HIERARCHICAL FREQUENCY TAGGING PROBE (HFTP): A UNIFIED APPROACH TO INVESTIGATE SYNTACTIC STRUCTURE REPRESENTATIONS IN LARGE LANGUAGE MODELS AND THE HUMAN BRAIN

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) have shown impressive capabilities across a range of language tasks. However, questions remain about whether LLMs effectively encode linguistic structures such as phrases and sentences and how closely these representations align with those in the human brain. Here, we introduce the Hierarchical Frequency Tagging Probe (HFTP) to probe the phrase and sentence representations in LLMs and the human brain in a unified manner. HFTP utilizes frequency-domain analysis to identify which LLM computational modules (multilayer perceptron (MLP) neurons) or human cortical areas encode phrases or sentences. Human brain activity is recorded using intracranial electrodes. The results revealed distinct sensitivities to sentences and phrases across various layers of LLMs (including GPT-2, Gemma, Llama 2, Llama 3.1, and GLM-4) and across different regions of the human brain. Notably, while LLMs tend to process sentences and phrases within similar layers, the human brain engages distinct regions to process these two syntactic levels. Additionally, representational similarity analysis (RSA) shows that the syntactic representations of all five LLMs are more aligned with neural representations in the left hemisphere—the dominant hemisphere for language processing. Among the LLMs, GPT-2 and Llama 2 show the greatest similarity to human brain syntactic representations, while Llama 3.1 demonstrates a weaker resemblance. Overall, our findings provide deeper insights into syntactic processing in LLMs and highlight the effectiveness of HFTP as a versatile tool for detecting syntactic structures across diverse LLM architectures and parameters, as well as in parallel analyses of human brains and LLMs, thereby bridging computational linguistics and cognitive neuroscience.

1 INTRODUCTION

Language is fundamental to human communication, thought, and cultural transmission. According to the framework proposed by Noam Chomsky, language is divided into three key components: semantics (meaning), phonology (sound), and syntax (hierarchical sentence structure) (Chomsky, 1965). Syntax is particularly crucial as it governs how words combine into meaningful sentences, enabling the recursive and generative properties unique to human language. The theory of universal grammar proposed by Noam Chomsky suggests that all human languages share innate syntactic rules (Chomsky, 1980). Building on this foundation, cognitive scientists have shown that syntactic processing is distinct from other language functions, relying on mechanisms specifically dedicated to organizing abstract grammatical structures (Matchin & Hickok, 2017; Pylkkanen & Bemis, 2011). Furthermore, as sentence complexity increases, the neural workload also intensifies, reflecting the capacity of the human brain to resolve syntactic ambiguities and manage grammatical dependencies (Pylkkanen & Bemis, 2011). As more advances are made in understanding human syntactic processing, language models in the field of artificial intelligence have been developed that aim to capture and comprehend human language.

In recent years, large language models (LLMs) have rapidly evolved, surpassing human-level capabilities in numerous tasks (Vaswani et al., 2017; Brown et al., 2020). Specially, the performance of

LLMs in natural language understanding, translation, and summarization has led to claims that they possess a remarkable degree of human-like linguistic ability, especially in generating language that adheres to the surface rules of syntax (Radford et al., 2019). However, the question of whether LLMs can process the hierarchical structures of sentences in a manner comparable to humans remains unclear. While some studies suggest that LLMs can successfully capture and manipulate hierarchical structures of sentences (Manning et al., 2020), others argue that LLMs lack the deeper syntactic processing capabilities observed in human cognition (Linzen et al., 2016; McCoy et al., 2019). This controversy stems from the absence of a unified framework to rigorously evaluate and compare the syntactic processing abilities of LLMs with those of the human brain, making it difficult to draw definitive conclusions about their true linguistic competence (Tenney et al., 2019; Warstadt et al., 2020). Therefore, it is crucial to develop a method that can simultaneously probe the hierarchical syntactic structure representations in both LLMs and the human brain, and on this basis, explore the internal similarities of syntactic representations between them.

Ding and colleagues (Ding et al., 2016) introduced the hierarchical frequency tagging (HFT) technique to uncover how the human brain processes hierarchical linguistic structures during natural speech comprehension. In this paradigm, monosyllabic words are presented at a rate of 4Hz to form phrases at 2Hz, which combine into sentences at 1Hz. Using frequency-domain analysis of electrophysiological signals, Ding and colleagues deconstruct the processing of linguistic structures such as phrases and sentences. Subsequent research further highlighted the importance of attentional mechanisms in processing these hierarchical structures (Ding et al., 2018), while other studies have shown that HFT captures distinct neural responses to different linguistic elements (Keitel et al., 2018). Moreover, Martin et al. extended this framework by exploring how the brain organizes complex linguistic stimuli (Martin & Doumas, 2017). These studies demonstrate the effectiveness of HFT in revealing language processing patterns. Given the differences in temporal processing between LLMs and the human brain, aligning their syntactic representations in the frequency domain is a feasible and necessary approach.

Building on the HFT paradigm Ding et al. (2016), here we developed the Hierarchical Frequency Tagging Probe (HFTP) to investigate whether specific computational modules within LLMs process hierarchical sentence structures. HFTP offers a unified approach to explore internal similarities and systematically examine the alignment of syntactic representations between LLMs and the human brain. The key contributions of this paper are: (1) We innovatively employed frequency-domain analysis using HFTP to characterize the syntactic representations of every computational unit in each layer of LLMs; (2) HFTP provides a simple and universally applicable approach for both LLMs and the human brain, establishing a platform for studying the alignment of syntactic structure representations between them; (3) Using syntactic templates derived from HFTP, we identified brain regions highly correlated with LLMs, predominantly located in key language-processing areas of the left hemisphere; (4) By comparing five LLMs, we found that the internal syntactic representations of GPT-2 and Llama 2 exhibit higher overall similarity to those of the human brain. In sum, HFTP effectively aligns representations between LLMs and the human brain, providing a novel platform for future alignment studies.

2 Related Work

Syntactic processing in the human brain In humans, syntactic processing involves a complex neural network that organizes words into hierarchical structures, enabling infinite expression from a finite set of elements. Research on brain damage and neuroimaging has indicated a left hemisphere advantage in language processing (Friederici & Brauer, 2009; Hagoort, 2013; Blank et al., 2016), with further evidence suggesting that this lateralization may stem from distinct sensitivity to temporal modulations crucial for speech perception (Albouy et al., 2020). However, recent evidence suggests that syntactic processing relies on a distributed network across frontal and temporal areas, with significant overlap in regions responsible for both syntactic and semantic functions (Blank et al., 2016; Fedorenko et al., 2020). These findings demonstrate that syntactic processing is not confined to isolated regions but is part of a broad, interconnected network.

Syntactic processing in language models Even before the development of LLMs, researchers found that simple LSTM language models could capture syntax-sensitive dependencies, such as subject-verb agreement (Linzen et al., 2016; Kuncoro et al., 2018). Using a technique called structural

probing, Manning and colleagues discovered that transformer-based models like BERT can encode hierarchical syntactic trees, enabling such models to implicitly represent complex syntax without direct training (Hewitt & Manning, 2019). These transformer-based models excel at tracking both local and long-range dependencies through specialized attention mechanisms, distributing syntactic knowledge across layers (Clark, 2019; Tenney et al., 2019; Manning et al., 2020). However, the methods employed in these studies of language models make it challenging to apply findings to the exploration of human brain activity.

Alignment between LLMs and the human brain Previous research on the alignment between LLMs and human brain representations has largely focused on next-word prediction without distinguishing between semantics and syntax. This approach compares the prediction probabilities of LLMs with human brain activity to assess shared computational principles (Schrimpf et al., 2021; Goldstein et al., 2022). Recent studies indicate that syntactic processing representations closely correlate with brain regions responsible for hierarchical syntax (Caucheteux & King, 2022; Oota et al., 2024). Moreover, efforts to disentangle syntax from semantics reveal distinct patterns in both neural and model representations (Caucheteux et al., 2021). While some alignment exists, variations in task complexity suggest that LLMs display more layer-specific syntactic processing (Tuckute et al., 2024). However, these studies utilize different corpora, alignment methods, and LLMs, making it difficult to conduct a systematic investigation of syntactic representations in large models. A unified tool is needed to probe syntactic structure representations across the human brain and various LLMs, enabling systematic comparisons of representation similarities at the neuronal population level.

3 METHODS

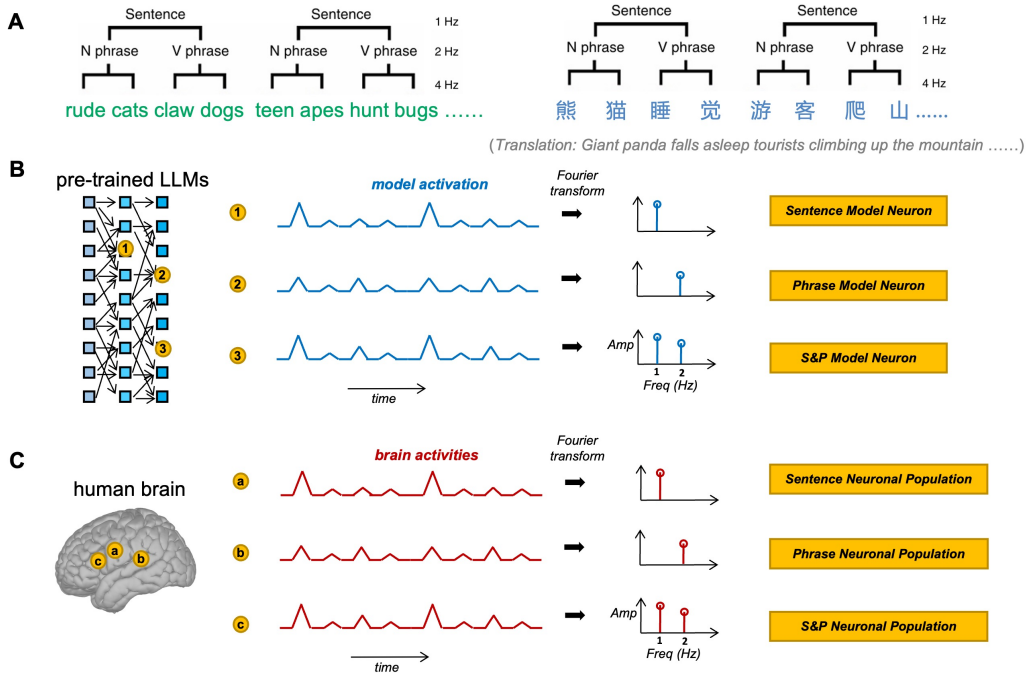


Figure 1: A framework for Hierarchical Frequency Tagging Probes (HFTP) and an illustration of neurons involved in different levels of hierarchical linguistic processing in both large LLMs and the human brain. **A**, hierarchical linguistic structure in English and Chinese including syllable, phrase, and sentence. **B**, hierarchical linguistic pattern observed both in LLMs and **C**, human brain.

We present the framework of the proposed HFTP methodology (see Figure 1). This framework is organized into four parts: Section 3.1 describes the syntactic corpora used and the LLM architectures; Section 3.2 details the application of HFTP to identify significant sentence and phrase neurons; Section 3.3 explains how the HFTP approach is applied to human intracranial stereo-electroencephalography (sEEG) data; and Section 3.4 correlates syntactic structure representations

in LLMs and the human brain by comparing frequency-domain representations and identifying similarities in how syntactic structures are encoded across both systems.

3.1 DATA AND LLMS

We utilized Chinese and English corpora adapted from (Ding et al., 2016), consisting of four-syllable sequences in Chinese or four-word sequences in English, where the first two and last two units form phrases (see Figure 1). Further details regarding the corpus can be found in Appendix A.6. For both the sEEG and model-brain alignment experiments, we used the same two Chinese corpora—the **sentence** and **phrase** corpora—from (Sheng et al., 2019). While these corpora share a similar structure to the Chinese syntactic corpus used in the LLM experiments, they differ in content. To control for semantic processing, we created a word-order randomized version for each corpus as a control condition. In this randomized version, the semantics of individual words are preserved, while the syntactic structure is disrupted.

We applied HFTP to five state-of-the-art LLMs—GPT-2, Gemma, Llama 2, Llama 3.1, and GLM-4—which vary in both architecture and parameter scale (see Table 3). As GPT-3.5 and its subsequent versions are not open-source, we were limited to using GPT-2 from the GPT series for LLM experiments. Notably, the term "MLP neuron" refers here to computational modules within the multilayer perceptron (MLP) layers of a Transformer model. These neurons are part of the feed-forward network, which follows the attention mechanism in each layer. The MLP consists of two fully connected layers, and the number of neurons corresponds to the hidden units that process the input after attention has been applied.

3.2 SYNTACTIC STRUCTURE PROBE IN LLMS

For each LLM, sequences from both the Chinese and English syntactic corpora were concatenated into a continuous text to capture neural-like activations. During this process, each Chinese syllable (or English word) outputted an activation value, allowing the signal corresponding to every individual linguistic unit to be traced. These time-domain activations were then transformed into frequency-domain information via fast-fourier transform (FFT). Due to the lack of time-course information inherent to LLM input structures, we artificially defined a time scale on which the activation values are output at a frequency of 4 Hz, and we also manually constrained the sampling rate to 4 Hz, limiting the observable frequencies to the 0-2 Hz range. This adjustment ensured that the syntactic rhythms analogous to those observed in human brain data could be captured within the model activations.

LLMs, with their multiple layers and thousands of MLP neurons per layer, require a systematic approach to identify which neurons are responsible for either sentence or phrase processing. We developed a method to identify significant syntactic processing units, applicable to both LLMs and human brain data. For the LLMs, we conducted a permutation test, randomizing the model activations derived from the structured input corpus 1000 times. The original frequency bins at 1 Hz and 2 Hz, representing sentence and phrase rhythms respectively (their real parts of amplitudes are denoted as $\mathrm{real}[\mathrm{amp}(1\mathrm{Hz})]$, $\mathrm{real}[\mathrm{amp}(2\mathrm{Hz})]$), were compared to the 95% confidence intervals (CI) generated by the distribution of permuted activations. Neurons whose $\mathrm{real}[\mathrm{amp}(1\mathrm{Hz})]$ and $\mathrm{real}[\mathrm{amp}(2\mathrm{Hz})]$ values exceeded this threshold were classified as *significant MLP neurons* (see 1), indicating their involvement in syntactic processing with statistical robustness against random noise.

Definition 1 (Significant MLP Neurons)**.** *For a fixed frequency f, a neuron is a significant MLP neuron, if and only if its FFT result satisfies*

$$\mathrm{real}[\mathrm{amp}(f)] \notin \textit{95\% CI of permuted distribution}. \tag{1}$$

The set containing all the significant MLP neurons in terms of frequency f is denoted as $\mathbb{S}_f$.

Since the *significant MLP neurons* are distributed almost uniformly across all layers, identifying the specific neurons that contribute to sentence and phrase processing requires a more objective and systematic method. To achieve this, we applied z-scores to the FFT amplitudes at 1 Hz and 2 Hz in both the experimental and control groups for all *significant MLP neurons* across layers. The z-score deviation between the experimental and control groups was then calculated for each neuron. Neurons associated with sentence processing and phrase processing were defined as those whose z-

scores deviated by more than two standard deviations from the mean, at 1 Hz and 2 Hz, respectively (see 2).

Definition 2 (Sentence MLP Neurons and Phrase MLP Neurons)**.** *A neuron n is defined as a sentence/phrase MLP neuron if it satisfies*

$$n \in \mathbb{S}_f, \quad z_f(n) \geq \mu_{z_f} + 2\sigma_{z_f}, \tag{2}$$

where $z_f(n)$ denotes the z-score of the FFT amplitude for neuron n at fruquency f, μ_{z_f} denotes the mean z-score across all neurons for the frequency f, σ_{z_f} denotes the standard deviation of z-scores across all neurons for the frequency f, and the frequency f is specified as 1Hz *and* 2Hz *for sentence and phrase MLP neuron respectively.*

Following this, we identified and analysed sentence and phrase MLP neurons across layers and LLMs, with full details provided in Section 4.1. We also conducted bilingual experiments to assess the ability of different LLMs to perceive syntactic structures across Chinese and English (see Appendix A.2).

3.3 Syntactic structure probe in the human brain

In the human sEEG experiment, we recorded sEEG signals from 26 native Chinese speakers, using two Chinese corpora: the **sentence** and **phrase** corpora. The Chinese corpus was presented to the human subjects in the auditory modality. In the **sentence** corpus, every nine four-syllable sequences were concatenated into a single trial, yielding 40 trials per subject. Similarly, in the **phrase** corpus, every 18 two-syllable phrases were concatenated, also resulting in 40 trials per subject. Each syllable had a duration of 250 ms, and sEEG signals were sampled at 512 Hz for most participants, except for one with a sampling rate of 2,048 Hz. Note that to reduce the strong neural responses at word onset during auditory presentation, we only used sEEG recordings from the final 32 syllables of each trial per subject.

To analyze the sEEG data, we employed inter-trial phase coherence (ITPC), a frequency-domain method relatively resistant to noise that quantifies the consistency of phase relationships in oscillatory brain activity across multiple trials (Cohen, 2014). SEEG Electrode localization was performed similarly to our previous studies (Xu et al., 2023; Wang et al., 2024); all electrodes were mapped to brain regions defined by the Automated Anatomical Labeling (AAL) system. We then grouped certain AAL regions to form 12 brain regions of interest (ROIs) (details in Appendix A.7). Subsequent experiments were conducted based on brain ROIs.

As previously outlined, the proposed HFTP approach is designed to be applicable to both LLMs and human brain data. For the human brain analysis, we employed the same permutation testing procedure on the time-domain sEEG data that captured cortical activity during listening to Chinese corpora. Specifically, ITPC results were randomized 1000 times for each channel in each subject. The original frequency bins, $\mathrm{real}[\mathrm{amp}(1\mathrm{Hz})]$ and $\mathrm{real}[\mathrm{amp}(2\mathrm{Hz})]$, were then assessed to determine whether they fell within the 95% confidence interval of the permuted ITPC distribution (see 3).

Definition 3. *A channel c is defined as a sentence/phrase channel if its ITPC result satisfies*

$$\mathrm{real}[\mathrm{amp}(f)] \notin \textit{95\% CI of permuted ITPC}, \tag{3}$$

where $f = 1\mathrm{Hz}$ for sentence channel and $f = 2\mathrm{Hz}$ for phrase channel.

Using this probe, we identified and analyzed the distribution of sentence and phrase channels across various brain ROIs, with full details provided in Section 4.2.

3.4 Alignment of Syntactic Structure Representations of LLMs with the Human Brain

To explore syntactic structure representation alignment between LLMs and the human brain, we compared their frequency-domain representations using the same **sentence** and **phrase** corpora. For each computational modular (a MLP neuron in LLMs or an sEEG electrode channel in the human brain), we extracted values in the frequency spectrum as a feature. This approach creates a multi-dimensional space based on frequency-domain features, where each syntactic structure corresponds to a specific point in this space (see Figure 2). We then compute the distances between

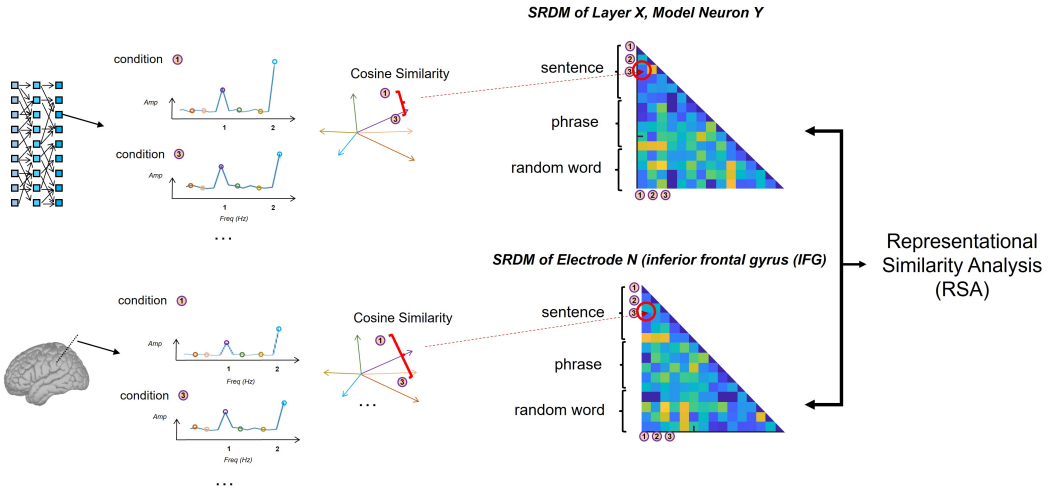


Figure 2: Alignment pipeline between LLMs and human brain. SRDMs are computed for both MLP neurons and brain channels using cosine similarity. RSA is then applied to quantify the similarity between model and brain representations.

these points for different syntactic structures within the same computational unit using cosine similarity. Through pairwise comparisons, we constructed Structure Representational Dissimilarity Matrices (SRDMs) for each computational module, which are similar to Representational Dissimilarity Matrices (RDMs) but specifically capture the representations of syntactic structures (Cichy et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014). We then applied Representational Similarity Analysis (RSA) to enable cross-modal comparisons between LLMs and brain data, correlating the representations in both systems (Kriegeskorte et al., 2008). This approach allowed us to quantify alignment and use statistical tests to identify significant overlaps. We introduced two key measures: model-brain similarity ($S_{m,b}$) and model-region similarity (S_{m,b_r}), to evaluate alignment globally and in specific brain ROI, and used the contribution ratio (CR_r) to assess the impact of each region on the alignment. For more details on the alignment pipeline, see Appendix A.1. The comprehensive discussion of the alignment results can be found in Section 4.3.

4 Experiments

In our experiments, we employed the Hierarchical Frequency Tagging Probe (HFTP) to investigate structural processing capabilities in both the human brain and LLMs, including GPT-2 (0.7B), Gemma (2B), Llama 2 (7B), Llama 3.1 (8B), and GLM-4 (9B). This unified approach allowed us to detect syntactic patterns across both systems and facilitate alignment between their representations.

4.1 MLP neurons represent sentences and phrases in LLMs

Using the HFTP method, we identified neurons in all five models that selectively represent sentences (sentence neurons), phrases (phrase neurons), and neurons that simultaneously represent both (sentence & phrase neurons). In the examples shown, we highlight MLP neurons which display distinct hierarchical frequency patterns. Figure 3 demonstrates four patterns: a significant peak at the sentence frequency (f_{sentence}), a significant peak at the phrase frequency (f_{phrase}), dual peaks at both f_{sentence} and f_{phrase}, and no significant peaks. Frequencies beyond 2 Hz have been artificially set to zero for smoothness in the representation.

For the five LLMs tested, we identified their sentence and phrase neurons using the HFTP method. Figure 4 shows the distribution of exclusive sentence/phrase neurons and those representing both across different layers, based on experiments using the Chinese syntactic corpus. All five models contain neurons dedicated to capturing sentences and phrases, demonstrating their ability to encode the syntactic hierarchies of human language. However, distinct distribution patterns suggest varied syntactic processing strategies: Llama and GLM primarily process syntactic information in later

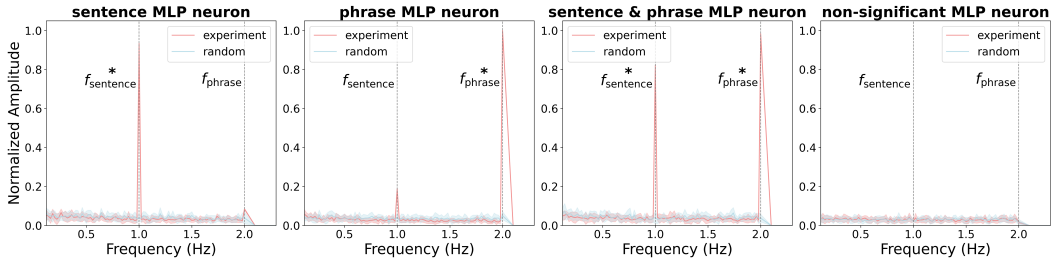
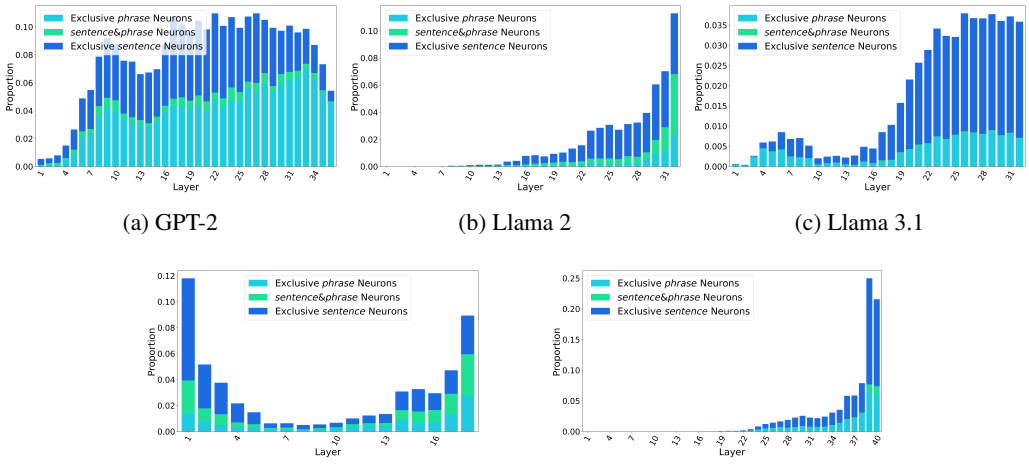


Figure 3: Hierarchical frequency patterns of MLP neurons selectively represent sentence features, phrase features, and shared features of both. Here, "experiment" denotes the original corpus, while "random" indicates the randomized version. Significant frequency peaks are marked ($*p < 0.05$).

layers, indicating a more integrated approach, while GPT has higher concentrations of sentence and phrase neurons in its middle layers. In contrast, Gemma presents a two-step process, with dense concentrations in both early and late layers.



(a) GPT-2 (b) Llama 2 (c) Llama 3.1

(d) Gemma (e) GLM-4

Figure 4: Statistics of exclusive sentence/phrase MLP neurons and sentence & phrase MLP neurons in each layer across five LLMs

A comparative analysis shows a notable decrease in the maximum proportions of sentence neurons (from 8.9% in Llama 2 to 3.0% in Llama 3.1) and phrase neurons (from 6.8% to 0.9%). Since Llama 3.1 is an updated version of Llama 2, this suggests a potential shift in computational resources. To improve performance on complex tasks—such as reasoning and higher cognitive functions—Llama 3.1 may reduce the specialized processing of syntactic structures (sentences and phrases), reallocating neurons to these advanced cognitive functions.

Additionally, a consistent covariant trend between sentence and phrase neurons across layers was observed for all five models, with high statistical correlations, including Gemma (r = 0.841), GPT-2 (r = 0.585), GLM (r = 0.993), Llama 2 (r = 0.912), and Llama 3.1 (r = 0.934). These findings suggest that LLMs share similar underlying mechanisms for sentence and phrase processing.

4.2 Sentences and Phrases Representations in the Human Brain

Using the HFTP approach, we identified neuron populations in the human brain that selectively represent sentences and phrases. Each sEEG channel captures collective responses from nearby neuron populations, providing high spatio-temporal resolution of neural activity. This allows us to assess sentence and phrase selectivity precisely. As shown in Figure 5, we found channels representing sentences and phrases, as well as channels with shared representations, while some channels did

not represent either. These findings align with those observed in LLMs, demonstrating that HFTP effectively investigates the internal representations of syntactic structures in both systems.

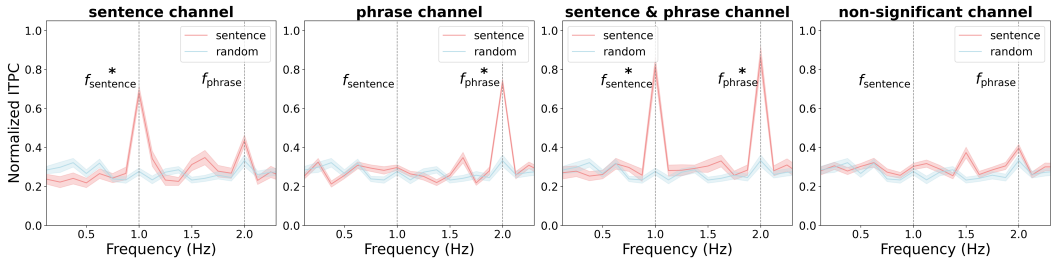
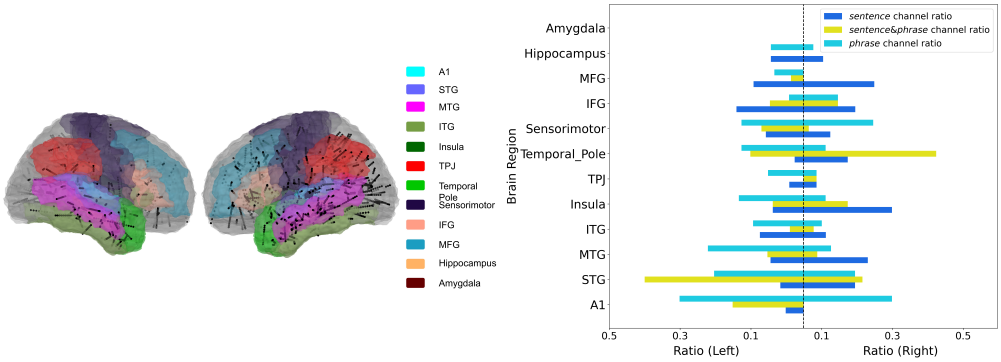


Figure 5: Hierarchical frequency patterns of MLP neurons selectively represent sentence features, phrase features, and shared features of both. Here, "experiment" denotes the original corpus, while "random" indicates the randomized version. Significant frequency peaks are marked ($*p < 0.05$).

Similar to our analysis of neuron types in LLM layers, we calculated the proportions of sentence and phrase channels within each brain ROI. As shown in Figure 6, phrase channels decrease from lower layers (A1) to higher layers (e.g., IFG), while sentence channels show the opposite trend, increasing at higher brain layers. This pattern aligns with earlier MEG studies Sheng et al. (2019), supporting distinct processing mechanisms for sentences and phrases. Correlations between sentence and phrase channels across brain ROIs in both hemispheres revealed no significant relationship (left: r = -0.1685, p = 0.606; right: r = -0.197, p = 0.539), suggesting that sentence and phrase processing operate independently. This contrasts with the behavior of LLMs, implying that while the human brain segregates sentence and phrase processing across different regions, LLMs integrate both syntactic levels within the same model layers. This difference further demonstrates that the layered representations of LLMs may not directly align with the differentiated processing roles observed in distinct brain ROIs.



(a) sEEG channel locations and Brain ROIs (b) Significant sEEG channel distribution

Figure 6: (a) Brain ROIs of the left and right hemispheres used in this study. The black electrodes represent the sEEG channel locations across all participants. (b) Distribution of significant exclusive sentence/phrase and sentence & phrase channels (**sentence** corpus) in different brain ROIs.

4.3 Alignment of Syntactic Structure Representations between LLMs and the Human Brain

Since the layered representations of LLMs do not align directly with the specific processing functions of various brain ROIs, we sought to investigate whether overall syntactic representations in LLMs are comparable to those in the human brain, both globally and across individual brain ROIs. To accomplish this, we employed Searchlight representational alignment, taking syntactic representations from LLMs as reference points. Specifically, we extracted sentence MLP neuron representations and correlated them with the SRDMs of each sEEG channel, identifying the top 100 most correlated channels to calculate average correlation values. This analysis provided model-brain

similarity ($S_{m,b}$) and model-region similarity (S_{m,b_r}) across the five LLMs in different hemispheres. Correlations were computed separately for exclusive sentence/phrase neurons and sentence & phrase neurons under the **sentence** corpus.

Table 1: Averaged top 100 Spearman correlation coefficients between SRDMs of sEEG channels and those of MLP neurons under the **sentence** corpus condition, separated by left (L) and right (R) hemispheres. Note that '/' denotes cases where the model lacks channels corresponding to the brain ROI in that hemisphere within the top 100 Spearman-ranked channels. The values in the rows corresponding to brain ROIs represent the model-region similarity S_{m,b_r}.

	GPT-2		Gemma		Llama 2		Llama 3.1		GLM-4	
	L	R	L	R	L	R	L	R	L	R
$S_{m,b}$	0.646	0.437	0.590	0.422	**0.651**	0.439	0.533	0.410	0.603	0.446
A1	0.674	0.475	**0.647**	0.331	0.605	0.635	0.511	/	**0.657**	/
STG	0.647	0.422	**0.628**	0.378	**0.689**	0.431	0.494	0.341	**0.620**	0.408
MTG	0.668	0.397	**0.609**	0.352	**0.679**	0.439	0.526	0.389	0.602	0.385
ITG	0.638	0.449	0.579	0.421	0.636	0.428	0.544	0.397	0.594	0.449
Insula	0.606	0.446	0.545	0.406	0.627	0.430	0.540	0.434	0.577	0.470
TPJ	0.598	0.472	0.543	0.357	0.620	0.338	0.557	0.460	0.573	0.469
Temporal Pole	**0.698**	0.450	0.563	0.512	0.591	0.466	/	**0.588**	0.595	0.425
Sensorimotor	0.600	0.465	0.567	0.429	0.638	0.468	0.527	0.406	0.593	0.503
IFG	**0.717**	0.410	0.596	0.522	**0.656**	0.487	0.538	0.434	**0.616**	0.507
MFG	0.631	0.411	0.534	0.536	0.602	/	**0.562**	0.363	0.584	0.395
Hippocampus	**0.755**	0.391	0.542	0.430	0.572	0.366	**0.590**	0.424	0.580	0.437
Amygdala	/	0.461	/	0.594	/	0.556	/	/	/	0.481

As shown in Table 1, we observed that Llama 2 (r=0.651) and GPT-2 (r=0.646) exhibited the highest average correlations with human brain activity at the sentence level. Unexpectedly, Llama 3.1 (r=0.533), an updated version of Llama 2, showed lower alignment with the human brain. Although Llama 3.1 has employed multiple techniques—such as task balancing and post-processing iterative alignment—to enhance its overall performance (Dubey et al., 2024), including its language capabilities, it remains unclear whether these improvements result in a closer similarity between the internal model representations and those of the human brain. The limitations of scaling up LLMs in semantic role understanding, as demonstrated by their struggles with complex arguments and nuanced differences in semantic roles (Cheng et al., 2024), suggest that the computational principles underlying language processing in Llama 3.1 may have gradually diverged from those of the human brain. This divergence potentially leads LLMs to develop their own unique patterns for language processing representations, differing from human semantic processing.

Additionally, the top brain ROI for each model—primarily in the left hemisphere—highlighted regions critical for syntactic processing, such as the left STG, MTG, and IFG, with Llama 2 showing particularly strong correlations in these areas. We found that other LLMs also exhibited relatively high S_{m,b_r} in these three brain ROIs. Similarly, phrase-level experiments (see Table 2) yielded comparable findings, reinforcing the robustness of the HFTP approach and suggesting its potential as a valuable tool for future studies of model-brain alignment.

5 Conclusion

In conclusion, this study advances our understanding of syntactic processing in both LLMs and the human brain. By introducing the Hierarchical Frequency Tagging Probe (HFTP), we provide a unified methodology for analyzing hierarchical syntax and exploring representational similarities between artificial and biological systems. The findings demonstrate that while LLMs capture some aspects of human syntactic processing, their underlying mechanisms diverge notably from those of the human brain. This highlights the need for further refinement of LLMs to better emulate human-like cognitive processes. This research bridges the gap between computational linguistics and cognitive neuroscience, paving the way for future interdisciplinary studies that can enhance both artificial intelligence and our understanding of human cognition.

REFERENCES

Philippe Albouy, Lucas Benjamin, Benjamin Morillon, and Robert J Zatorre. Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science*, 367 (6481):1043–1047, 2020.

Idan Blank, Zuzanna Z Balewski, Kyle Mahowald, and Evelina Fedorenko. Syntactic processing is distributed across the language system. *NeuroImage*, 127:307–323, 2016.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134, 2022.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Disentangling syntax and semantics in the brain with deep networks. In *International conference on machine learning*, pp. 1336–1348. PMLR, 2021.

Ning Cheng, Zhaohui Yan, Ziming Wang, Zhijie Li, Jiaming Yu, Zilong Zheng, Kewei Tu, Jinan Xu, and Wenjuan Han. Potential and limitations of llms in capturing structured semantics: A case study on srl. In *International Conference on Intelligent Computing*, pp. 50–61. Springer, 2024.

Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, 1965.

Noam Chomsky. *Rules and Representations*. Columbia University Press, 1980.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *arXiv preprint arXiv:1406.3284*, 2014.

Kevin Clark. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

Michael X. Cohen. *Analyzing Neural Time Series Data: Theory and Practice*. MIT Press, 2014.

Nai Ding, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, 19(1):158–164, January 2016.

Nai Ding, Xunyi Pan, Cheng Luo, Naifei Su, Wen Zhang, and Jianfeng Zhang. Attention is required for knowledge-based sequential grouping: insights from the integration of syllables into words. *Journal of Neuroscience*, 38(5):1178–1188, 2018.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Evelina Fedorenko, Idan Asher Blank, Matthew Siegelman, and Zachary Mineroff. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, 203: 104348, 2020.

Angela D Friederici and Jens Brauer. Syntactic complexity in the brain. In *Functional neuroimaging of syntactic processing*, pp. 491–506. John Benjamins Publishing, 2009.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

Peter Hagoort. Muc (memory, unification, control) and beyond. *Frontiers in Psychology*, 4:416, 2013.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Anne Keitel, Joachim Gross, and Christoph Kayser. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS biology*, 16(3):e2004473, 2018.

Seyed Mohammad Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11): e1003915, 2014.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.

Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. Lstms can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1426–1436, 2018.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020.

Andrea E Martin and Leonidas AA Doumas. A mechanism for the cortical computation of hierarchical linguistic structure. *PLoS biology*, 15(3):e2000663, 2017.

William Matchin and Gregory Hickok. Distinguishing syntactic operations in the brain: Dependency and phrase-structure parsing. *Neurobiology of Language*, pp. 345–362, 2017.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019.

SubbaReddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Liina Pylkkanen and David K Bemis. Building by syntax: The neural basis of minimal linguistic structures. *Cerebral Cortex*, 21:265–273, 2011.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Edmund T Rolls, Marc Joliot, and Nathalie Tzourio-Mazoyer. Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *Neuroimage*, 122:1–5, 2015.

Martin Schrimpf, Idan A. Blank, Greta Tuckute, Conrad Kauf, Eliana A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Jingwei Sheng, Li Zheng, Bingjiang Lyu, Zhehang Cen, Lang Qin, Li Hai Tan, Ming-Xiong Huang, Nai Ding, and Jia-Hong Gao. The cortical maps of hierarchical linguistic structures during speech perception. *Cerebral cortex*, 29(8):3232–3240, 2019.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models, February 2024.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601. ACL, 2019.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.

Qian Wang, Lu Luo, Na Xu, Jing Wang, Ruolin Yang, Guanpeng Chen, Jie Ren, Guoming Luan, and Fang Fang. Neural response properties predict perceived contents and locations elicited by intracranial electrical stimulation of human auditory cortex. *Cerebral Cortex*, 34(2):bhad517, 2024.

Alex Warstadt, Alicia Parrish, Haokun Liu, Aditi Mohananey, Wei Peng, and Samuel R. Bowman. Learning which features matter: Roberta acquires a preference for linguistic generalizations (eventually). *arXiv preprint arXiv:2004.14847*, 2020.

Na Xu, Baotian Zhao, Lu Luo, Kai Zhang, Xiaoqiu Shao, Guoming Luan, Qian Wang, Wenhan Hu, and Qun Wang. Two stages of speech envelope tracking in human auditory cortex modulated by speech intelligibility. *Cerebral Cortex*, 33(5):2215–2228, 2023.

A Appendix

A.1 Alignment pipeline for syntactic processing between LLMs and the human brain

This appendix we provide the detailed pipeline used to align the syntactic representations in LLMs with those in the human brain, focusing on identifying and comparing sentence/phrase representations across both systems.

Data and Experimental Setup To maintain consistency between the LLM and human experiments, we used the same two corpora: a **sentence** corpus (four-syllable Chinese sequence) and a **phrase** corpus (two-syllable Chinese sequence). The word-order randomized version of each corpus was used as a control condition, as detailed in Section 3.1. Each corpus included 40 trials and each trial contains 36 syllables. For SRDM calculation, the corpora were divided into six experimental conditions, each with 20 trials. Sentence/phrase representations of the last 32 syllables were extracted from both LLMs and human subjects to reduce the strong responses at word onset in both systems. The representations are then transformed into the frequency domain.

Frequency-Domain Transformation and Similarity Metrics For the LLMs, neuron activations were transformed using FFT to capture the frequency components of structure processing across the six conditions. From this transformation, we calculated the cosine similarity between each pair of conditions, constructing an SRDM for each MLP neuron. Similarly, for the human brain, we calculated the ITPC to capture frequency-domain representations for each brain channel, producing a channel SRDM.

To assess the structure alignment between LLMs and the human brain, we computed the Spearman correlation between the SRDM of each LLM layer and the SRDM of each brain channel. We then grouped the model SRDMs at the layer level by averaging the cosine similarity across neurons for a fixed layer. The top 100 most relevant brain channels for each model layer were identified based on Spearman correlation, and the overlap of sentence/phrase channels in these top 100 channels was evaluated using a chi-square test. A significant overlap indicated alignment in structure processing between LLMs and brain ROIs.

Model-Brain Similarity and Model-Region Similarity Two key metrics were defined to quantify the structure alignment between LLMs and the human brain. The first, model-brain similarity ($S_{m,b}$), represents the overall similarity of syntactic processing between an LLM and the human brain. It is calculated as the average Spearman correlation between the SRDM of each LLM layer and the top 100 most relevant brain channels:

$$S_{m,b} = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{100}\sum_{i\in \mathrm{top}(j)}^{100}\rho(L_j, C_i), \tag{4}$$

where M is the number of layers of an LLM; L_j and C_i denote the model layer and the brain channel with the indices j and i respectively; $\mathrm{top}(j)$ means the indices of top 100 channels in terms of model layer L_j; and $\rho(L_j, C_i)$ denotes the Spearman correlation between the model SRDM at layer L_j and the SRDM for brain channel C_i.

The second metric, model-region similarity (S_{m,b_r}), measures the alignment between LLMs and specific brain ROIs. This is calculated by averaging the Spearman correlation for the top 100 channels within a particular brain ROI:

$$S_{m,b_r} = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{n(j,r)}\sum_{i\in \mathrm{top}(j)\cap\mathbb{C}_r}^{n(j,r)}\rho(L_j, C_i), \tag{5}$$

where $\mathbb{C}_r$ denotes the indices of all channels belonging to the specific region r, and $n(j,r)$ means the total number of indices in $\mathrm{top}(j)\cap\mathbb{C}_r$, namely the number of channels belonging to region r and at the same time within the top 100 channels in terms of model layer L_j.

Contribution Ratio of Brain ROIs To further investigate the role of specific brain ROIs in syntactic processing, we introduced the contribution ratio (CR_r). The contribution ratio highlights which brain ROIs contribute most significantly to the syntactic alignment between LLMs and the human brain. Fixing a model layer, this metric quantifies the influence of each brain ROIs by calculating the proportion of channels from a given region within the top 100 most relevant channels, normalized by the overall representation of the ROIs (results can be found in Appendix A.4). The contribution ratio is defined as:

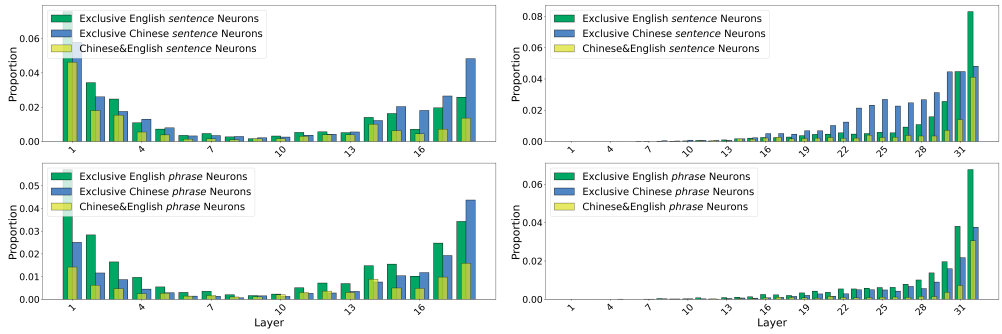
$$CR_r(L_j) = \frac{N_r^{\text{top}}(L_j)/N^{\text{top}}}{N_r^{\text{total}}/N^{\text{total}}}, \tag{6}$$

where $N_r^{\text{top}}(L_j)$ is the number of channels in region r within the top 100 channels in terms of the LLM layer L_j, N^{top} is the total number of top channels, which is specified as 100 in this case, N_r^{total} is the total number of channels in region r, and N^{total} is the total number of brain channels.

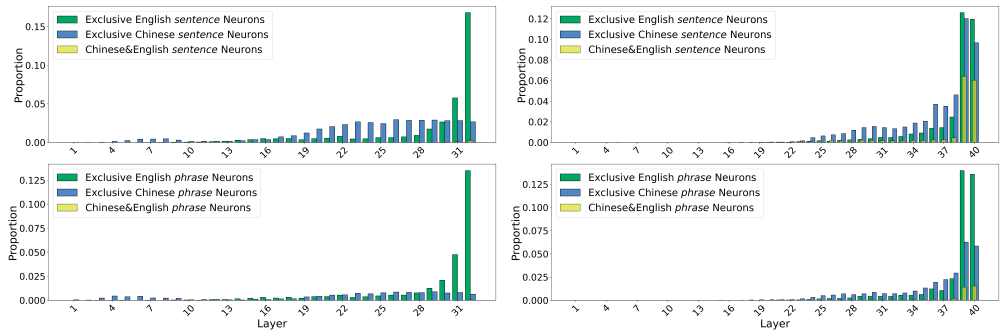
A.2 Sentences and phrases representations in multilingual LLMs

Previous studies have explored how LLMs handle different languages, concluding that while most neurons are shared across languages, a smaller subset of neurons is dedicated to processing specific languages (Tang et al., 2024). But does this hold true for syntactic structure perception? This appendix provides insights into this question. The results in Figure 7 suggest that language-specific syntactic neurons (i.e., exclusive sentence/phrase neurons) tend to cluster toward the final layers of Llama 2, Llama 3.1, and GLM-4, with the proportion of bilingual neurons (Chinese & English) increasing progressively in deeper layers. In contrast, Gemma displays a different pattern, where

both language-specific and bilingual neurons are found not only in deeper layers but also in the initial layers.



(a) Gemma (b) Llama 2



(c) Llama 3.1 (d) GLM-4

Figure 7: Cross-language neural representations extracted from four multilingual models (Gemma, Llama 2, Llama 3.1, and GLM) depicting syntactic processing capabilities.

Interestingly, Llama 3.1 shows a notably lower count of Chinese-specific neurons compared to English-specific neurons, and fewer Chinese & English neurons than the other three multilingual LLMs. Although Llama 3.1 was pre-trained on 176 languages (Dubey et al., 2024), it appears to have less specialization in Chinese, which may explain the reduced presence of Chinese-specific neurons and, consequently, fewer bilingual neurons. It is important to note that GPT-2 is a monolingual model designed for English, so bilingual representation comparisons were not applicable for this model.

A.3 Alignment results on phrase-level syntactic representations

In this appendix, we present the alignment results for phase-level syntactic representations. These results closely mirror those observed in the sentence-level analysis (see Table 1), reinforcing the overall consistency of our findings. As we can see from Table 2, the alignment results highlight the effectiveness of the HFTP in capturing phase-level syntactic representations. GPT-2 exhibits a strong average correlation in the left hemisphere (r = 0.647), while Llama 2 shows a comparable alignment (r = 0.645). Similar to the findings for sentence-level processing, notably, the syntactic structure representations of Llama 2 also align most closely with the left STG, MTG, and IFG. In contrast, Llama 3.1 exhibits a lower correlation (r = 0.516), suggesting that enhancements in model architecture do not necessarily lead to better alignment with human brain activity. Additionally, other LLMs also demonstrate relatively high S_{m,b_r} values in these key brain ROIs.

A.4 Contribution Ratios of LLMs

In this appendix we present the contribution ratio results for five large language models (LLMs) used in this study: GPT-2, Gemma, Llama 2, Llama 3.1, and GLM-4. The contribution ratios for

Table 2: Averaged top 100 Spearman correlation coefficients between SRDMs of sEEG channels and those of MLP neurons under the **phrase** corpus condition, separated by left (L) and right (R) hemispheres. Note that '/' denotes cases where the model lacks channels corresponding to the brain ROIs in that hemisphere within the top 100 Spearman-ranked channels The values in the rows corresponding to brain ROIs represent the model-region similarity S_{m,b_r}.

	GPT-2		Gemma		Llama 2		Llama 3.1		GLM-4	
	L	R	L	R	L	R	L	R	L	R
$S_{m,b}$	**0.647**	0.455	0.591	0.434	0.645	0.440	0.516	0.400	0.600	0.439
A1	**0.678**	0.389	**0.625**	0.339	0.594	0.625	0.525	/	0.585	/
STG	0.652	0.430	0.599	0.361	**0.661**	0.443	0.505	0.357	**0.636**	0.414
MTG	0.652	0.396	**0.637**	0.398	**0.696**	0.461	0.527	0.260	**0.603**	0.421
ITG	0.628	0.464	0.577	0.425	0.642	0.426	0.510	0.422	0.597	0.430
Insula	0.605	0.495	0.547	0.438	0.619	0.472	0.492	0.388	0.577	0.454
TPJ	0.617	0.447	0.586	0.335	0.623	0.357	**0.544**	0.378	0.565	0.483
Temporal Pole	**0.675**	0.462	0.580	0.514	0.603	0.462	0.461	**0.540**	0.582	0.405
Sensorimotor	0.629	0.487	0.587	0.426	0.615	0.442	0.516	0.358	0.573	0.494
IFG	**0.712**	0.458	0.577	0.490	**0.650**	0.5	0.475	0.458	**0.602**	0.438
MFG	0.610	0.421	0.603	**0.635**	0.558	0.300	0.475	0.388	0.517	0.398
Hippocampus	0.632	0.399	0.525	0.479	0.564	0.372	**0.571**	0.370	0.503	0.454
Amygdala	/	0.471	/	0.560	/	0.516	/	0.382	/	0.451

each model were computed in a manner consistent with the methodology outlined in the main paper. Specifically, the contribution ratio for each model was calculated based on the number of top 100 significant channels within each brain ROIs, as described in Appendix A.1. Below, we present the results for both the left (L) and right (R) hemispheres of each model (See Figures 8, 9, 10, 11 and 12). These figures offer further insights into how different LLMs align with human brain ROIs in terms of syntactic processing.

From these figures, we observe that across all LLMs, regions such as A1 and STG in the left hemisphere, and the Insula, Temporal Pole, and Amygdala in the right hemisphere contribute more significantly to the alignment with human brain syntactic processing. These regions are known to be involved in language-specific processes in the human brain, particularly in the left hemisphere, where the STG and A1 are crucial for auditory and syntactic processing. The alignment between these brain ROIs and the LLMs suggests that these models may be capturing aspects of hierarchical syntactic structures in ways that are functionally similar to human neural mechanisms. The Insula, Temporal Pole, and Amygdala, though not traditionally highlighted as primary language regions, may also play supporting roles in language comprehension, possibly through emotion and memory-related pathways. This suggests that LLMs might engage both language-specific and auxiliary brain ROIs to process syntax, mirroring the integrated and distributed nature of human brain networks involved in language processing.

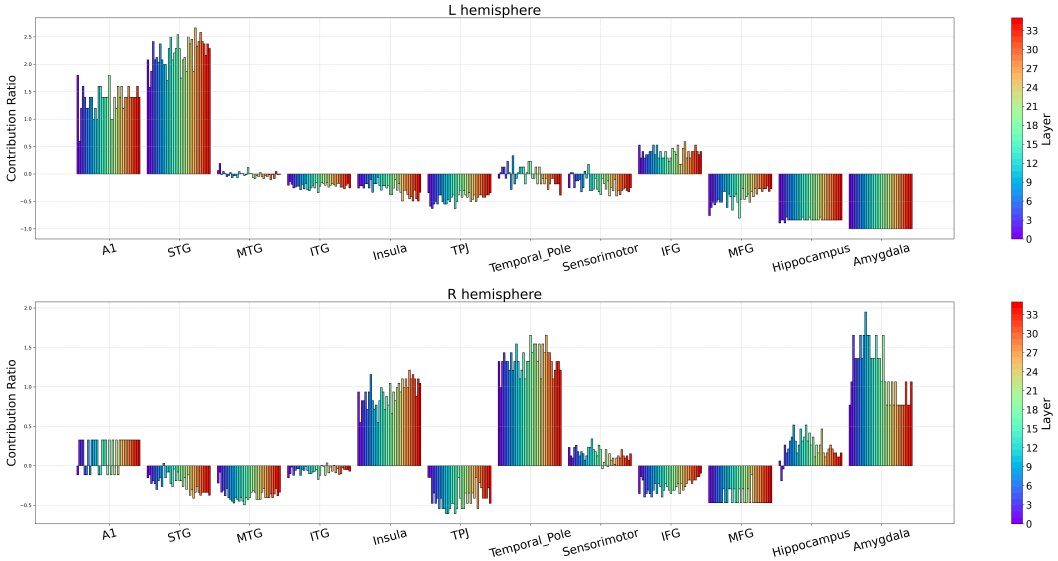


Figure 8: Contribution ratios for GPT-2 Chinese model: Left hemisphere (top) and Right hemisphere (bottom).

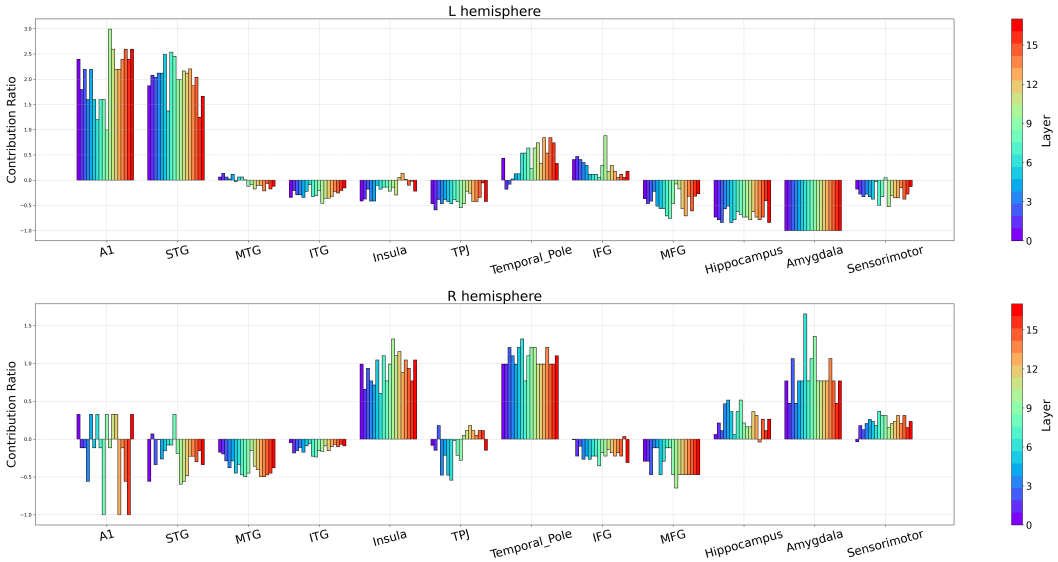


Figure 9: Contribution ratios for Gemma model: Left hemisphere (top) and Right hemisphere (bottom).

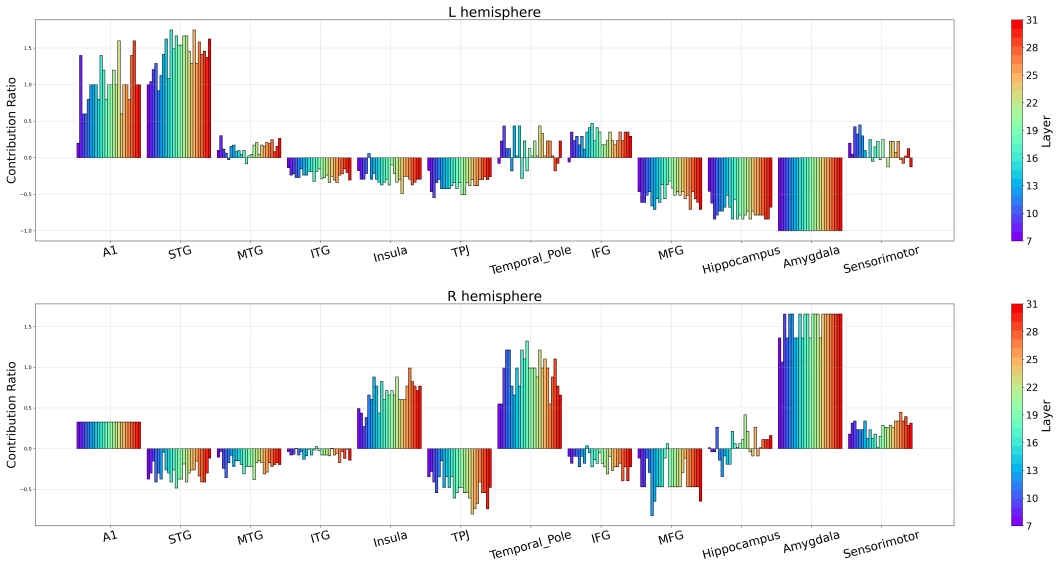


Figure 10: The contribution ratios of the left (up) and right (bottom) hemispheres in the Llama-2-7b model. The upper bar highlights the brain ROIs that contribute most significantly to syntactic processing.

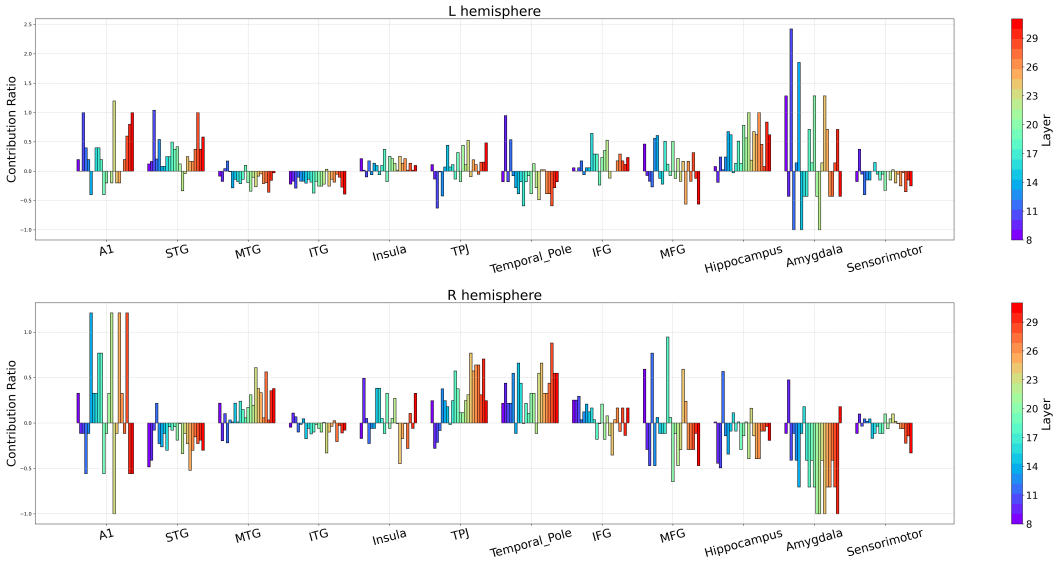


Figure 11: Contribution ratios for Llama 3.1 model: Left hemisphere (top) and Right hemisphere (bottom).

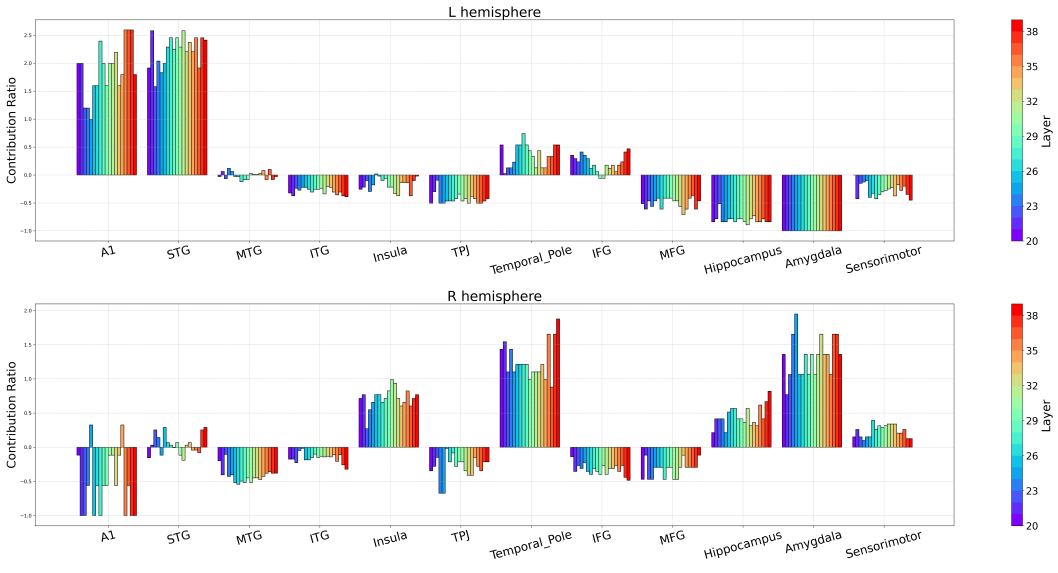


Figure 12: Contribution ratios for GLM-4B model: Left hemisphere (top) and Right hemisphere (bottom).

A.5 Model details

In this appendix, we present the details of the LLMs used in this study. Table 3 summarizes key parameters, including model size, number of layers, attention heads, and MLP neurons.

Table 3: Comparison of model parameters.

Model	Size	Layer	Attention head	MLP neuron
GPT-2 (Radford et al., 2019)	774M	36	20	5120
Gemma (Team et al., 2024)	2B	18	8	16384
Llama 2 (Touvron et al., 2023)	7B	32	32	11008
Llama 3.1 (Dubey et al., 2024)	8B	32	32	14336
GLM 4 (GLM et al., 2024)	9B	40	32	13696

A.6 Syntactic corpora

For the HFTP experiments in LLMs, we reconstructed two syntactic corpora based on (Ding et al., 2016): one comprising sentences with four-syllable sequences in Chinese and the other with four-word sequences in English. These corpora were utilized to assess the syntactic processing capabilities of the models (see Table 4 and Table 5).

For the HFTP experiment in the human brain, we utilized two Chinese corpora: the **sentence** and **phrase** corpora. To ensure consistent analysis of syntactic processing across both LLMs and the human brain, the same corpora were applied to the alignment experiment. The corpus are originated from (Sheng et al., 2019).

Table 4: Chinese syntactic corpus.

Four-syllable Sequences				
老牛耕地	朋友请客	厨师做饭	竹鼠啃笋	农民种菜
青草发芽	和尚念经	老师讲课	鲸鱼喷水	绵羊吃草
英雄救火	游客爬山	鸭子划水	蜘蛛结网	祖父下棋
医生看病	护士打针	母鸡下蛋	行人过街	法官判案
狮子吃肉	老鹰捕鱼	蜜蜂采花	小孩读书	猴子摘果
画家作画	船夫摇桨	诗人吟诗	麻雀筑巢	猴子摘桃
渔夫撒网	骆驼饮水	狐狸捕鼠	海豹顶球	小猫抓鱼
老马拉车	鸽子衔枝	孩童拾贝	雏鸡啄米	山雀捉虫
青鸟啄木	樵夫砍柴	黑熊爬树	土狼挖洞	军鸽传信
燕雀喂仔	野猪拱地	渔民划船	蚯蚓钻土	蚕蛾吐丝

Table 5: English syntactic corpus.

Four-word Sequences		
fat rat sensed fear	wood shelf holds cans	tan girls drove trucks
gold lamps shine light	dry fur rubs skin	sly fox stole eggs
top chefs cook steak	our boss wrote notes	two teams plant trees
all moms love kids	new plans give hope	large ants built nests
teen apes hunt bugs	rude cats claw dogs	rich cooks brewed tea
fun games waste time	pink toys hurt girls	huge waves hit ships
deaf ears hear you	his aunt tied shoes	kind words warm hearts
long fight caused hate	dead sharks leak blood	smart dogs dig holes
slim kids like jeans	sick boys fail tests	rear doors hide cups
pale hands make bread	bad smells fill town	mad foes smack chefs
quiet lamb ate grass	soft fork brings food	green frogs miss flies
black skies show stars	tall guys flee camp	gray goat climb hills
iced beer costs cents	old kings gave speech	blue eyes shed tears
white cars need gas	young child closed doors	thin threads hang plates
their store sold cars	cute cubs drink milk	six farms lost cows
sharp knife cuts cheese	round soap killed germs	loud sound scared mom
weird clowns wear hats	her sons paint walls	

A.7 BRAIN ROIS

As discussed in Section 3.3, we reorganized the original sEEG data by grouping the Automated Anatomical Labeling (AAL) annotations into newly defined brain ROIs for our experiments (Rolls et al., 2015). In this appendix, we provide the full names of AAL regions, the corresponding AAL labels used in the sEEG data, and their mapped brain ROIs in Table 6 .

Table 6: Automated Anatomical Labeling (AAL) annotations from the original sEEG data, along with their mapped brain ROIs. Note that the regions are distinguished by the left and right hemispheres.

AAL Full Name	**AAL Label**	**ROI**
Heschl Gyrus	Heschl	A1
Superior Temporal Gyrus	Temporal_Sup	STG
Middle Temporal Gyrus	Temporal_Mid	MTG
Inferior Temporal Gyrus	Temporal_Inf	ITG
Parahippocampal Gyrus	ParaHippocampal	ITG
Fusiform Gyrus	Fusiform	ITG
Insular Cortex	Insula	Insula
Angular Gyrus	Angular	TPJ
Supramarginal Gyrus	SupraMarginal	TPJ
Inferior Parietal Lobule	Parietal_Inf	TPJ
Superior Temporal Pole	Temporal_Pole_Sup	Temporal_Pole
Middle Temporal Pole	Temporal_Pole_Mid	Temporal_Pole
Paracentral Lobule	Paracentral_Lobule	Sensorimotor
Supplementary Motor Area	Supp_Motor_Area	Sensorimotor
Rolandic Operculum	Rolandic_Oper	Sensorimotor
Precentral Gyrus	Precentral	Sensorimotor
Postcentral Gyrus	Postcentral	Sensorimotor
Inferior Frontal Gyrus, Opercular part	Frontal_Inf_Oper	IFG
Inferior Frontal Gyrus, Triangular part	Frontal_Inf_Tri	IFG
Inferior Frontal Gyrus, Orbital part	Frontal_Inf_Orb	IFG
Middle Frontal Gyrus	Frontal_Mid	MFG
Middle Frontal Gyrus, Orbital part	Frontal_Mid_Orb	MFG
Hippocampus	Hippocampus	Hippocampus
Amygdala	Amygdala	Amygdala