
Scaling and context steer LLMs along the same computational path as the human brain

Joséphine Raugel

Meta AI

Laboratoire de Neurosciences Cognitives
et Computationnelles (Inserm U960)
Ecole Normale Supérieure - PSL

Stéphane d’Ascoli

Meta AI

Jérémy Rapin

Meta AI

Valentin Wyart*

Laboratoire de Neurosciences Cognitives
et Computationnelles (Inserm U960)
Ecole Normale Supérieure - PSL

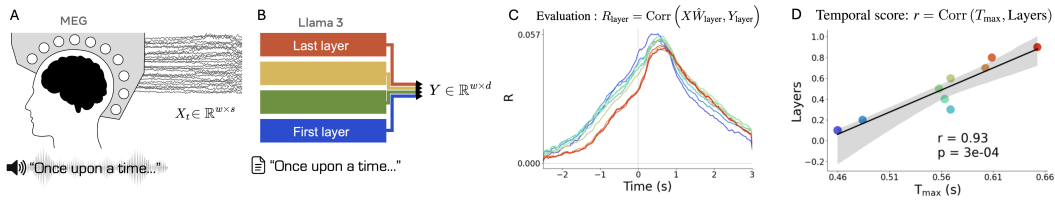
Jean-Rémi King*

Meta AI

*shared senior authorship

Abstract

Recent studies suggest that the representations learned by large language models (LLMs) are partially aligned to those of the human brain. However, whether and why this alignment score arises from a similar sequence of computations remains elusive. In this study, we explore this question by examining temporally-resolved brain signals of participants listening to 10 hours of an audiobook. We study these neural dynamics jointly with a benchmark encompassing 17 LLMs varying in size and architecture type. Our analyses confirm that LLMs and the brain generate representations in a similar order: specifically, activations in the initial layers of LLMs tend to best align with early brain responses, while the deeper layers of LLMs tend to best align with later brain responses. This brain-LLM alignment is consistent across transformers and recurrent architectures. However, its emergence depends on both model size and context length. Overall, this study sheds light on the sequential nature of computations and the factors underlying the partial convergence between biological and artificial neural networks.



1 Introduction

Motivation. While large language models (LLMs) are not designed to resemble the human brain, recent studies show that their activations share similarities with those of the brain in response to speech [1, 2, 3, 4, 5, 6]. In the same way bats and birds independently evolved wings [7], LLMs and the human brain thus seem to follow a partial convergence [4].

State of the art. Recent studies showed that an anatomical alignment exists between LLM layers and functional regions of the human brain, in the sense that their first layers tend to align with low-level areas of the brain such as primary sensory cortices whereas their deeper layers tend to align with higher-level areas such as secondary sensory or associative cortices [5, 8, 3, 9, 1].

Remaining challenges. While this *anatomical* alignment between LLMs and the brain is increasingly established, the *order* in which these representations emerge remains poorly understood. While it has been shown by [6] that GPT2-XL exhibits a form of temporal alignment, (i) whether this temporal alignment with the brain is systematically found across LLMs, (ii) whether this alignment depends on the type of architecture, (iii) on its size, or (iv) on the length of its context remains currently unknown. In sum, the factors that lead an LLM to adopt a computational path analogous to the human brain’s remain unknown.

Approach. To address these issues, we analyze the temporally resolved brain signals of healthy individuals recorded with magnetoencephalography (MEG) [10], while they listened to 10 hours of audiobooks. We then systematically analyze these neural dynamics in conjunction with a benchmark of 17 LLMs varying in architecture, size and training choices.

2 Methods

Problem formalization. We compare (i) the representations of the human brain in response to natural speech, to (ii) the representations of LLMs in response to the corresponding textual input. Brain activity, here measured with magnetoencephalography (MEG), leads to a high-dimensional time series that depends, in part, on speech input. To test whether the order of computations is similar between the brain and LLMs, we test whether the orders in which representations are generated are correlated between the two systems.

Linear mapping. Following others [11, 12, 13], we operationally define a neural “representation” as “linearly readable information”. As there is no one-to-one alignment between each MEG sensor and each activation of an LLM, we compare the representations across these two systems through a linear mapping. Specifically, we fit a ridge regression to predict LLM activations ($Y \in \mathbb{R}^{w \times d}$) from brain activity ($X_t \in \mathbb{R}^{w \times s}$):

$$\hat{W} = \arg \min_W \{ \|Y - XW\|_2^2 + \lambda \|W\|_2^2 \}$$

with w the number of words, d the number of LLM activations, s the number of MEG sensors and t the time point relative to word onset. For this, we use scikit-learn’s RidgeCV, with tuning of logarithmically spaced regularization strength through a grid search approach ($\alpha = 10^{-4}$ to 10^8 , tuned for each dimension independently).

Alignment score. To evaluate this alignment score between an LLM and a human brain, we compute, for each time sample relative to word onset, a Pearson correlation between Y and WX_t on a held-out test set. We repeat this procedure across all five train–test folds of the cross-validation.

Temporal alignment. After computing this LLM-brain alignment score of $l = 9$ equally spaced layers of the LLM, we evaluate whether the time at which the score peaks correlates with the depth of the layer in the LLM considered. We refer to this as *temporal alignment*. Specifically, we compute, for each layer $T_{\max}$, the mean of the temporal window during which $\tilde{R} \geq 95\%$, where $\tilde{R}$ is the normalized alignment score of the layer, obtained by dividing the alignment score by its maximum value across time. Finally, we compute the Pearson correlation between the $T_{\max}$ and the relative depth of the 9 layers. The result of this correlation is hereafter referred to as *temporal score*.

Brain data and preprocessing. We here focus on a large within-subject MEG dataset publicly available [10]. This dataset consists of three healthy participants who listened to 10 h of audio books

in a CTF MEG scanner. To limit the impact of noise we apply a band-pass filter between 0.1 and 20 Hz, down-sample the signal at 30 Hz, time-lock the brain responses to individual words, and epoch the corresponding neural data between -2.5 s and +3 s relative to word onset using MNE-Python [14]. Finally, we z-score MEG signals across words, for each MEG channel and each time point independently.

LLM activations and preprocessing. We use a selection of SOTA LLMs to ensure a comprehensive evaluation ranging through architectures, scales, design and training choices. Specifically, we benchmark models such as Llama-3-8B [15, 16], Llama-3.2 (1B, 3B) [15, 17, 18], Mistral-7B-v0.1 [19], Gemma-7B [20], Qwen1.5-7B [21], and GPT-2-XL [22, 23], the latter serving as a historical reference. We also investigate dynamics at play in SOTA state space models: Mamba-1.4B-hf [24, 25] and RecurrentGemma-9B [26]. Additionally, we leverage the Pythia family [27] consisting of 8 models of increasing size and same training setup. Except when explicitly stated, context length is 50 words. For each LLM, we investigate 9 layers linearly distributed between 10% and 90% of the model hierarchy. To mitigate the issue of heterogeneous sizes, we transform these activation patterns with a Principal Component Analysis ($n=50$) with scikit-learn [28].

Text preprocessing. To ensure the processing of the most semantically meaningful words, we study only content words (as opposed to function words), specifically those which belong to the following part-of-speech categories as defined by Spacy [29]: NOUN, VERB, ADJ, ADV. We ensure the replicability of findings for all words (function and content) via control figures in App. F.

Compute resources. Evaluating the largest LLM on our dataset to extract both activation patterns for all 9 layers and next-word probability requires 3 V100 GPU-hours and 12 CPU-hours. Pre-processing both LLMs’ and brain’s activations, fitting and testing a decoding model on our dataset, for each layer and each time step along our temporal window of 5.5s requires in total 4 hours x 9 layers x 15CPUs (Allocated memory: 100 GB) x 3 subjects = 1,620 CPU-hours - for the largest LLM. In total, experiments on 17 LLMs and 9 varying contexts required ~75 GPU-hours and ~42,500 CPU-hours. An internal cluster was used for all experiments. Licenses for the models and dataset used are presented in App. J.

3 Results

Alignment score. We compare the representations of LLMs to those of the human brain in response to natural speech. For this, we fit a linear model to predict the LLMs’ contextual representations from the MEG activations, time locked to word onset, and we evaluate this linear mapping with a correlation between true and predicted activations on a held-out test set. The results show that the alignment scores between each layer of the LLMs and the brain increase around 0.4s post-word onset (Fig. 1A-C).

Temporal alignment. We next examine the existence of brain-like dynamics of computations in the nine LLMs specified above. On average, these models show a “temporal score” of $r = 0.99$ ($p < 1e-06$), between the depth of their layers and the MEG responses to words (Fig. 1A-B). This temporal alignment is observed in all studied models, even non-transformer models like RecurrentGemma-9B and Mamba-1.4B. (Fig. 2C-D). Oldest model GPT2-XL exhibits the lowest Temporal score, though significant, $r = 0.85$. While being grounded in the same seminal transformer design, GPT2-XL is smaller and lacks modern advancements present in the other studied transformers [30, 31]. When not pretrained, the LLMs do not show this alignment, and encode very poorly the brain activity. Together, these results suggest that the order of computations activated through recent LLMs’ layers is similar to the order of computations of the human brain listening to natural speech.

Impact of causality. To assess the impact of contextual directionality, we compare two bidirectional LLMs - BERT [32] and RoBERTa [33] - and one bidirectional speech model - Wav2vec2.0 [34] - to the previous causal models, more faithful to the brain’s causal mechanisms of language processing. While alignment scores are comparable, the temporal scores of these bidirectional models are substantially lower than those of causal LLMs (see App. B).

Impact of model size. Various architectures, scales and training choices all yield above-chance representational and temporal alignments. To identify the factors that impact the emergence of this phenomenon, we repeat these analyses on a family of LLMs that solely vary in model size. For this,

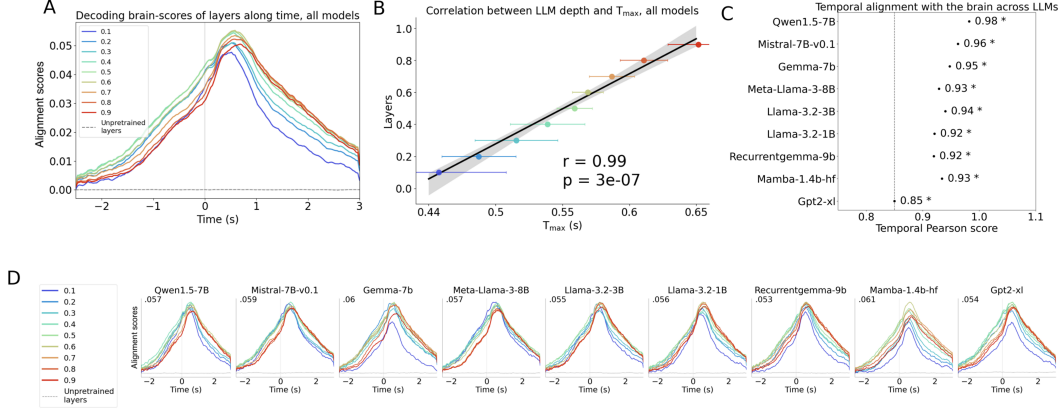


Figure 2: Human brain and LLMs exhibit temporal alignment. Correlation between time of peaking alignment scores (T_{max} , x-axis) and layer depth shows a highly significant temporal alignment. A. Alignment scores of 9 representative layers across each of the 9 studied LLMs, as a function of word-onset ($t=0$). Alignment scores have been averaged across models. In dashed gray curves, layers from unpretrained versions of these models, averaged over models. B. The time steps of peaking alignment scores (T_{max} , x-axis) are plotted for each representative layer (y-axis), averaged across models. The Temporal score r and associated p are printed on the figure. The grey area indicates the confidence intervals of the regression estimate. Error bars across subjects could not be computed due to the low number of subjects and the need to average across subjects to denoise neural data, though we ensure reproducibility of results across subjects in App. G,C. Here, colored error bars indicate standard deviations of the layer-wise distributions of T_{max} across the 9 presented models. C. Temporal scores are computed and presented for each model studied independently. An asterisk next to the Temporal score indicates the score is significant with $p < 5e-3$. D. Alignment scores of 9 representative layers across each of the 9 presented LLMs, as a function of word-onset ($t=0$). Each figure presents one model studied independently.

we leverage the Pythia family [27]: eight models of increasing scales. These models are trained in the same way, on the same data, in a highly controlled setup where only model size varies. We find that both the representational and the temporal alignment increase with model size (Fig. 3), from a non-significant temporal score ($r = 0.44$, $p > 0.05$) for the smallest model of 14M parameters, to a highly significant temporal score ($r = 0.96$, $p < 1e-4$) for the biggest model of 12 billion parameters. The correlation between the temporal score and the log model size reaches $r = 0.87$ ($p = 0.01$). This emergence follows a logarithmic trend, where the temporal and alignment scores of the biggest models tend to plateau. A similar trend is observed for alignment score.

Impact of context size. When listening and processing language, humans accumulate narrative context in the form of evidence to extract the richest meaning out of the currently heard word, and anticipate the next one [35]. Motivated by this incremental nature of human language processing, we postulate that brain-like inference dynamics in LLMs emerge with context size. To test how context size impacts representational and temporal alignments, we repeat our analyses on a single model (Llama-3.2 3B) while varying the amount of words in the input context. The results show that representational and temporal alignments increase with context size ($r = 0.81$, $p < 5e-2$, Fig. 4), from a non-significant temporal score without context ($r = 0.19$, $p > 0.5$) to a highly significant temporal score for context of 1000 words ($r = 0.93$, $p = 3e-4$). This temporal score increases logarithmically and considerably slows down from context lengths of 50 words. A similar - though lower - correlation is observed for alignment score. We find similar logarithmic increases of both temporal and alignment scores for state-space model Mamba (see in App. E).

Impact of word predictability. Autoregressive LLMs are trained to predict incoming words/tokens. In a similar way, predictive coding theory suggests that the brain anticipates upcoming words [36]. Could this similarity be the driving factor of temporal alignment? To test this possibility, we evaluate how temporal alignment varies with word’s predictability: if this hypothesis was true,

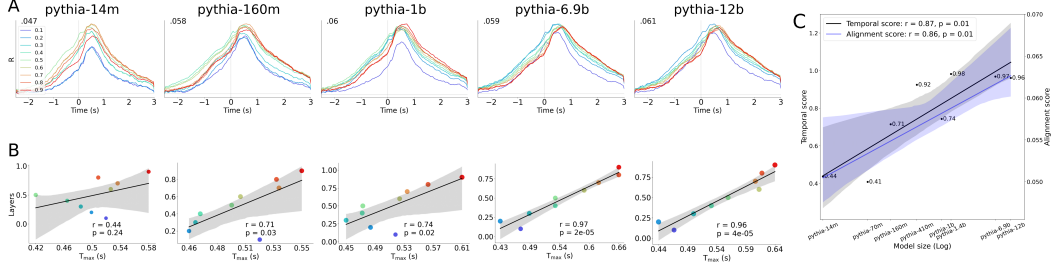


Figure 3: Temporal alignment emerges with model size. Colors indicate layer depth. A. Each of the 5 figures on the horizontal axis presents results for a specific model belonging to the Pythia family and studied independently, of size 14m, 160m, 1b, 6.9b and 12b parameters respectively (left to right). The Pythia family hosts 8 models of increasing size, all trained with the same data amount and parameter choices. Figures present evolution of alignment scores R_{layer} of 9 representative layers, from 10% to 90% of model depth, as a function of word-onset ($t=0$). B. Each of the 5 figures on the horizontal axis presents results for a specific models belonging to the Pythia family and studied independently. The time steps of peaking alignment scores (T_{max} , x-axis) are plotted for each representative layer (y-axis). The Temporal score r and associated p are printed on the figure. The grey area indicates the confidence intervals of the regression estimate. C. Temporal and alignment scores as functions of model size, for the 8 models forming the Pythia family. The model names (x-axis) are displayed on a logarithmic scale corresponding to their respective size. The Pearson scores r and associated p quantifying these correlations are printed on the figure. The grey and blue areas indicate the confidence intervals of the regression estimates.

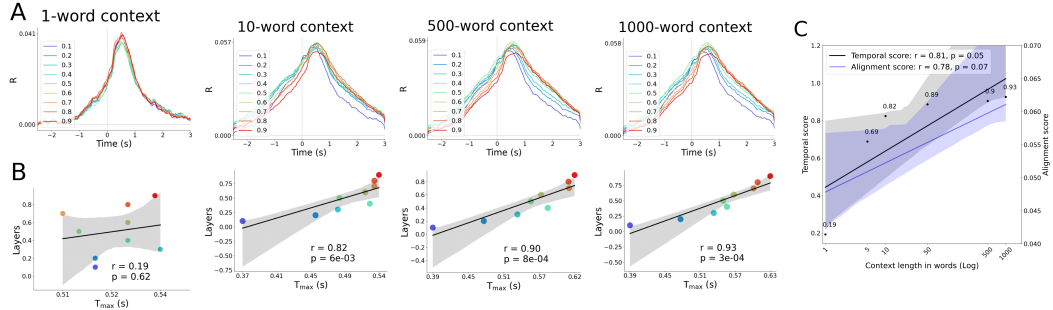


Figure 4: Temporal alignment increases with the length of the context provided to the LLM. Colors indicate layer depth. A. Each of the four figures on the horizontal axis presents results for a specific context length provided to Llama-3.2 3B, 1-word, 10-word, 500-word and 1000-word contexts respectively (from left to right). Figures present evolution of alignment scores R_{layer} of 9 representative layers, from 10% to 90% of model depth, as a function of word-onset ($t=0$). B. Each of the four figures on the horizontal axis presents results for a specific context length provided to Llama-3.2 3B. The time steps of peaking alignment scores (T_{max} , x-axis) are plotted for each representative layer (y-axis). The Temporal score r and associated p are printed on the figure. The grey area indicates the confidence intervals of the regression estimate. C. Temporal and alignment scores as functions of context length when given to Llama-3.2 3B, for six context lengths (x-axis). Context lengths are displayed on a logarithmic scale. The Pearson scores r and associated p quantifying these correlations are printed on the figure. The grey and blue areas indicate the confidence intervals of the regression estimates.

highly unpredictable words should exhibit low temporal alignment. To control for the fact that predictability can be impacted by the word being a content or function word, for this analysis we include all words of our dataset. For this, we first retrieve the predictability of each word in its context from the softmax-transformed logits of Llama-3-8B. We then separate these words into four

predictability quartiles. Finally, we evaluate temporal alignment for each quartile independently. The results show that most and least predictable words both lead to above chance representational and temporal alignments: $r = 0.92$, $p < 1e-3$ for the quartile "most expected" and $r = 0.83$, $p < 1e-2$ for the quartile "most surprising", respectively (Fig. 5). The temporal score for the most surprising quartile does exhibit a lower r value and p . However, when computing the Pearson correlation of the layer-wise differences of $T_{\max}$ between "most expected" and "most surprising" quartiles, we do not find a significant impact of contextual predictability on Temporal score ($p = 0.61$, Fig. 5E). This result holds with more layers too (see App. H). Overall, this control analysis indicates that word-predictability alone does not explain the emergence of temporal alignments between LLMs and the brain.

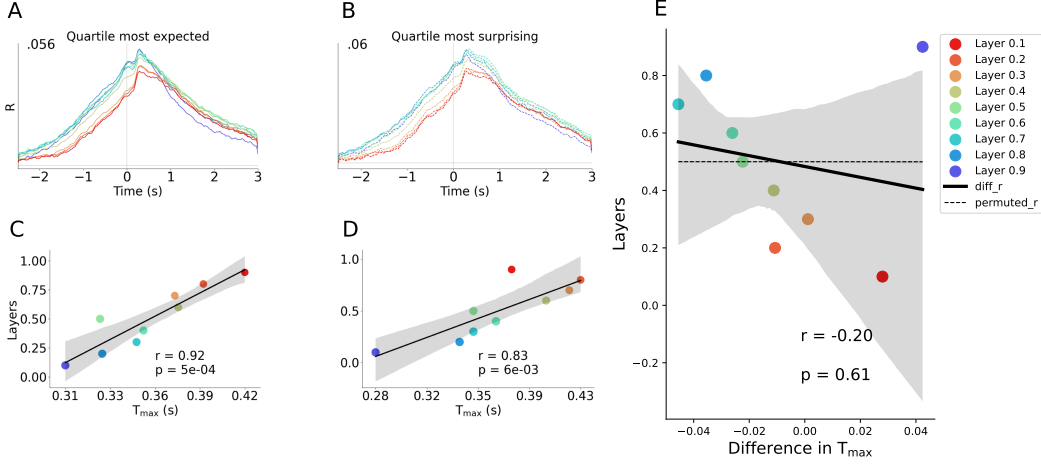


Figure 5: Temporal alignment holds independently of word predictability. Colors indicate layer depth. A. Alignment scores of 9 representative layers of Llama-3-8B, from 10% to 90% of layer depth, as a function of word-onset ($t=0$). Alignment score dynamic curves resulting from evaluating only the quartile of most expected words (from context) among the $\sim 270\,000$ words forming the dataset. The contextual predictability of words is computed through Llama-3-8B. B. The time steps of peaking alignment scores ($T_{\max}$, x-axis) of the quartile of most expected words are plotted for each representative layer (y-axis) of Llama-3-8B. The Temporal score r and associated p are printed on the figure. The grey area indicates the confidence intervals of the regression estimate. C. Alignment scores of 9 representative layers of Llama-3-8B, from 10% to 90% of layer depth, as a function of word-onset ($t=0$). Alignment score dynamic curves resulting from evaluating only the quartile of least expected (i.e. more surprising) words from context, among the $\sim 270\,000$ words forming the dataset. The contextual predictability of words is computed through Llama-3-8B. D. The time steps of peaking alignment scores ($T_{\max}$, x-axis) of the quartile of least expected words are plotted for each representative layer (y-axis) of Llama-3-8B. The Temporal score r and associated p are printed on the figure. The grey area indicates the confidence intervals of the regression estimate. E. The pairwise differences between time steps of peaking alignment scores (Difference in $T_{\max}$, x-axis) per layer, between the quartile of most expected words and the quartile of least expected words, for each representative layer (y-axis) of Llama-3-8B. The Pearson score r and associated p quantifying this correlation are printed on the figure. The grey area indicates the confidence intervals of the regression estimate.

Correlation between temporal and alignment scores. We investigate whether, and why, temporal and alignment scores are correlated, while arising from different measurements. We study jointly the temporal scores across all previously studied models - clustered by families - and context lengths, with their respective maximal alignment score - i.e. through their best predictive layer. When plotting both alignment and temporal scores across families of models (Fig. 6A), we find that both alignment and temporal scores increase with model size and context length. When correlating both alignment and temporal scores, we find that temporal score indeed correlates significantly with alignment score (Pearson score = 0.54, $p = 9e-04$) (Fig. 6B). This result indicates that the capacity of an

LLM to predict neural signals of the human brain is correlated to how tightly aligned its pathway of computations is with the one at play in the brain, when processing language.

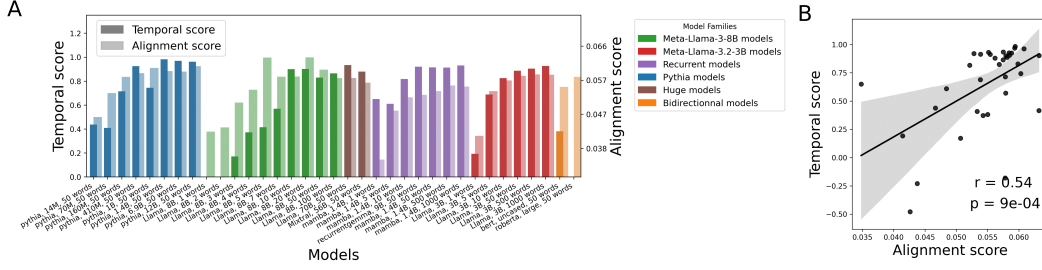


Figure 6: **Temporal and alignment scores are significantly correlated.** A. Temporal score and alignment score evolve similarly across models and context lengths, along the x-axis, clustered by families of models. Models are presented in the format [model name, model size, context length]. B. Temporal score (y-axis) and alignment score (x-axis) are significantly correlated over all models and variance in contexts size presented in this research. The Pearson score r and associated p quantifying this correlation are printed on the figure. The grey area indicates the confidence interval of the regression estimate.

4 Discussion

Temporal alignment. This work investigates whether and when LLMs generate representations in an order similar to the human brain’s during natural speech listening. This study provides three main contributions. First, the order of representations generated by LLMs’ layers strongly correlates with the sequence of neural activations observed in the human brain recorded with MEG. This *temporal* alignment complements previous work on the *anatomical* alignment observed between LLMs’ layers and functional regions of the human brain [5, 8, 3, 9, 1], and systematize early report of temporal alignment between language models and the brain [4, 6].

A shared computational path. We show that temporal scores is independent of the word predictability (Fig. 5). Beyond indicating the autoregressive generation capacities of the LLM, the temporal score thus seems to indicate its brain-like inference dynamics. Additionally, as previously reported [2, 4, 3, 37], the best alignment scores are achieved in the intermediary layers. This result suggests that the intermediary representations (layer=0.6 of architecture) – as opposed to the input and the prediction – are best aligned between brains and LLMs. Finally, while part of the alignment score can be explained by the latent representations present from the very first layer — accounting for the relatively high alignment score observed from the first layers of the LLM — the temporal score accounts for evolving representations, a sequentially ordered mechanism of processing. Together, these results indicate that the brain and LLMs don’t just share similar representations, but a similar computational path too. The increase of alignment score in more brain-like models may stem from a shared sequential structure of computations between the LLMs and the brain, as captured by the temporal score.

Impact of model architecture. Second, this LLM-brain alignment is observed in non-transformer architectures, such as state-space models. While other architectures (e.g. fully connected networks [38], Kolmogorov-Arnold Networks [39]) remain to be investigated, this finding suggests that the convergence between LLMs and the brain is architecture-independent. If confirmed, this result provides additional evidence that alignment score does not result from a trivial inductive bias. It remains to be investigated, however, whether it relates to model objective (next token prediction) or to the structure of language [40].

Impact of context and model size. Third, our experiments show that this alignment directly depends on (i) context size and (ii) model size – although with a saturation beyond 70B parameters models (App. D). These MEG results extend previous works on scaling laws in neuro-ai [37, 41, 42].

In particular, [37, 43] showed that LLMs that best predict functional magnetic resonance imaging (fMRI) responses to natural speech are those with the largest amount of parameters. In parallel, [44, 45] showed that context size improved the alignment between brain and fMRI. Here, we further show the effect of context size and model size to increase logarithmically, hence pointing to diminishing returns, if not a plateau. Extending our analyses to larger LLMs, e.g. Llama 3.3 70B, does not yield major improvement as compared to smaller LLMs e.g. Llama 3.2 3B, hence pointing to a plateau effect of scaling laws identifiable with MEG. Together, these findings clarify the specific conditions required for brain-like representations and computations to emerge. It remains unclear, however, whether context and size act directly on the alignment, or are confounded by other uncontrolled variables, such as linguistic performance. For example, in App. I, we find that performance is correlated with temporal alignment for specific conditions - LLM with increasing context lengths - but not others - LLMs with increasing sizes. Disentangling the causal chain that links these factors remains a major research avenue.

Working memory. Interestingly, the comparison between State Space Models (SSMs) and LLMs offers a new perspective to investigate context-size. Indeed, transformers compute contextual representations, thanks to a non linear combination of the past context, and hence require very large memory buffers that are implausible in the brain. By contrast, SSMs can be thought of as recurrent neural networks (RNNs) with hidden states that linearly evolve over time. At each time point, they thus represent the full context with their hidden state – an approach presumably similar to the brain. Our results show that SSMs do exhibit modest yet consistent improvements in temporal scores even at very large context sizes (App. E), and thus show that it is, in fact possible, to build and maintain a long context in a single hidden state. However, further research remains necessary to evaluate the degradation of such memory in the absence of meaningful context, such as during the memorization of a random digit sequence like a phone number – a task recognized as highly constrained in human subjects [46].

Limitations. Three main limitations should be highlighted. First, MEG has a limited spatial resolution of brain activity. This recording device does not allow a single-neuron recording, and is largely unable to pick up deep sources. Consequently, the precise neuronal bases of the present findings remain to be further explored with intracranial recordings. Second, the present dataset only consists of 3 subjects. This design choice was motivated by the fact that subjects each listen to 10 h of audiobooks, making this dataset the largest per-subject MEG listening dataset authors found, among multi-subject datasets. As alignment models are trained subject-wise, the biggest possible per-subject amount of data allows the most precise brain-alignment with LLMs dynamics. While App. G shows similar findings can be identified within each of these three subjects, how these alignments *vary across individuals* thus remains unknown. Finally, the present work only focuses on pretrained text-models. Yet, humans necessarily process language through sensory modality. Consequently, future work remains necessary to investigate the similarities between brains and speech models [47, 48, 5, 9].

Impact. The present findings reveal striking similarities between human brain responses and sequential representations in multiple kinds of large language models (LLMs). Yet, their architecture, sensory modality, learning objective, and training regime are remarkably disjoint [49, 50, 51, 52]. Furthermore, specific features like syntactic structures and semantic roles may be represented fundamentally differently [53, 54]. That partially converging computational paths arise despite these differences is thus all the more surprising, and highlights the necessity to clarify the computational principles that lead language processing to be partially shared between biological and artificial systems. We hope that systematically investigating the factors that steer LLMs to function similarly to the brain will help reduce this major gap, and ultimately chart a path toward building artificial systems that learn as efficiently as the human brain.

5 Acknowledgments

We warmly thank the authors of [10] for building the MEG dataset on which we relied for this study.

We wish to thank every member of Meta’s legal team for helping us to process quickly through the legal process in order to submit in time to the Neurips 2025 conference.

References

- [1] Aditya R Vaidya, Shailee Jain, and Alexander G Huth. Self-supervised models of audio effectively explain human cortical responses to speech, 2022.
- [2] Shailee Jain, Vy Vo, Shivangi Mahto, Amanda LeBel, Javier S. Turek, and Alexander G. Huth. Interpretable multi-timescale models for predicting fmri responses to continuous natural speech. In *Advances in Neural Information Processing Systems*, volume 33, pages 13738–13749, 2020.
- [3] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain), 2019.
- [4] Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134, 2022.
- [5] Juliette Millet, Charlotte Caucheteux, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, Jean-Remi King, et al. Toward a realistic model of speech processing in the brain with self-supervised learning, 2022.
- [6] Ariel Goldstein, Eric Ham, Samuel A Nastase, Zaid Zada, Avigail Grinstein-Dabus, Bobbi Aubrey, Mariano Schain, Harshvardhan Gazula, Amir Feder, Werner Doyle, et al. Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain. *BioRxiv*, pages 2022–07, 2022.
- [7] Neil Shubin, Cliff Tabin, and Sean Carroll. Fossils, genes and the evolution of animal limbs. *Nature*, 388(6643):639–648, 1997.
- [8] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441, 2023.
- [9] Yuaning Li, Gopala K Anumanchipalli, Abdelrahman Mohamed, Peili Chen, Laurel H Carney, Junfeng Lu, Jinsong Wu, and Edward F Chang. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, 26(12):2213–2225, 2023.
- [10] Kristijan Armeni, Umut Güçlü, Marcel van Gerven, and Jan-Mathijs Schoffelen. A 10-hour within-participant magnetoencephalography narrative dataset to test models of language comprehension. *Scientific Data*, 9(1):278, 2022.
- [11] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.
- [12] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.
- [13] Jean-Rémi King and Stanislas Dehaene. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences*, 18(4):203–210, 2014.
- [14] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, 7:267, 2013.
- [15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024.
- [16] Meta AI. Llama 3: Open foundation and instruction models. <https://huggingface.co/meta-llama/Meta-Llama-3-8B>, 2024.
- [17] Meta AI. Llama 3.2 1b. <https://huggingface.co/meta-llama/Llama-3.2-1B>, 2024.
- [18] Meta AI. Llama 3.2 3b. <https://huggingface.co/meta-llama/Llama-3.2-3B>, 2024.
- [19] Mistral AI. Mistral 7b. <https://huggingface.co/mistralai/Mistral-7B-v0.1>, 2023.

- [20] 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, 2024.
- [21] Alibaba Cloud. Qwen1.5 7b. <https://huggingface.co/Qwen/Qwen1.5-7B>, 2024.
- [22] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners, 2019.
- [23] OpenAI. Gpt-2 xl. <https://huggingface.co/openai-community/gpt2-xl>, 2019.
- [24] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024.
- [25] Stateful AI. Mamba-1.4b-hf. <https://huggingface.co/state-spaces/mamba-1.4b-hf>, 2024.
- [26] Aleksandar Botev Griffin Team, Soham De, Samuel L Smith, Anushan Fernando, George-Christian Muraru, Ruba Haroun, and Leonard Berrada et al. Recurrentgemma 2b. <https://huggingface.co/google/recurrentgemma-2b>, 2024.
- [27] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling, 2023.
- [28] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- [29] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 1–7. Association for Computational Linguistics, 2020.
- [30] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- [31] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.
- [32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- [33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.
- [34] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.
- [35] Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P De Lange. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119, 2022.
- [36] Karl Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836, 2005.
- [37] Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36:21895–21907, 2023.
- [38] Alexander G. Schwing and Raquel Urtasun. Fully connected deep structured networks. *CoRR*, abs/1503.02351, 2015.

- [39] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.
- [40] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.
- [41] Stéphane d’Ascoli, Corentin Bel, Jérémy Rapin, Hubert Banville, Yohann Benchetrit, Christophe Pallier, and Jean-Rémi King. Decoding individual words from non-invasive brain recordings across 723 participants. *arXiv preprint arXiv:2412.17829*, 2024.
- [42] Hubert Banville, Yohann Benchetrit, Stéphane d’Ascoli, Jérémy Rapin, and Jean-Rémi King. Scaling laws for decoding images from brain activity. *arXiv preprint arXiv:2501.15322*, 2025.
- [43] Laurent Bonnasse-Gahot and Christophe Pallier. fmri predictors based on language models of increasing complexity recover brain left lateralization. *Advances in Neural Information Processing Systems*, 37:125231–125263, 2024.
- [44] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Deep language algorithms predict semantic comprehension from brain activity. *Scientific reports*, 12(1):16327, 2022.
- [45] Refael Tikochinski, Ariel Goldstein, Yoav Meiri, Uri Hasson, and Roi Reichart. Incremental accumulation of linguistic context in artificial and biological neural networks. *Nature communications*, 16(1):803, 2025.
- [46] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.
- [47] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021.
- [48] Alexei Baeviski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.
- [49] Linnea Evanson, Christine Bulteau, Mathilde Chipaux, Georg Dorfmueller, Sarah Ferrand-Sorbets, Emmanuel Raffo, Sarah Rosenberg, Pierre Bourdillon, and Jean-Rémi King. Emergence of language in the developing brain. Manuscript, May 2025. Equal contribution by Pierre Bourdillon and Jean-Rémi King.
- [50] Emmanuel Dupoux. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59, 2018.
- [51] Rachel R Romeo, Joshua Segaran, Julia A Leonard, Sydney T Robinson, Martin R West, Allyson P Mackey, Anastasia Yendiki, Meredith L Rowe, and John DE Gabrieli. Language exposure relates to structural neural connectivity in childhood. *Journal of Neuroscience*, 38(36):7870–7877, 2018.
- [52] Mohammad Amin Ghanizadeh and Mohammad Javad Dousti. Towards data-efficient language models: A child-inspired approach to language learning. *arXiv preprint arXiv:2503.04611*, 2025.
- [53] James Fodor, Carsten Murawski, and Shinsuke Suzuki. When word order matters: human brains represent sentence meaning differently from large language models. *bioRxiv*, pages 2025–07, 2025.
- [54] Daria Kryvosheieva and Roger Levy. Controlled evaluation of syntactic knowledge in multilingual language models. *arXiv preprint arXiv:2411.07474*, 2024.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract claims that activations in early-to-late layers of LLMs best align with early-to-late brain responses when processing a word. The abstract also claims that at least two factors allow this emergence in an LLM, size and context. The Result section of the paper show evidences for these claims. P-values are provided in all figures, either explicitly or noted in the form of asterixs, and ensure the significancy of these results.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The paper does discuss three limitations of the work performed by the authors in the section **Discussion**, paragraph **Limitations**.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper, in the Method section, as well as in figure captions and specific subsections for result-specific methods.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper is based on an open-source dataset cited in the Introduction section and in the Methods section. The paper cannot currently release the associated code as it is based on an internal codebase, however said codebase is planned to be released in short-term future.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The experimental setting, including training and test details, is presented in the core of the paper in Section Methods.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper does report statistical significance of every result in the form of P_values in each corresponding figure as well as in the main text of the paper. Error bars are computed when possible, defined in Methods and justified in Methods when not possible to compute (e.g. too few subjects). Confidence intervals are also computed in each figure where technically possible.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper provides sufficient information on the computer resources needed to reproduce the experiments in the Methods Section, in paragraph LLM activations and preprocessing.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The research conducted in this paper conforms to all points presented in the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: The scope of the paper being foundational research only, the authors cannot find any direct path to any negative applications of the results presented in the paper.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper does not release any new data or model.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All creators or original owners of assets used in the paper are cited in the main text, next to the mentioned asset.

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The present paper uses only one dataset regarding research with human subjects, and this dataset has already been published for three years, it is open-source and available to all, and cited in the paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The present paper uses only one dataset regarding research with human subjects, and this dataset has already been published for three years, it is open-source and available to all, and cited in the paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: An LLM was used only for editing, or formatting purposes, not used as a tool in the core methods of this research. LLMs were used in the scope of this research not as tools but as subjects, their dynamics were studied in comparison to human brain's dynamics during language processing as the scope of the present paper, but authors believe the field of research of the paper to not be the scope of this question.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Supplementary Material and Technical Appendices

A Encoding scores for two LLMs.

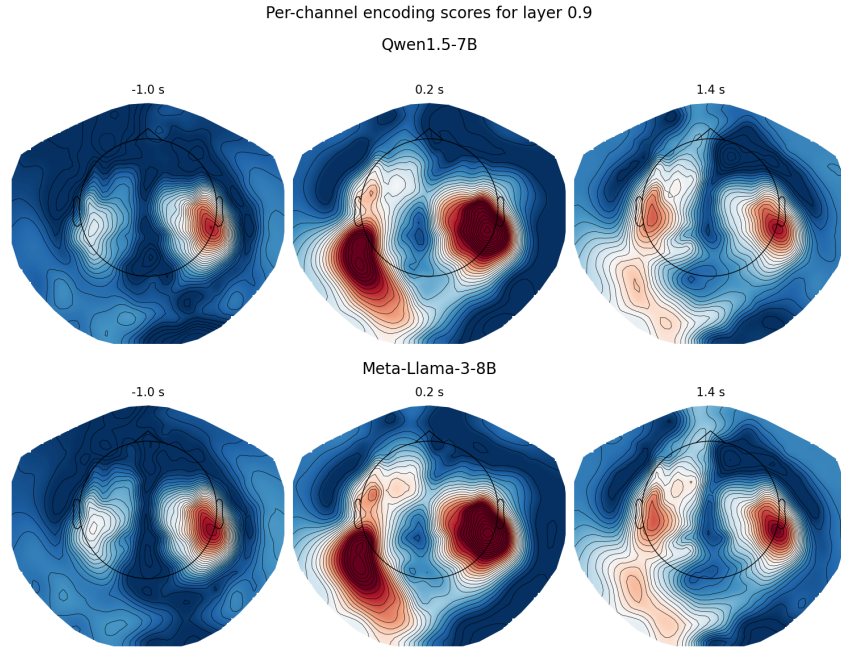


Figure 7: **Encoding scores for two LLMs.** Per-channel encoding scores for the layer $l=0.9$ of two LLMs, Qwen 1.5-7B (up) and LLama-3-8B (down), at several time points during word prediction pre word-onset (-1.0s, left), and processing post word onset (0.2s middle, and 1.4s right).

B Temporal and alignment scores of bidirectional models.

Below are presented results for bidirectional LLMs BERT and RoBERTa, alongside bidirectional speech model Wav2vec2. All three models show alignment scores comparable to larger, more recent and causal LLMs, but much lower, non-significant temporal scores.

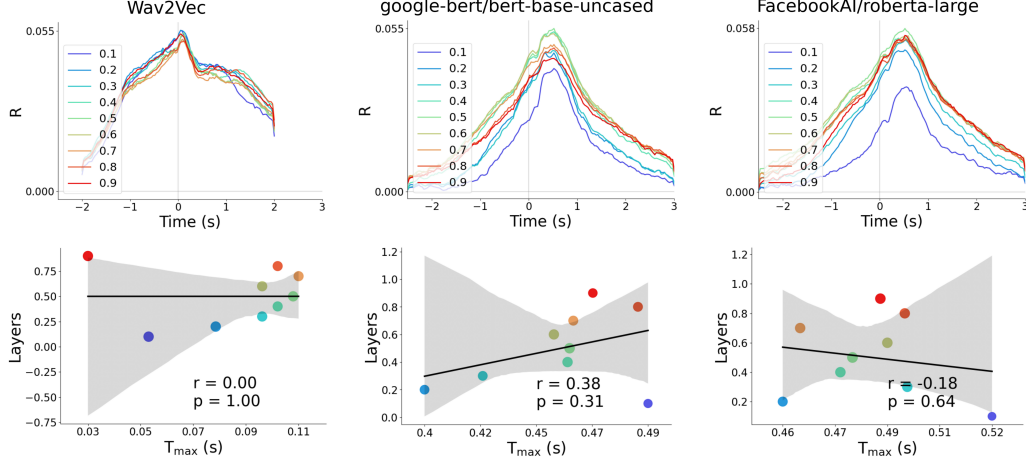


Figure 8: **Temporal and alignment scores of bidirectional models.** Each of the 3 figures presents results for a Wav2vec2, BERT and RoBERTa. Upper figures on the horizontal axis present evolution of alignment scores R_{layer} of 9 representative layers, from 10% to 90% of model depth, as a function of word-onset ($t=0$). Lower figures on the horizontal axis presents the time steps of peaking alignment scores (T_{max} , x-axis) plotted against each representative layer (y-axis). The Temporal score r and associated p are printed on the figure. The grey area indicates the confidence intervals of the regression estimate.

C Alignment scores across subjects, per LLM

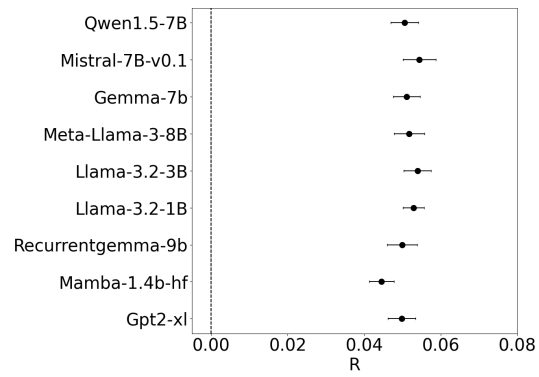


Figure 9: **Alignment scores across subjects, per LLM.** Maximum alignment scores per model, Each errorbar is a standard deviation across the three subjects of the dataset.

D Temporal and alignment scores of the largest models

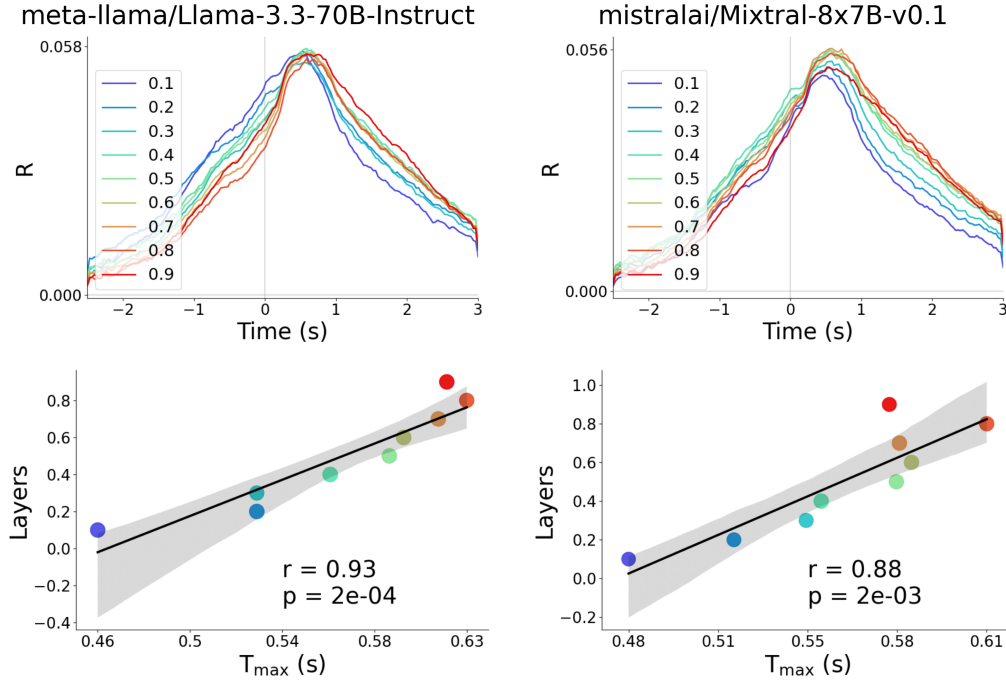


Figure 10: **Temporal and alignment scores of the largest models.** Each of the 2 figures presents results for a Llama3.3-70B-Instruct and Mixtral-8x7B-v0.1. Upper figures on the horizontal axis present evolution of alignment scores R_{layer} of 9 representative layers, from 10% to 90% of model depth, as a function of word-onset ($t=0$). Lower figures on the horizontal axis presents the time steps of peaking alignment scores (T_{max} , x-axis) plotted against each representative layer (y-axis). The Temporal score r and associated p are printed on the figure. The grey area indicates the confidence intervals of the regression estimate.

E Temporal alignment increases with the length of the context provided to State-Space Model Mamba-1.4B

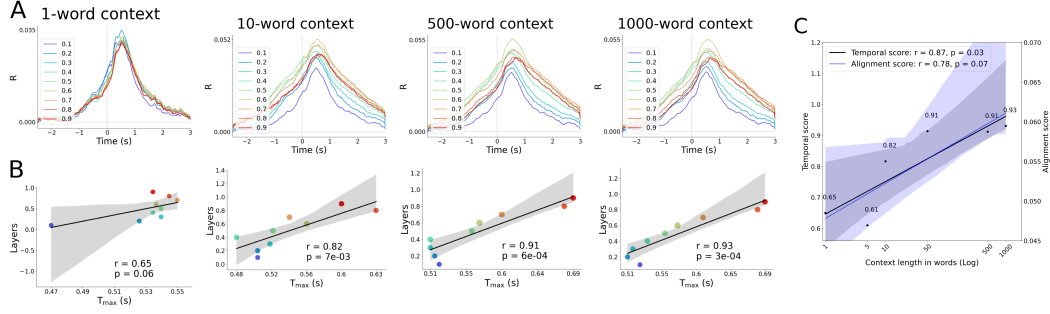


Figure 11: **Temporal alignment increases with the length of the context provided to State-Space Model Mamba-1.4B.** Colors indicate layer depth. A. Each of the four figures on the horizontal axis presents results for a specific context length provided to State-Space Model Mamba-1.4B, 1-word, 10-word, 500-word and 1000-word contexts respectively (from left to right). Figures present evolution of alignment scores R_{layer} of 9 representative layers, from 10% to 90% of model depth, as a function of word-onset ($t=0$). B. Each of the four figures on the horizontal axis presents results for a specific context length provided to Mamba-1.4B. The time steps of peaking alignment scores (T_{max} , x-axis) are plotted for each representative layer (y-axis). The Temporal score r and associated p are printed on the figure. The grey area indicates the confidence intervals of the regression estimate. C. Temporal and alignment scores as functions of context length when given to Mamba-1.4B, for six context lengths (x-axis). Context lengths are displayed on a logarithmic scale. The Pearson scores r and associated p quantifying these correlations are printed on the figure. The grey and blue areas indicate the confidence intervals of the regression estimates.

F Human brain and LLMs exhibit temporal alignment. Analysis performed for all types of words: function and content words

To ensure the processing of the most semantically meaningful words, we often study only content words in the main text of this paper (as opposed to function words), specifically those which belong to the following part-of-speech categories as defined by Spacy: NOUN, VERB, ADJ, ADV. Here, we test for replicability of findings for all words (function and content).

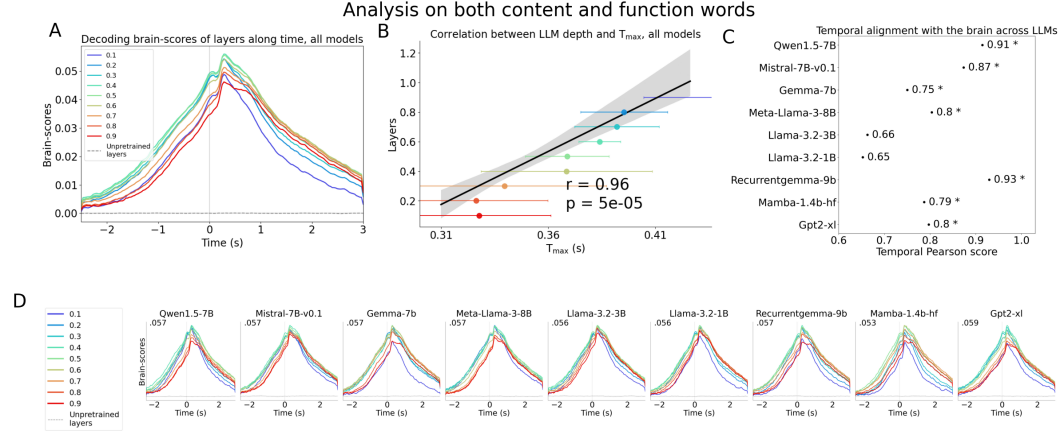


Figure 12: Human brain and LLMs exhibit temporal alignment. Analysis performed for all types of words: function and content words. Correlation between time of peaking alignment scores (T_{max} , x-axis) and layer depth shows a highly significant temporal alignment. **A.** Alignment scores of 9 representative layers across each of the 9 studied LLMs, as a function of word-onset ($t=0$). Alignment scores have been averaged across models. **B.** The time steps of peaking alignment scores (T_{max} , x-axis) are plotted for each representative layer (y-axis), averaged across models. The Temporal score r and associated P_{value} are printed on the figure. The grey area indicates the confidence intervals of the regression estimate. Here, colored error bars indicate standard deviations of the layer-wise distributions of T_{max} across the 9 presented models. **C.** Temporal scores are computed and presented for each model studied independently. An asterisk next to the Temporal score indicates the score is significant with $P_{value} < 5e-3$. **D.** Alignment scores of 9 representative layers across each of the 9 presented LLMs, as a function of word-onset ($t=0$). Each figure presents one model studied independently.

G Human brain and LLMs exhibit temporal alignment – Subject 1, 2 and 3 plotted individually

We ensure reproducibility of results presented in the main text of this paper – averaged across subjects – this time for individual subjects.

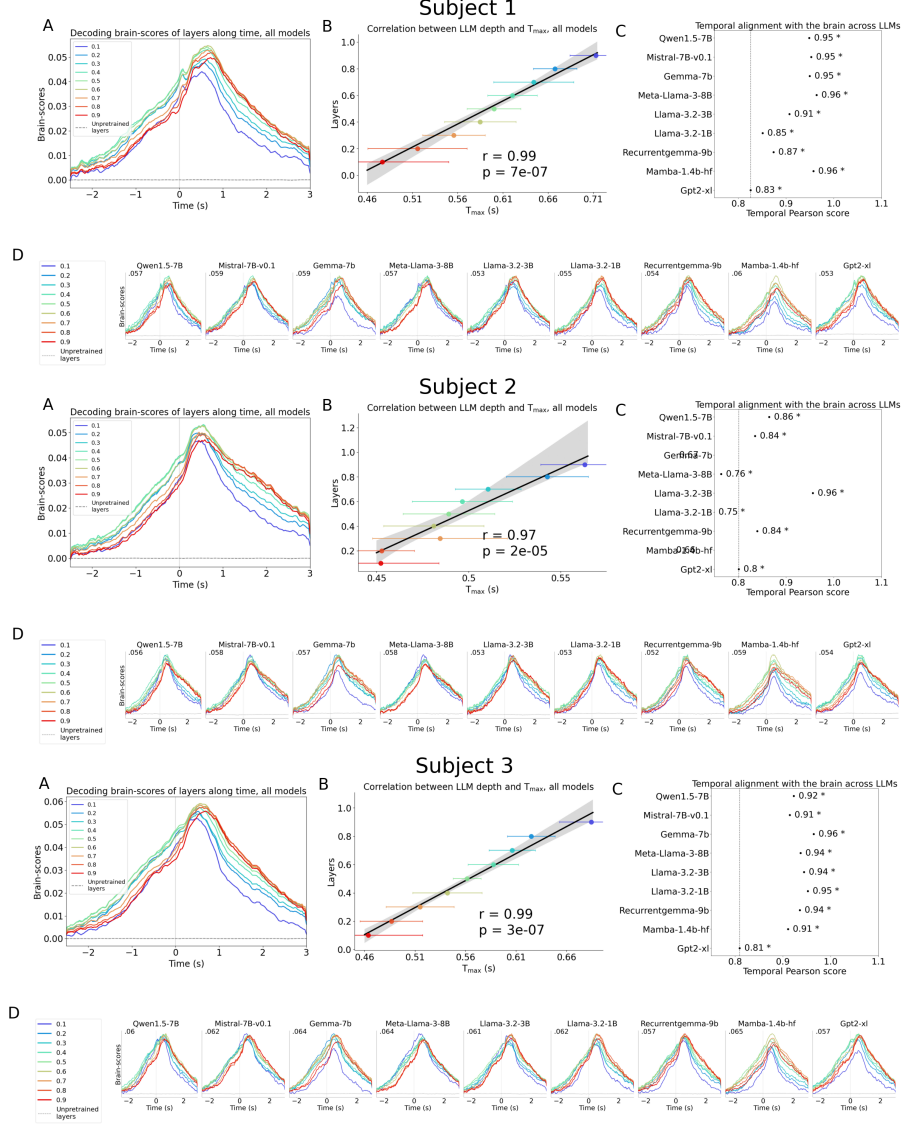


Figure 13: **Human brain and LLMs exhibit temporal alignment – Subject 1, 2 and 3 plotted individually**

Correlation between time of peaking alignment scores (T_{max} , x-axis) and layer depth shows a highly significant temporal alignment. A. Alignment scores of 9 representative layers across each of the 9 studied LLMs, as a function of word-onset (t=0). Alignment scores have been averaged across models. B. The time steps of peaking alignment scores (T_{max} , x-axis) are plotted for each representative layer (y-axis), averaged across models. The Temporal score r and associated P_{value} are printed on the figure. The grey area indicates the confidence intervals of the regression estimate. Here, colored error bars indicate standard deviations of the layer-wise distributions of T_{max} across the 9 presented models. C. Temporal scores are computed and presented for each model studied independently. An asterisk next to the Temporal score indicates the score is significant with $P_{value} < 5e-3$. D. Alignment scores of 9 representative layers across each of the 9 presented LLMs, as a function of word-onset (t=0). Each figure presents one model studied independently.

H Temporal alignment holds independently of word predictability

When computing the Pearson correlation of the layer-wise differences of $T_{\max}$ between "most expected" and "most surprising" quartiles, we do not find a significant impact of contextual predictability on Temporal score. To ensure this result, we perform here the same analysis on twice as many layers across the depth of Llama-3-8B architecture.

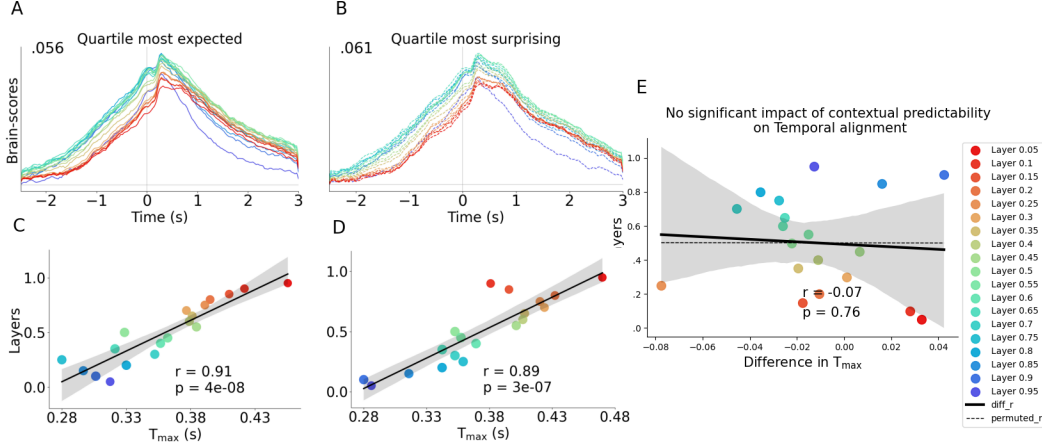


Figure 14: **Temporal alignment holds independently of word predictability.** Colors indicate layer depth. A. Alignment scores of 19 representative layers of Llama-3-8B, from 10% to 90% of layer depth, as a function of word-onset ($t=0$). Alignment score dynamic curves resulting from evaluating only the quartile of most expected words (from context) among the ~270 000 words forming the dataset. The contextual predictability of words is computed through Llama-3-8B. B. The time steps of peaking alignment scores ($T_{\max}$, x-axis) of the quartile of most expected words are plotted for each representative layer (y-axis) of Llama-3-8B. The Temporal score r and associated P_{value} are printed on the figure. The grey area indicates the confidence intervals of the regression estimate. C. Alignment scores of 19 representative layers of Llama-3-8B, from 10% to 90% of layer depth, as a function of word-onset ($t=0$). Alignment score dynamic curves resulting from evaluating only the quartile of least expected (i.e. more surprising) words from context, among the ~270 000 words forming the dataset. The contextual predictability of words is computed through Llama-3-8B. D. The time steps of peaking alignment scores ($T_{\max}$, x-axis) of the quartile of least expected words are plotted for each representative layer (y-axis) of Llama-3-8B. The Temporal score r and associated P_{value} are printed on the figure. The grey area indicates the confidence intervals of the regression estimate. E. The pairwise differences between time steps of peaking alignment scores (Difference in $T_{\max}$, x-axis) per layer, between the quartile of most expected words and the quartile of least expected words, for each representative layer (y-axis) of Llama-3-8B. The Pearson score r and associated P_{value} quantifying this correlation are printed on the figure. The grey area indicates the confidence intervals of the regression estimate.

I Temporal scores along top-3 accuracies evaluated on our dataset for next-token prediction across LLMs

We find that performance is correlated with temporal alignment for specific conditions – LLMs with increasing context lengths – but not others – LLMs with increasing sizes. Regression estimates illustrate how scaling model size or context affect temporal alignment with brain activity. Increasing context lengths significantly correlates with temporal alignment ($p = 0.01$). Increasing model sizes does not significantly correlate with temporal alignment.

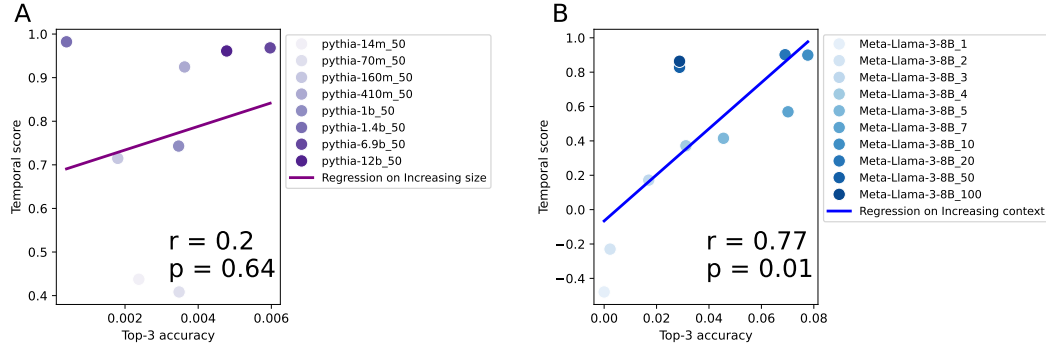


Figure 15: **Temporal scores along top-3 accuracies evaluated on our dataset for next-token prediction across LLMs.** A and B show the correlation between Temporal score (y-axis) and top-k accuracy (x-axis), across two subsets of models: A. Pythia models of increasing size, evaluated at top-k = 3. B. Meta-Llama-3-8B model with increasing context lengths, evaluated at top-k = 3.

J Licenses

The dataset used in this paper is licensed under a Creative Commons Attribution 4.0 International License.

Regarding the models used, here follows a list of their licenses:

- Qwen/Qwen1.5-7B: Tongyi Qianwen License
- mistralai/Mistral-7B-v0.1: Apache 2.0 License
- google/gemma-7b: Gemma License
- meta-llama/Meta-Llama-3-8B: Llama 3 Community License Agreement
- meta-llama/Llama-3.2-3B: Llama 3.2 Community License Agreement
- meta-llama/Llama-3.2-1B: Llama 3.2 Community License Agreement
- google/recurrentgemma-9b: Gemma License
- state-spaces/mamba-1.4b-hf: Apache 2.0 License
- openai-community/gpt2-xl: MIT License
- Pythia family of models: Apache 2.0 License