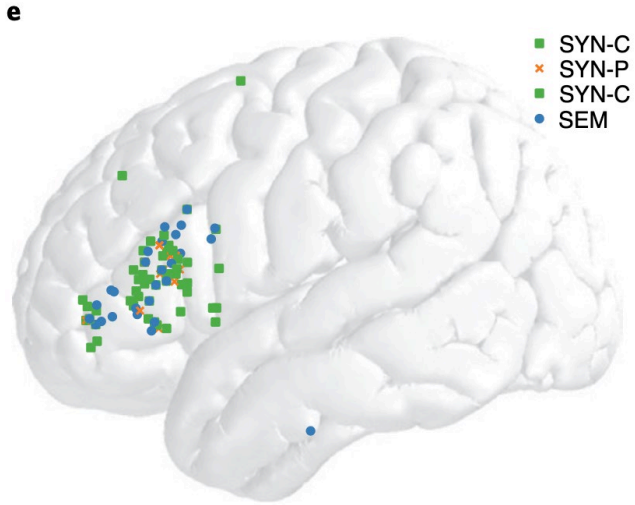
Hierarchical Frequency Tagging Probe (HFTP): A Unified Approach to Investigate Syntactic Structure Representations in Large Language Models and the Human Brain

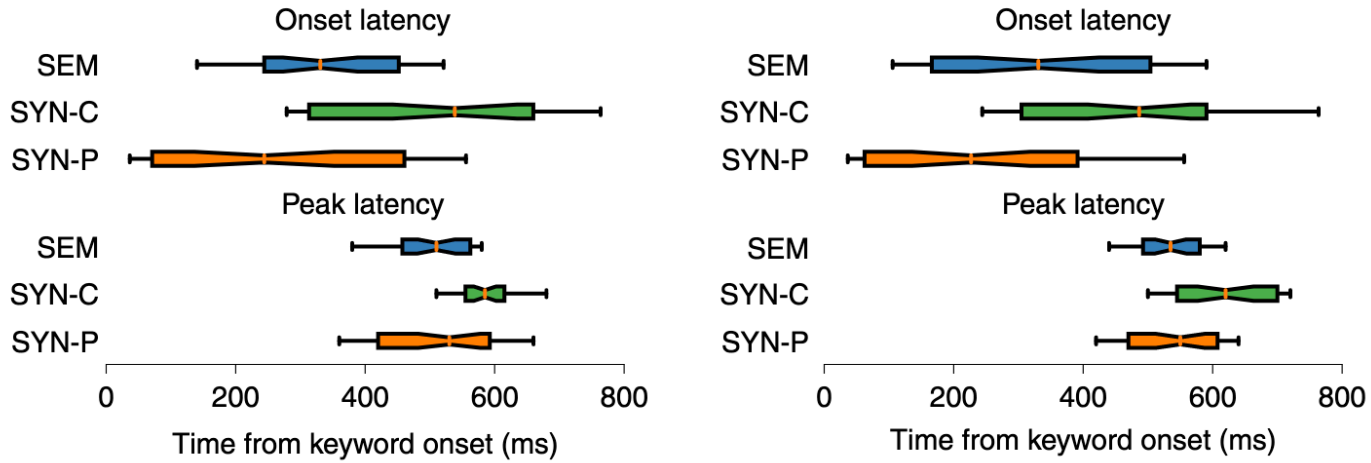
Jingmin AN

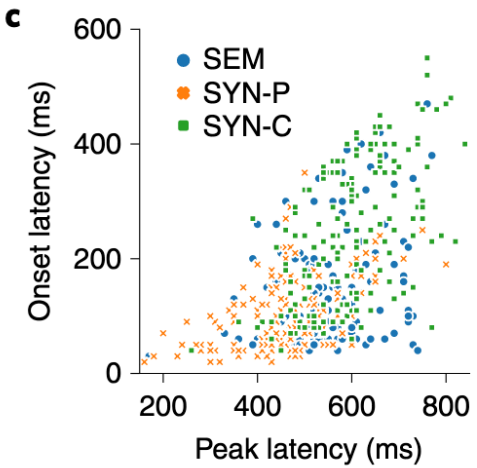
Center for Life Sciences, Peking University

Beijing, P.R. China

Background - Syntactic study in the human brain



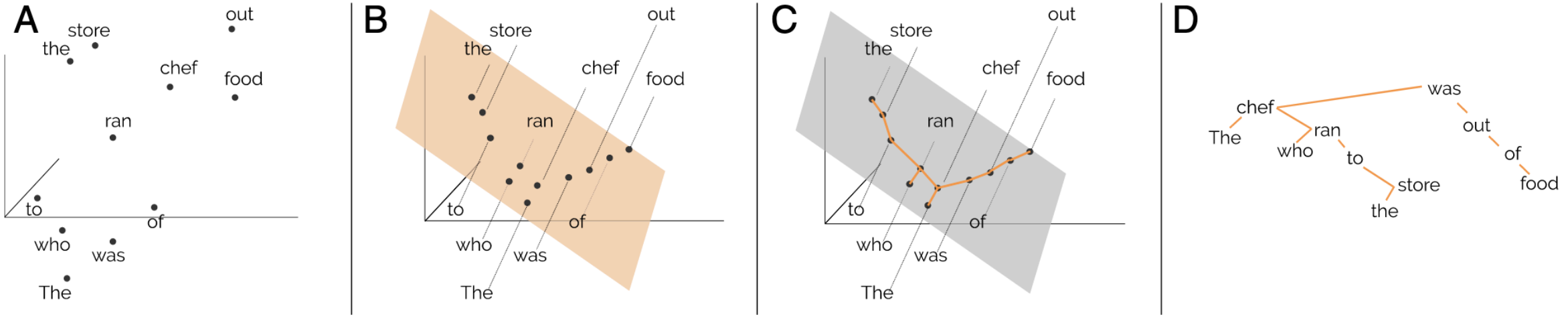




- Using syntactic-violation and semantic paradigms, single-electrode analyses in the **IFG** indicate that, **Chinese syntax is processed before semantics**; at larger scales, however, the two are **intertwined**.

Zhu, Y. et al. Distinct spatiotemporal patterns of syntactic and semantic processing in human inferior frontal gyrus. Nat Hum Behav (2022).

Background - Structural probe in language models



Each of the words of the sentence *The chef who ran to the store was out of food* is internally represented in context as a vector.

A structural probe finds a linear transform of that space under which squared L_2 distance between vectors best reconstructs tree path distance between words.

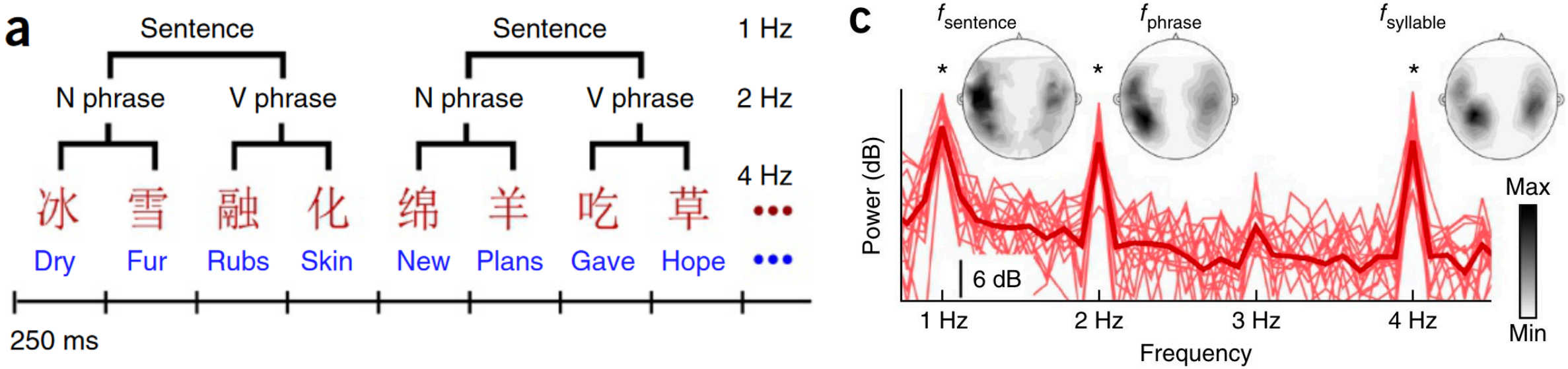
Once in this latent space, the structure of the tree is globally represented by the geometry of the vector space, meaning words that are close in the space are close in the tree.

In fact, the tree can be approximately recovered by taking a minimum spanning tree in the latent syntax space.

- A structural probe **linearly projects language-model vectors so distances approximate parse paths**, enabling a minimum spanning tree to recover syntax.

Manning, C. D. et al. Emergent linguistic structure in artificial neural networks trained by self-supervision. PNAS (2020).

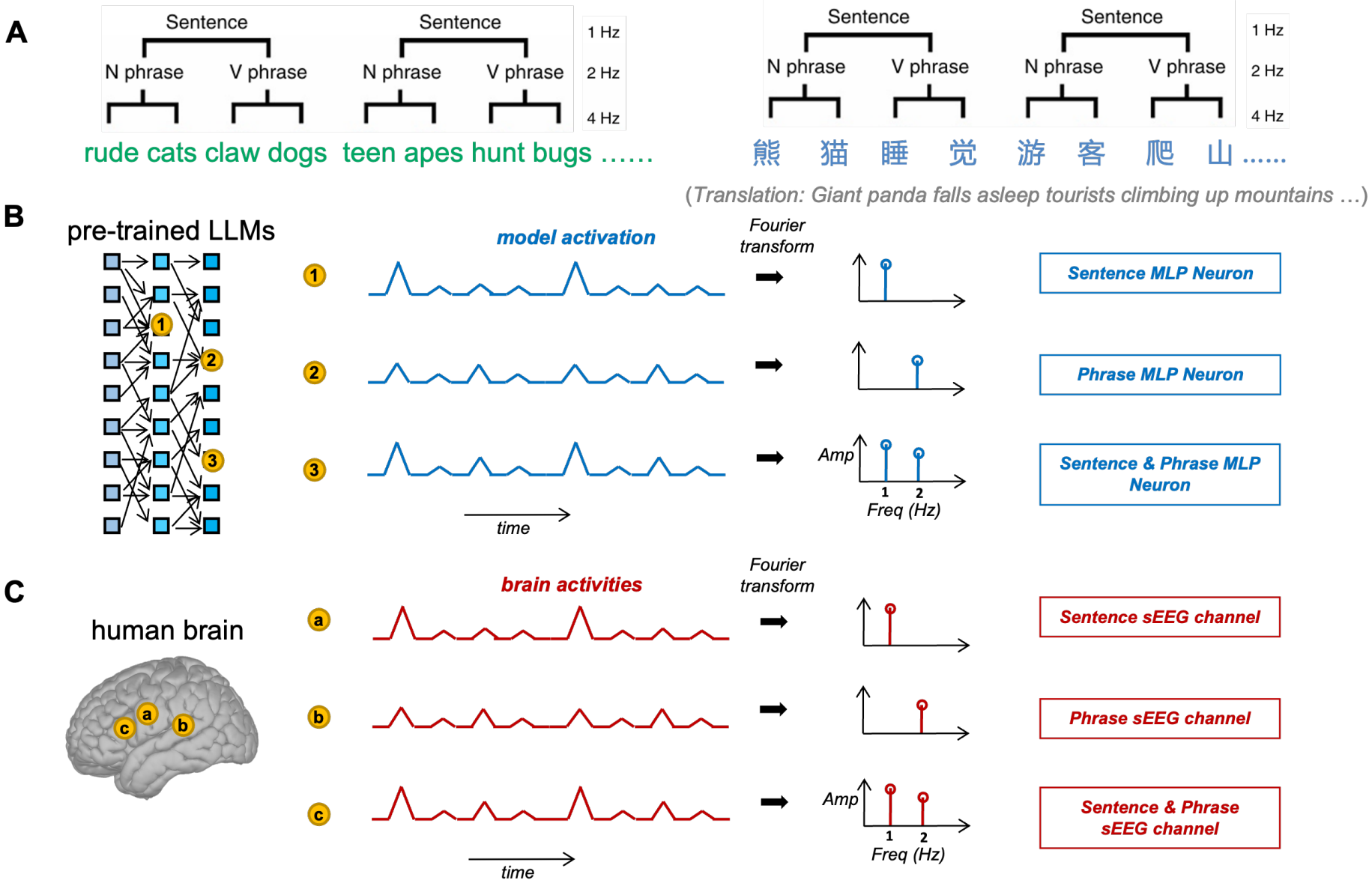
Background - Hierarchical Frequency Tagging (HFT)



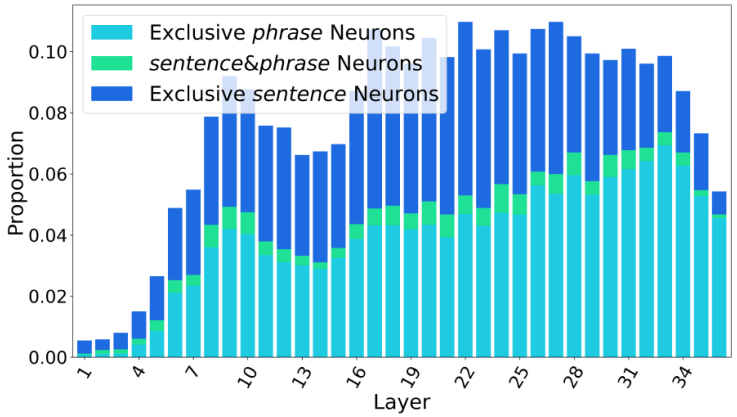
- For every four-word collocation, **participants** can intrinsically encode **four syllables (4Hz)**, **two phrases (2Hz)** and **one sentence (1Hz)**, and this hierarchical pattern tracking **generalizes across languages**.

Ding, N., et al. Cortical tracking of hierarchical linguistic structures in connected speech. Nature Neuroscience (2016)

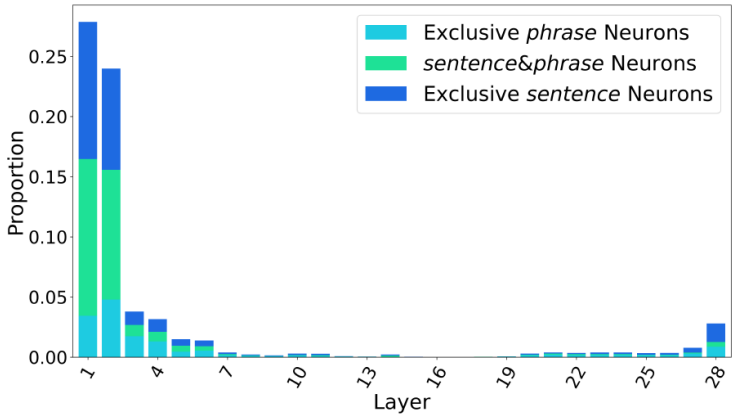
Workflow


A
Sentence
N phrase
V phrase
Sentence
N phrase
V phrase
1 Hz
2 Hz
4 Hz
rude cats claw dogs teen apes hunt bugs ……
Sentence
N phrase
V phrase
Sentence
N phrase
V phrase
1 Hz
2 Hz
4 Hz
熊 猫 睡 觉 游 客 爬 山 ……
(Translation: Giant panda falls asleep tourists climbing up mountains …)
B
pre-trained LLMs
1
2
3
model activation
Fourier transform
Sentence MLP Neuron
Phrase MLP Neuron
Amp
1 2
Freq (Hz)
Sentence & Phrase MLP Neuron
time
C
human brain
a
b
c
brain activities
Fourier transform
Sentence sEEG channel
Phrase sEEG channel
Amp
1 2
Freq (Hz)
Sentence & Phrase sEEG channel
time

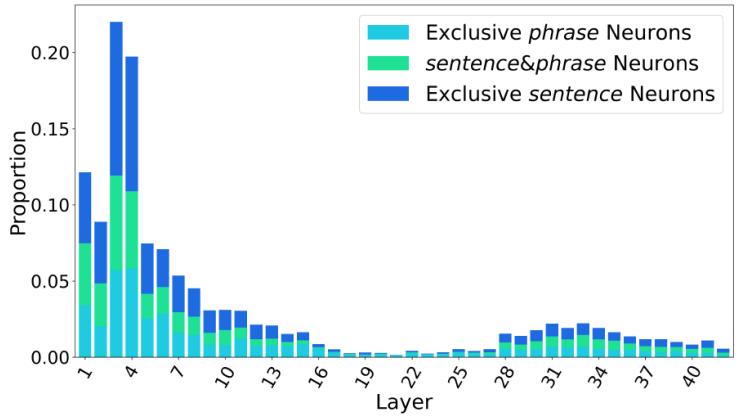
Results - Distribution of syntactic neurons



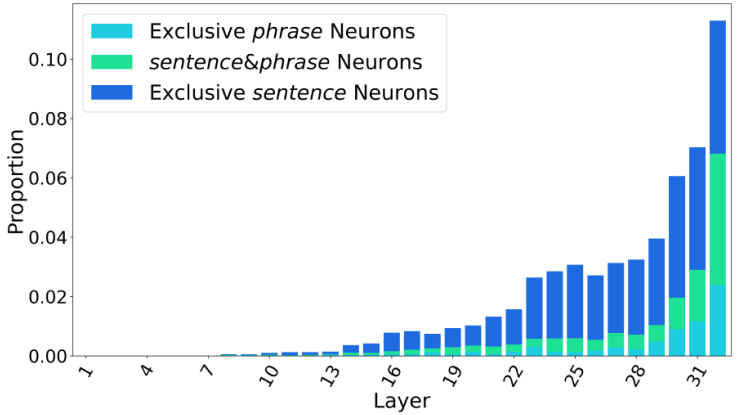
(a) GPT-2



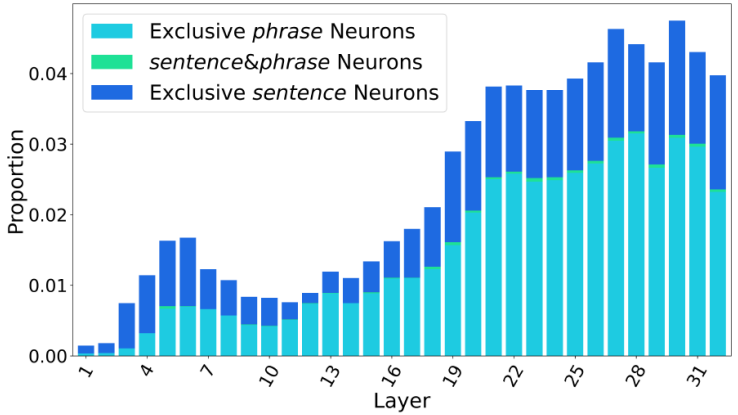
(b) Gemma



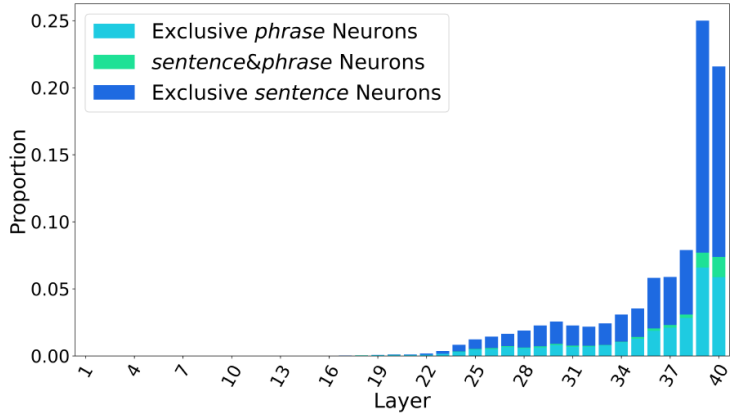
(c) Gemma 2



(d) Llama 2

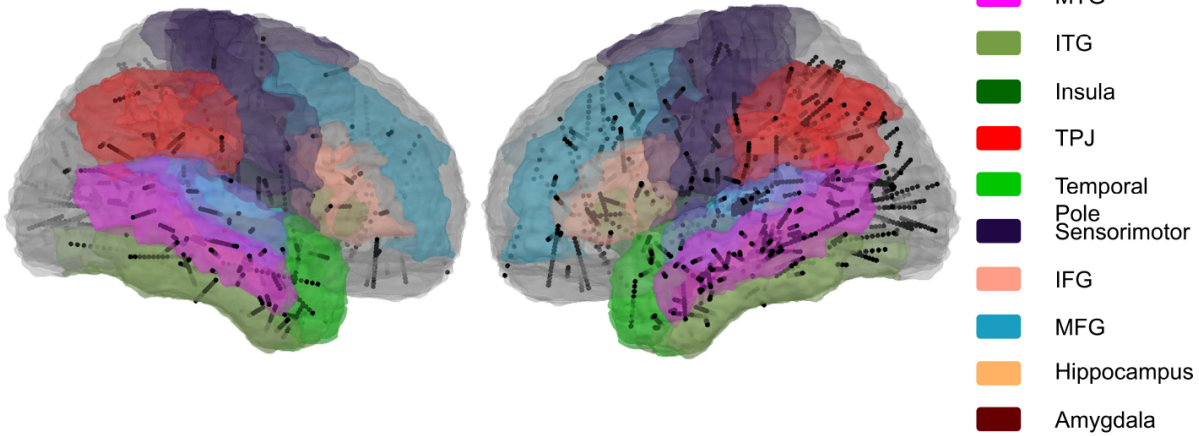


(e) Llama 3.1

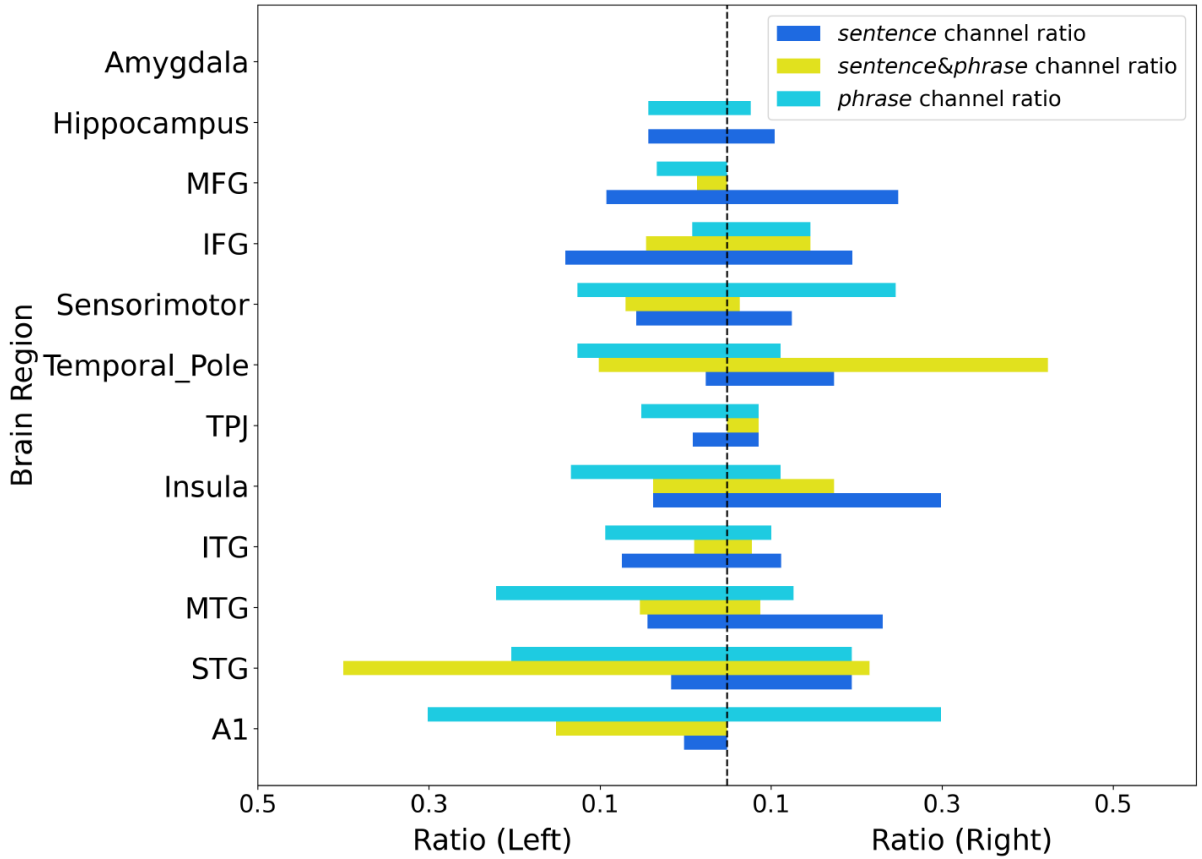


(f) GLM-4

Results - Distribution of syntactic channels and Brain ROIs

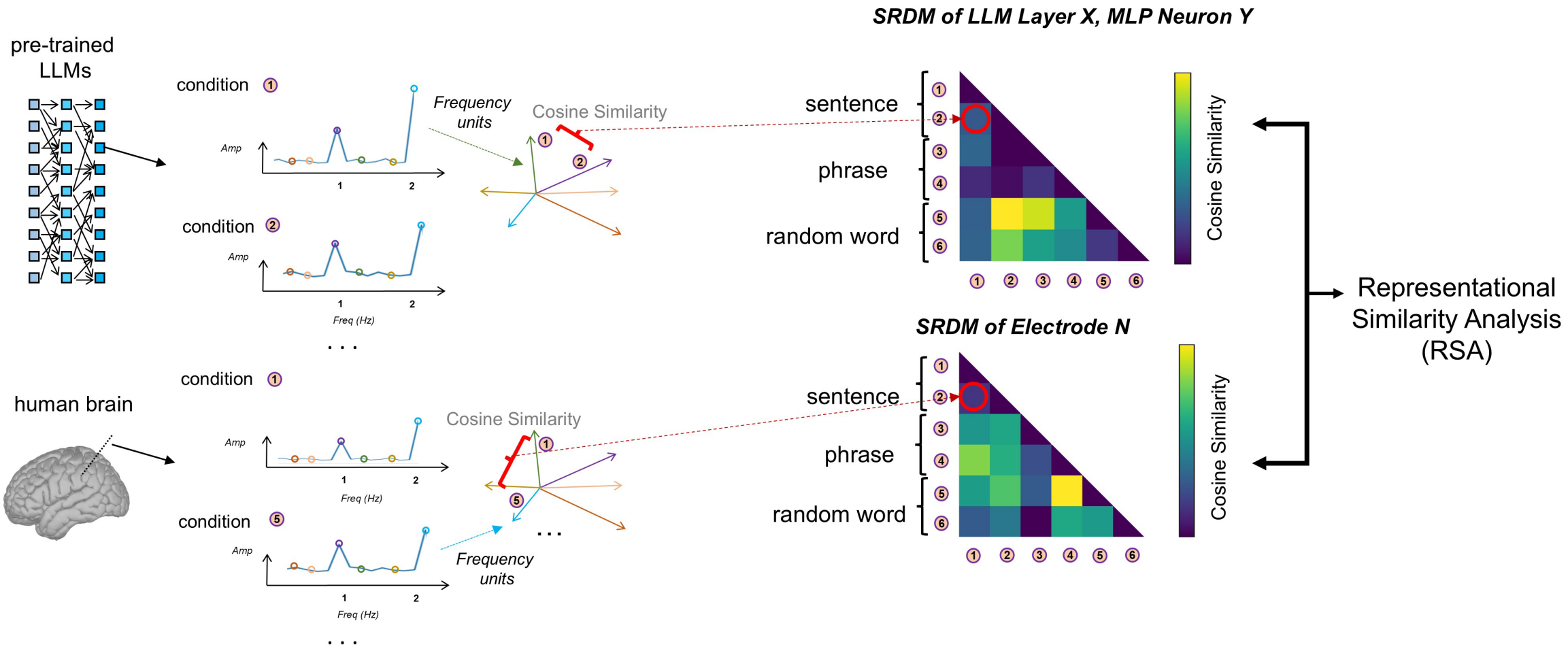


(a) sEEG channel locations and Brain ROIs



(b) Significant sEEG channel distribution

Alignment pipeline



- We compute **Structure Representational Dissimilarity Matrixes (SRDMs)** from cosine similarities across conditions for sentence, phrase units in LLMs and the human brain, then apply **Representational Similarity Analysis (RSA)** between model and brain SRDMs to quantify model–brain correspondence.

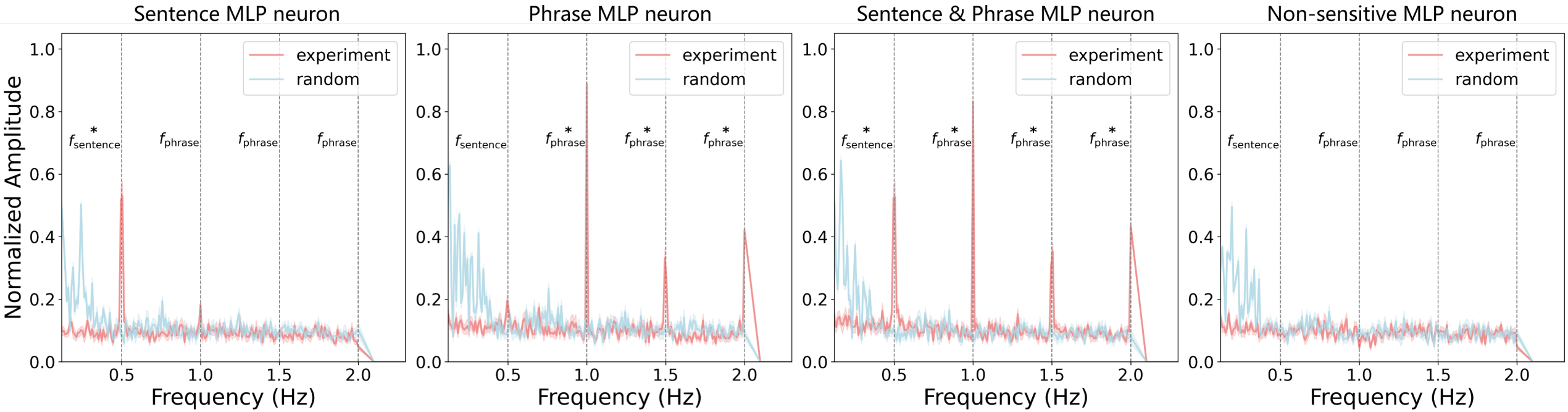
Results – Representational Alignment

	GPT-2		Gemma		Gemma 2		Llama 2		Llama 3.1		GLM-4	
	L	R	L	R	L	R	L	R	L	R	L	R
$S(m, b)$	**0.654**	0.442	0.582	0.411	0.644	0.450	0.645	0.439	0.514	0.405	0.630	0.445
A1	**0.683**	0.423	**0.642**	0.358	**0.702**	0.333	0.649	0.547	0.514	0.403	**0.664**	0.374
STG	0.667	0.422	**0.593**	0.386	0.654	0.410	**0.672**	0.453	0.507	0.392	**0.647**	0.412
MTG	0.674	0.392	0.584	0.383	**0.659**	0.411	**0.67**4	0.409	**0.521**	0.408	0.645	0.397
ITG	0.637	0.444	0.578	0.406	0.631	0.448	0.629	0.426	0.509	0.401	0.615	0.439
Insula	0.624	0.460	0.551	0.425	0.600	0.476	0.630	0.446	**0.518**	0.422	0.604	0.475
TPJ	0.610	0.452	0.566	0.373	0.641	0.410	0.619	0.400	**0.518**	0.408	0.606	0.438
Temporal Pole	0.648	0.473	0.556	0.470	0.643	0.558	0.610	0.469	0.494	0.448	0.616	0.483
Sensorimotor	0.637	0.462	0.567	0.426	0.622	0.448	0.624	0.446	0.505	0.396	0.617	0.463
IFG	**0.694**	0.463	**0.603**	0.466	**0.670**	0.496	**0.665**	0.491	0.513	0.410	**0.646**	0.490
MFG	0.615	0.436	0.557	0.401	0.585	0.489	0.597	0.367	0.510	0.397	0.588	0.473
Hippocampus	**0.698**	0.405	0.553	0.408	0.626	0.428	0.657	0.413	0.534	0.390	0.613	0.434
Amygdala	/	0.489	0.566	0.454	/	0.472	/	0.558	0.496	0.377	/	0.508

* **sentence** corpus

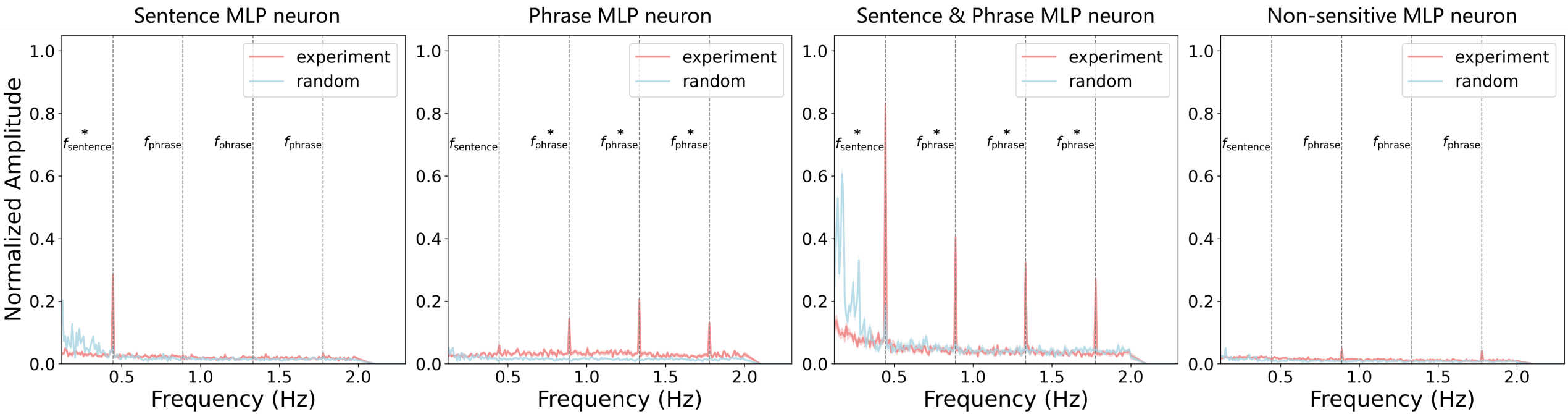
- **GPT-2** showed the strongest alignment in left hemisphere.
- Advancing models **diverge** in model-brain alignment (**e.g., Gemma 2** ↑, **Llama 3.1** ↓), uncoupling performance gains from syntactic brain-likeness.

Results – HFTP on naturalistic corpus (8-word)



- In the 8-word English naturalistic corpus, LLMs exhibit robust peaks at **the sentence rate (0.5 Hz)** and at **phrase rates (1, 1.5, 2 Hz)**, demonstrating the generalizability of our method.

Results – HFTP on naturalistic corpus (9-word)



- In the 9-word English naturalistic corpus, LLMs exhibit robust peaks at **the sentence rate (~0.44 Hz)** and at **phrase rates (~0.89, 1.33, 1.78 Hz)**, demonstrating the generalizability of our method.

Conclusion

- We introduced the **Hierarchical Frequency-Tagging Probe (HFTP)**, a unified framework that **probes internal representational structure** and **systematically assesses the alignment of syntactic representations** between LLMs and the human brain.
- For LLMs, syntactic units cluster at different layers across models: some concentrate in **early layers** while some cluster in **later layers**.
- For the human brain, sentence- and phrase-selective sEEG channels **cluster in left hemisphere—A1, STG, MTG, IFG—with fewer right-hemisphere**.
- Representational alignment is strongest in left-lateralized language regions and varies by model family; **upgrades do not monotonically improve brain alignment**.