FDR-significant tests (p<0.05)	**11**
R^2 Distribution (FDR-significant)	
Mean R^2	**0.1493**
Median R^2	**0.1201**
Std R^2	**0.0543**
Min R^2	**0.1024**
Max R^2	**0.2751**
25th percentile	**0.1110**
75th percentile	**0.1745**
$R^2 > 0.1$ Breakdown	
Tests with $R^2 > 0.1$	**395**
$R^2 > 0.1$ AND uncorrected p<0.05	**66**
$R^2 > 0.1$ AND FDR significant	**11**
Uncorrected-Significant Tests by Participant	
Subject 3	**44 (66.7%)**
Subject 4	**1 (1.5%)**
Subject 5	**3 (4.5%)**
Subject 6	**1 (1.5%)**
Subject 8	**2 (3.0%)**
Subject 9	**1 (1.5%)**
Subject 10	**1 (1.5%)**
Subject 13	**6 (9.1%)**
Subject 14	**5 (7.6%)**
Subject 15	**1 (1.5%)**
Subject 17	**1 (1.5%)**
FDR-Significant Tests by Participant	
Subject 3	**9 (81.8%)**
Subject 10	**1 (9.1%)**
Subject 13	**1 (9.1%)**

Figure 12. High alignment models FDR-Corrected and Uncorrected Significant Model Summary Statistics (Participant level and General Level)

Results Summary

Overall	
Total tests	**297,112**
R^2 Distribution	
Mean R^2	**-0.0577**
Median R^2	**-0.0160**
Std R^2	**0.1233**
Min R^2	**-0.9999**
Max R^2	**0.4137**
25th percentile	**-0.0518**
75th percentile	**-0.0038**

Figure 13. Test r^2 distribution for all encoding models. Note: Due to the inclusion of all cortical regions (even non-lingustically driven ones) many models performed worse than the average (negative r^2) and therefore negligible to no alignment between embeddings and neural representations.