

Triple Phase Transitions: Understanding the Learning Dynamics of Large Language Models from a Neuroscience Perspective

Yuko Nakagi^{1,2*}, Keigo Tada^{1,2*}, Sota Yoshino^{1,2*}, Shinji Nishimoto^{1,2†}, Yu Takagi^{1,2,3‡}

¹Osaka University, Japan

²National Institute of Information and Communications Technology, Japan

³National Institute of Informatics, Japan

Correspondence: nishimoto.shinji.fbs@osaka-u.ac.jp, yutakagi322@gmail.com

Abstract

Large language models (LLMs) often exhibit abrupt emergent behavior, whereby new abilities arise at certain points during their training. This phenomenon, commonly referred to as a “phase transition”, remains poorly understood. In this study, we conduct an integrative analysis of such phase transitions by examining three interconnected perspectives: the similarity between LLMs and the human brain, the internal states of LLMs, and downstream task performance. We propose a novel interpretation for the learning dynamics of LLMs that vary in both training data and architecture, revealing that three phase transitions commonly emerge across these models during training: (1) alignment with the entire brain surges as LLMs begin adhering to task instructions (*Brain Alignment and Instruction Following*), (2) unexpectedly, LLMs diverge from the brain during a period in which downstream task accuracy temporarily stagnates (*Brain Detachment and Stagnation*), and (3) alignment with the brain reoccurs as LLMs become capable of solving the downstream tasks (*Brain Realignment and Consolidation*). These findings illuminate the underlying mechanisms of phase transitions in LLMs, while opening new avenues for interdisciplinary research bridging AI and neuroscience.

1 Introduction

Large language models (LLMs) often exhibit abrupt emergent behaviors, whereby new abilities arise from scaling the model size, the amount of training data, or the number of training steps (Wei et al., 2022a). This behavior has predominantly been discovered through evaluations of model outputs (e.g., downstream task performance), leading to numerous breakthroughs (Wei et al., 2022b; Caballero et al., 2023; Du et al., 2024). Recent efforts in mechanistic interpretability have begun to uncover the underlying internal changes that occur during training process, aiming to elucidate these emergent phenomena (Olsson et al., 2022; Chen et al., 2023). However, prior studies have largely examined these transitions in isolated perspectives, making it unclear how these transitions, along with other additional aspects, interact with each other.

In seeking to clarify the underlying principles of deep neural networks (DNNs), neuroscience research has yielded human-centered insights into DNNs through comparisons between their internal representations and human brain activity (Yamins et al., 2014; Güçlü & van Gerven, 2015). This approach has recently been extended to LLMs (Jain & Huth, 2018;

*Equal first author.

†Team lead.

‡Equal last author.

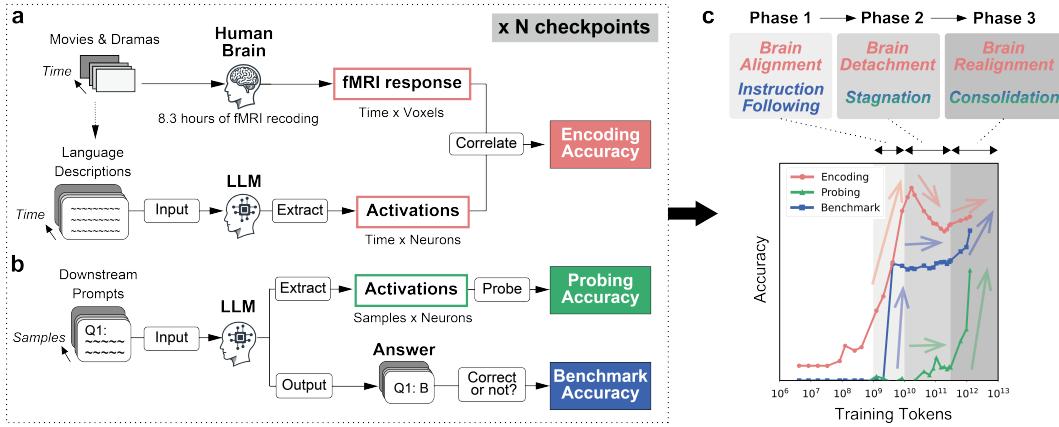


Figure 1: Overview of the study. **a** *Brain encoding* analysis. **b** *Probing* analysis (top) and *benchmark* analysis (bottom). **c** Three phase-transition phenomena during the learning process of LLMs, as identified through the results of the encoding, probing, and benchmark analyses. Red, green, and blue lines indicate encoding, probing, and benchmark accuracies, respectively, across all LLM checkpoints.

(Goldstein et al., 2022; Oota et al., 2022), revealing that brain activity has a stronger alignment with larger- and higher-performance models (Antonello et al., 2023). These findings highlight the potential of the human brain, an intricate system with diverse functions, to serve as a tool for model evaluation. Because most existing studies have only examined trained models, however, they offer limited insight into how emergent phenomena arise during the learning process.

In this study, we focus on the learning dynamics of LLMs in an attempt to provide a comprehensive understanding of emergent phenomena by integrating three analytical perspectives: *brain encoding* analysis, which assesses alignment with human brain activity; *probing* analysis, which detects shifts in internal representations; and *benchmark* analysis, which measures downstream task performance¹. We demonstrate that multiple LLMs, each characterized by distinct architectures and training data, exhibit a robust, common three-stage phase transition in their learning dynamics, including the precise timing of these transitions. Specifically, (1) alignment with the entire brain surges once the LLMs begin to follow downstream task instructions (*Brain Alignment and Instruction Following*); (2) surprisingly, the LLMs diverge from the brain during a period in which their downstream task accuracy temporarily stagnates (*Brain Detachment and Stagnation*); and (3) alignment with the brain reoccurs once LLMs become capable of solving the downstream tasks (*Brain Realignment and Consolidation*). Although prior work has reported stronger brain alignment with larger and higher performance models (Antonello et al., 2023), our findings reveal that the learning trajectory of LLMs is more complex than previously thought, consisting of multiple phases. By examining multiple LLMs trained on distinct language datasets, we highlight the influence of the training data on these learning dynamics. Taken together, our findings demonstrate that incorporating human brain activity as a biologically grounded benchmark reveals how the emergent capabilities of LLMs form and consolidate during training, offering essential insights for safer, more interpretable language models.

2 Related work

Phase transitions in LLMs LLMs acquire distinct abilities throughout their training process as their model size becomes larger, the amount of training data rises, and the number of training steps increases. While some abilities emerge gradually (Kaplan et al., 2020;

¹All experimental resources (source code, training data, and model checkpoints) will be publicly available to promote reproducibility and future research. See Sections A.1 and A.3 for the details.

Hoffmann et al., 2022), others appear abruptly (Ganguli et al., 2022; Wei et al., 2022a). Previous studies have identified “emergent abilities”, such as the onset of chain-of-thought prompting beyond a certain scale (Wei et al., 2022a). To understand LLMs internally, mechanistic interpretability seeks to unveil the internal computations learned by neural networks (Elhage et al., 2021; Dai et al., 2022). Internal phase transitions in the learning process have also recently been observed in this field (Olsson et al., 2022; Chen et al., 2023; Park et al., 2024), shedding light on how abrupt performance gains align with shifts in internal representations. For example, Olsson et al. (2022) found “phase changes” in Transformers, where induction heads suddenly arise and enable extended contexts to be handled through in-context learning. However, prior studies have largely focused on transitions in isolated perspectives, leaving it unclear how these shifts, alongside other relevant aspects, interact.

Interpretability from neuroscience insights From a neuroscience perspective, related work has compared the internal representations of DNNs with human brain activity (Yamins et al., 2014; Güçlü & van Gerven, 2015), demonstrating that the hierarchical structure of high-performing DNNs mirrors the hierarchical processing of the human visual cortex. More recent efforts have employed linear mappings from LLM internal representations to brain activity, probing language processing similarities between LLMs and the human brain (Jain & Huth, 2018; Schrimpf et al., 2021; Goldstein et al., 2022; Oota et al., 2022). Notable examples include comparisons of learning characteristics between LLMs and the human brain (Millet et al., 2022; Caucheteux & King, 2022; Caucheteux et al., 2023; Aw et al., 2023; Aw & Toneva, 2023; Antonello et al., 2023; Tuckute et al., 2024a,b; Antonello & Huth, 2024; de Varda et al., 2025) and discussions of the changes in latent dimensionality within LLMs (Cheng & Antonello, 2024).

3 Methods

3.1 Large language models

We analyze the learning dynamics of LLMs from three distinct perspectives. For this, we use four pre-trained models with available training checkpoints: OLMo-2 (OLMo et al., 2024), OLMo-0724 (Groeneveld et al., 2024), and LLM-jp (LLM-jp et al., 2024) for the main analysis, and Amber (Liu et al., 2023) for the control analysis. The number of checkpoints used for each model ranges from 18 to 28. Table 1 presents an overview of these LLMs. OLMo-2, OLMo-0724, and Amber are English-centric LLMs trained on publicly accessible datasets, while LLM-jp is a bilingual Japanese-English model trained on a roughly equal mix of Japanese and English. Each model uses a different training corpus and tokenizer, with vocabulary sizes ranging from 32,000 to 100,352. These LLMs have between 6.74–7.3B parameters, 32 hidden layers, and a hidden dimension of 4,096. They are all based on a decoder-only Transformer architecture (Vaswani et al., 2017), but each LLM incorporates a few critical modifications. There are also notable differences across these LLMs in terms of their layer normalization and attention mechanisms. See Section A.1 for further details on the models used. The multilayer perceptron (MLP) layers require most parameters and represent essential features (Bereska & Gavves, 2024; Geva et al., 2021). Thus, we use their neural activation as the internal representations of each model.

3.2 Brain encoding models

Our initial approach for analyzing the learning dynamics of the LLMs involves an investigation of how their activations progressively align with brain activity during the training process. Specifically, for each checkpoint described in Sections 3.1 and A.1, we perform an encoding analysis by evaluating the prediction accuracy of a learned linear mapping from each layer’s activations to brain activity (Naselaris et al., 2011; Nishimoto et al., 2011; Huth et al., 2012) (Figure 1a). These analyses are conducted separately for each participant.

fMRI datasets We use the Narrative Movie fMRI Dataset (Yamaguchi et al., 2024; Nakagi et al., 2024) for our brain encoding analysis. This dataset provides brain activity data from six healthy participants with normal or corrected-normal vision (three females; ages

22–40, mean = 28.7), while they freely watched 8.3 hours of movies or drama series inside a 3T functional Magnetic Resonance Imaging (fMRI) scanner. All participants were native Japanese speakers. The dataset includes nine video clips of movies or drama series as stimuli: eight international titles and one Japanese title. The international titles were dubbed into Japanese, allowing the participants to view all clips in Japanese. The dataset also contains three types of natural language annotation from the videos. We use the *Narrative Content* annotation, which describes the background story of the scene at 5-s intervals, for the main analysis and the *Objective Information* annotation, which describes the objects in the scene every second, for the control analysis, both in English and Japanese. See Section A.2 for additional dataset details. In this study, we use 29,993 seconds of data covering all six participants. We divide the data into training and test datasets. All prediction performance results are computed using the test dataset. Specifically, we use the fMRI scanning sessions corresponding to the last split of each movie or drama series (7,737 seconds in total) as the test data. The remainder of the sessions (22,262 seconds in total) forms the training data.

Model construction We extract the activations of the LLMs for the annotations from each hidden layer. They consist of several tokens for each time point. We then average the activations across tokens. Because multiple annotations exist for each second, we extract the activations for each annotator and then average them across all annotators. We then train an L2-regularized linear model to predict the voxel-level brain activity from the corresponding activations of the LLM neurons. We estimate the model weights from the training data, then apply them to the test data. The regularization parameters are determined for each voxel by cross-validation during training. We then evaluate the prediction accuracy by computing the Pearson’s correlation coefficients between the predicted and measured fMRI signals. Statistical significance is assessed using a blockwise permutation test that compares the correlations between predicted and measured signals against the correlations obtained after shuffling the measured signals. We set the threshold for statistical significance to $p < 0.05$ and correct for multiple comparisons using the FDR procedure. We model the hemodynamic delay in the BOLD signal, assumed to be 8–10 seconds. See Section A.3 for additional details about model construction and region-of-interest (ROI) analyses. In the main analysis, we focus on later layers within each LLM. Specifically, we focus on layer 25 in both OLMo-2 and LLM-jp, and layer 30 in OLMo-0724, which show the greatest checkpoint-wise changes and are strongly involved in each model’s emergent behavior across all evaluations (encoding, probing, and benchmark analyses; see Section A.6 for details). We confirm that the layers surrounding the main focal layer exhibit similar patterns of results (see Figure B.5).

3.3 Probing

As the second approach, we investigate how downstream task-relevant representations are progressively acquired within the LLMs during the training process. To this end, we perform a probing analysis by evaluating how well linear models can predict neuron activations from the answer labels of downstream tasks across all hidden layers (Figure 1b, top). These analyses are conducted separately for each LLM checkpoint.

Downstream datasets We use four downstream tasks: Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2020), CommonsenseQA (CSQA) (Talmor et al., 2019), AI2 Reasoning Challenge (ARC) (Clark et al., 2018), and HellaSwag (Zellers et al., 2019). For both the probing and benchmark analyses (see Section 3.4), we use 5-shot prompts in both English and Japanese. We exclude prompts exceeding each model’s maximum context length. See Section A.4 for additional details about downstream datasets and Section A.7 for examples of the prompts used in these datasets. In the probing analysis, each dataset is split into training and test datasets at a 4:1 ratio. All prediction performance results are computed using the test dataset.

Probing for MLP activations We first feed the downstream task prompts into the LLMs and extract the final-token neuron activations in each layer. For each sample in the downstream tasks, we create an answer matrix (the number of samples \times the number of choices), where the correct and incorrect choices are labeled as 1 and 0, respectively. Our objective is

to learn a mapping from this answer matrix to the activations across all layers of the LLMs. To this end, we use L2-regularized linear regression to estimate the linear weights that transform the answer matrix corresponding to the training data into the observed activations in the LLMs, and then apply these learned weights to the test data. The regularization parameter is optimized via 4-fold cross-validation for each neuron in the training dataset. We then evaluate the prediction accuracy by computing the Pearson’s correlation coefficients between the predicted and actual activations. The amount of information necessary for each task that is retained at each LLM checkpoint is quantitatively evaluated at the neuron level. In this study, we focus on the same layers as in the encoding analysis (Section 3.2).

3.4 Benchmark

As the third approach for analyzing the learning dynamics of LLMs, we investigate how the ability to solve downstream tasks emerges as the training process continues. Specifically, we use each LLM checkpoint to evaluate performance on the downstream task datasets from Section 3.3 (Figure 1b, bottom). The evaluation metric is the ratio of correctly answered items, i.e., the fraction of samples for which the model outputs exactly match the correct answers. We use the *llm-jp-eval* library² for this analysis.

3.5 Direct analysis of MLP activations

To investigate how internal representations in LLMs change over the course of training, we compare their properties across checkpoints using the correlation and dimensionality of activations. We compute the Pearson’s correlation coefficients of the activations across LLM checkpoints. Specifically, we use the same dataset as for the encoding analysis, extract the activations from the same layers, and calculate their correlations across different checkpoints. We also quantify the dimensionality of the activations at each checkpoint and examine changes across checkpoints. Here, we adopt the Intrinsic Dimension (ID), which has garnered attention as a means of characterizing the nature of LLM internal representations. We estimate the IDs at each checkpoint by applying the Generalized Ratios Intrinsic Dimension Estimator (GRIDE) (Denti et al., 2022) to each layer’s activations of the same dataset used for the encoding analysis (see Section A.5 for details).

4 Results

4.1 Learning dynamics of LLMs

We investigate how each accuracy metric evolves over the course of training using the three analytical approaches from Sections 3.2, 3.3, and 3.4. Figure 2 clearly shows that later layers of the LLMs have three phase transitions during training. The first phase transition emerges after around 10^9 – 10^{10} training tokens, where both the encoding and benchmark accuracy suddenly surge. In particular, the benchmark accuracy (blue line) reveals that the model’s ability to perform downstream tasks improves, suggesting that the LLMs begin following task instructions. Simultaneously, the brain encoding accuracy (red line) reveals enhanced overall alignment of the LLMs with the entire brain (see Section 4.2). We refer to this first phase as the *Brain Alignment and Instruction Following* Phase. The second phase transition arises after around 10^{10} – $3 \cdot 10^{11}$ training tokens, where the benchmark accuracy stagnates. Strikingly, in this phase transition, the brain encoding accuracy **declines**, indicating reduced alignment between the LLMs and the brain (see Section 4.2). Accordingly, we label this second phase as the *Brain Detachment and Stagnation* Phase. The third phase transition occurs beyond approximately $3 \cdot 10^{11}$ training tokens, where the benchmark and probing accuracy increase sharply, accompanied by a slight upward trend in the brain encoding accuracy. At this point, the increase in the benchmark accuracy suggests that the LLMs gradually acquire the capability to solve tasks, whereas the change in the brain encoding accuracy implies a renewed enhancement in alignment with the brain (see Section 4.2). We

²<https://github.com/llm-jp/llm-jp-eval>

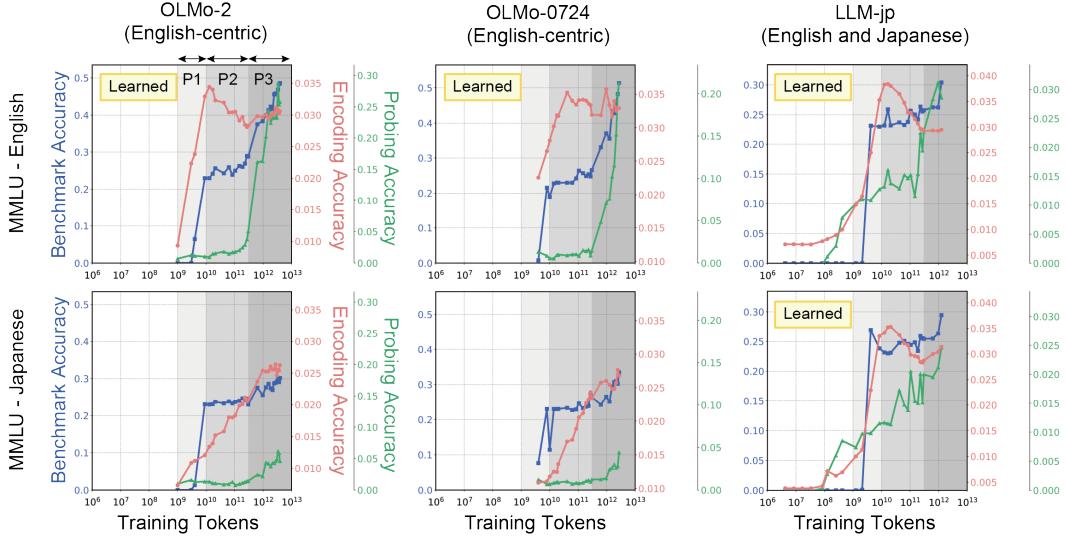


Figure 2: Learning dynamics of LLMs exhibiting three phase transitions. The horizontal axis denotes the number of training tokens. The vertical axis denotes the average encoding accuracy for all voxels of a single participant (DM06) (red lines), the benchmark accuracy (blue lines), and the average probing accuracy for all LLM neurons calculated using MMLU (green lines). We select layers 25, 30, and 25 from OLMo-2, OLMo-0724, and LLM-jp, respectively, to capture the transitions that occur at each phase of the learning dynamics. The background color indicates the LLM phase. The legend indicates whether the language has been learned sufficiently by the model. No checkpoints preceding the 10^9 training tokens have been made publicly available aside from LLM-jp.

thus refer to this third phase as the *Brain Realignment and Consolidation* Phase. Notably, these dynamics are only observed when the LLMs process a language that they have learned sufficiently: English and Japanese for LLM-jp, and English only for OLMo-2 and OLMo-0724.

Figure 2 presents the results for OLMo-2, OLMo-0724, and LLM-jp when using MMLU for the probing and benchmark analyses. The brain encoding result shows the average prediction accuracy across all brain voxels inferred from LLM neurons in the specified layer, whereas the probing result shows the average prediction accuracy across all LLM neurons. Section B.1 provides analogous results for adjacent layers, other downstream tasks (CSQA, ARC, HellaSwag), other LLMs (Amber), other languages (Chinese), and other annotations.

4.2 What happens during each phase?

Section 4.1 characterized the three phase transitions that arise during LLM training, along with interpretive insights drawn from encoding, probing, and benchmark analyses. Nevertheless, the exact internal changes that occur during each training phase remain unclear. To gain a deeper understanding and provide further interpretation, we examine how the relationship with the brain changes at the voxel level and how the internal representations in the LLMs shift at the neuron level in each phase.

Changes in the relationship with the brain To characterize the voxel-level relationship between LLMs and the brain across the three phases, we calculate the difference in brain encoding accuracy for each voxel between two representative training checkpoints that capture each phase transition, and project these changes onto each participant’s cerebral cortex. Figure 3 shows that, using the activations of OLMo-2, the voxel-level accuracy patterns shift distinctly from one phase to another. Specifically, in Phase 1—where the average prediction accuracy increases sharply—the accuracy gains are broadly distributed throughout the cerebral cortex, especially the temporal, occipital, and frontal cortex. By

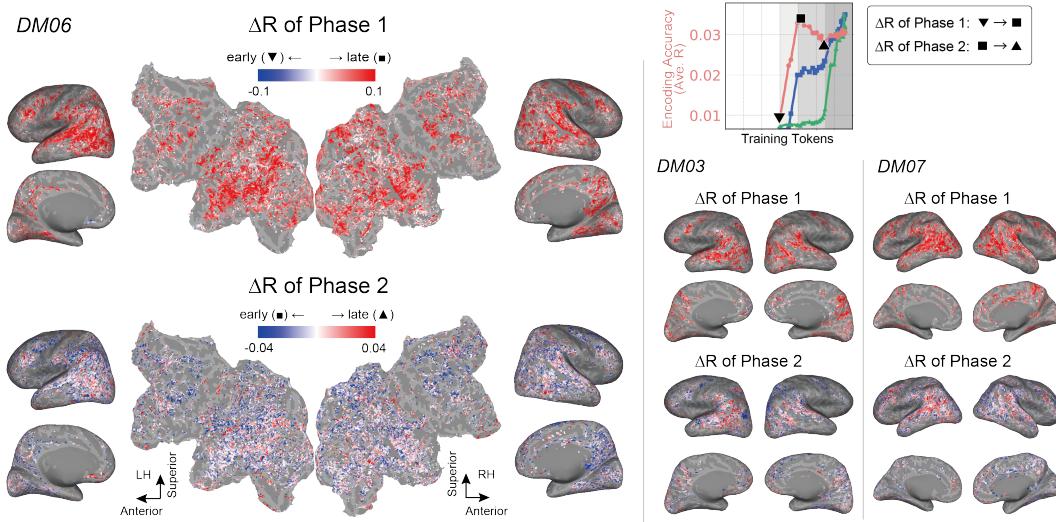


Figure 3: Changes in the relationship with the brain. Differences in encoding accuracy among checkpoints for three participants (DM06, DM03, and DM07), projected onto the inflated (top, lateral, and medial views) and flattened cortical surface (occipital areas are at the center, only for DM06), for both the left and right hemispheres. We have chosen OLMo-2’s checkpoints that capture the transitions at each phase of the learning dynamics. Brain regions with significant accuracy at either of the two checkpoints are colored ($p < 0.05$, FDR-corrected). Voxels exhibiting higher accuracy at the later checkpoint are indicated in red, whereas those exhibiting higher accuracy at the earlier checkpoint are indicated in blue.

contrast, in Phase 2—marked by a decrease in overall brain encoding accuracy—the changes tend to vary more on a voxel-by-voxel basis, with some voxels within the temporal cortex still showing improvement. In Phase 3, the accuracy rises again across broad brain regions (see Figure B.11). Furthermore, comparing the post-transition prediction accuracies in Phase 3 with those in Phase 1 reveals that, in certain voxels within the temporal cortex, the post-Phase 3 accuracies are higher than the post-Phase 1 accuracies (see Figure B.11).

Overall, these findings indicate that once the LLMs begin following the task instructions (Phase 1), their internal representations exhibit better global alignment with brain activity. During the subsequent stagnation phase (Phase 2), this global alignment partially recedes, and after Phase 3, the LLMs realign with the brain, reinforcing the similarity to brain regions involved in semantic processing, in particular. Notably, these patterns only emerge when the input language matches one on which the model has been trained. Further details for all participants, ROIs, other phases, and LLMs can be found in Figures B.11–B.16.

Evolution of LLM internal representations for downstream tasks We now investigate how the neuron-level internal representations evolve over the course of training. For this purpose, we examine changes in the distribution of probing accuracy for each neuron at the early, middle, and final checkpoints (302B, 1259B, 3896B training tokens in OLMo-2) of Phase 3. Figure 4a presents the results for neurons in layers 5, 15, 25, and 30 of OLMo-2 when using MMLU. From these results, we find that, in the English MMLU setting, later-layer neurons progressively increase in probing accuracy, indicating that, as the LLMs acquire downstream task proficiency, they also develop more specialized neuron activations. Figures B.17–B.19 show the results corresponding to all downstream tasks for OLMo-2, OLMo-0724, and LLM-jp, confirming that a similar trend emerges across every downstream task when each model is provided with the language on which it has been trained.

We can also explore whether different sets of neurons acquire specialized representations for distinct downstream tasks, focusing on the same checkpoints examined in Figure 4a. Figure 4b illustrates the relationships among the per-neuron probing accuracies of OLMo-2 for the English MMLU, CSQA, and ARC. Across all three tasks, it is evident that certain

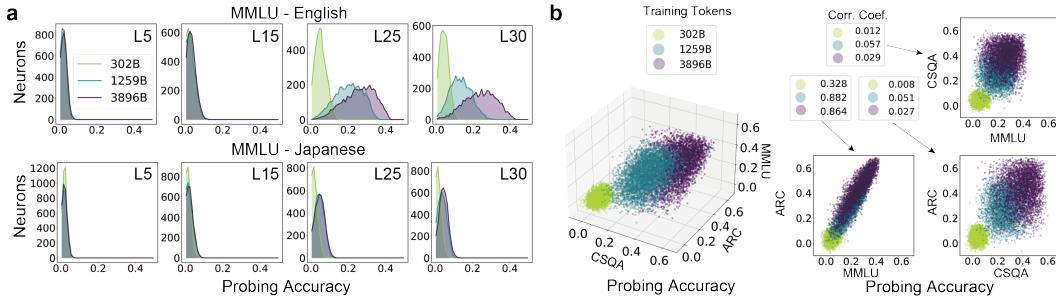


Figure 4: Probing results. **a** Evolution of probing accuracy for English/Japanese MMLU throughout the training process, assessed at layers 5, 15, 25, and 30 of OLMo-2. The horizontal axis indicates the probing accuracy. The vertical axis indicates the number of neurons that fall within each 0.01 accuracy bin. The legends corresponds to the number of training tokens. **b** Relationship between probing accuracy in the neurons of OLMo-2 (layer 25) across English MMLU, CSQA, and ARC. Each axis denotes the probing accuracy for the respective task, and the color gradient reflects the number of training tokens. The legend shows the correlation coefficient between certain two tasks.

neurons begin to manifest representations specialized for each downstream task as training progresses. Intriguingly, some neurons develop such specialized representations for all tasks, while others remain specific to only a subset of tasks. Moreover, the per-neuron accuracies for MMLU and ARC are highly correlated ($r = 0.864$), whereas those for CSQA demonstrate weak correlations with the other two tasks ($r = 0.029$ with MMLU, $r = 0.027$ with ARC). Furthermore, focusing on OLMo-0724—which, much like OLMo-2, demonstrates high benchmark accuracy on downstream task performance—we confirm that OLMo-0724 yields results that are analogous to those of OLMo-2 (Figure B.20). This implies that substantial transformations, specific to each task’s characteristics, are emerging within these models. The results for all downstream tasks, including HellaSwag, are discussed in Section B.3.

4.3 Changes in the nature of activations

Based on brain encoding and probing analyses, the preceding results have demonstrated that shifts in neuronal activation occur over the course of training. We now examine whether these alterations reflect fundamental shifts in the activations of the LLMs themselves.

As shown in Figure 5a, within-phase activations are highly similar, but substantial changes emerge at each phase transition. Interestingly, although the brain encoding accuracy does not vary dramatically between Phases 2 and 3, the underlying activations differ considerably. Next, we quantify the dimensionality of these activations using the IDs defined by GRIDE. Figure 5b shows that there is a strong similarity between the brain encoding accuracy and IDs during Phases 1 and 2, consistent with earlier findings that single-checkpoint encoding accuracy and IDs are highly correlated (Cheng & Antonello, 2024). Our findings extend this result by demonstrating a similarly robust correlation across multiple checkpoints. Nevertheless, the brain encoding accuracy and IDs are not perfectly congruent (see Phase 3), leaving open questions about the aspects of activation changes that intrinsic dimensionality alone does not capture.

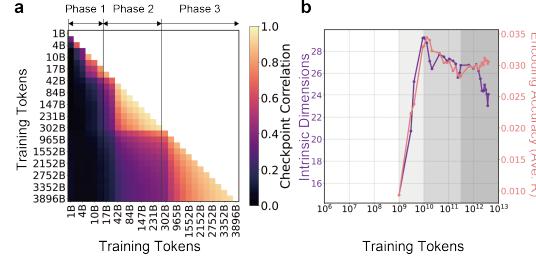


Figure 5: The Nature of Activations. **a** Variations in correlation coefficients of the activations of OLMo-2 across checkpoints. **b** IDs (purple line) and average encoding accuracy for all voxels of a single participant (DM06) (red line) across checkpoints.

5 Discussion and conclusions

In this study, we interpreted the learning dynamics of LLMs from three perspectives: alignment between LLMs and the brain, internal states associated with downstream tasks, and downstream task performance. Our novel approach elucidated a three-stage phase transition—including the precise timing of each transition—that emerges consistently during training, even among LLMs that diverge substantially in both training data and architecture.

The *Brain Alignment and Instruction Following* Phase and *Brain Realignment and Consolidation* Phase resemble the emergent phenomena reported by Wei et al. (2022a;b); Caballero et al. (2023); Olsson et al. (2022), whereby models abruptly gain novel abilities after passing a threshold. Prior work has largely focused on changes in outputs or the emergence of specific mechanisms. By integrating human brain alignment and shifts in LLM internal representations, we have shown that these factors progress in unison, highlighting multi-scale indicators behind abrupt gains and stagnations in model performance. Furthermore, the marked emergence of these phenomena in languages extensively represented within the model’s training corpus implies that deep internalization of their statistical patterns can trigger significant shifts in both brain alignment and overall performance. Hence, the presence or absence of these phase transitions may indicate whether an LLM has adequately assimilated and comprehended a language’s statistical structure.

Additionally, the learning dynamics that we have revealed are reminiscent of the “synaptic overproduction and pruning” posited in neuroscience, where an initial synaptic overgrowth eventually refines the neural circuits (Peter R., 1979; Bourgeois et al., 1994). These transient changes in the brain suggest the potential for more efficient neural information processing, leading to enhanced cognitive function. Thus, a promising avenue for future work involves examining whether the phenomenon observed in this study arises from the LLM pruning superfluous internal representations in pursuit of more refined ones.

This study has also revealed that voxel-level brain alignment does not simply increase over the training process, but fluctuates across distinct phases. Previous research linking model representations to brain activity (Yamins et al., 2014; Güclü & van Gerven, 2015; Schrimpf et al., 2021; Caucheteux et al., 2023) has not fully addressed such learning dynamics. While Antonello et al. (2023) reported a rise in alignment with increasing model size, our findings depict a more intricate path: an initial surge in brain alignment, a decline, and a final resurgence. This suggests that LLMs may adopt distinct computational strategies at various stages, rather than gradually acquiring brain-like language representations. Moreover, alignment with semantic regions in the temporal cortex generally rises over the course of training, implying a dynamic reorganization in how models correspond to brain activity.

By capturing shifts in LLM neuron activations related to downstream tasks, we elucidated how task-specific neurons form and how much specialization they share among tasks. These findings align with previous studies showing neurons that encode particular linguistic information (Tenney et al., 2019; Dai et al., 2022; Wang et al., 2022; Gurnee et al., 2023). We observed that neurons contributing substantially to certain tasks emerge abruptly in later layers as training progresses. Moreover, the variations in neuron sharing across tasks suggest that differences in activation patterns—reflecting task-specific characteristics (e.g., required capabilities, difficulty levels, and answer formats)—are indeed discernible at the neuronal level. Taken together, our results offer a more comprehensive view of LLM neurons during the learning process, building on past work that focused solely on trained models. Further in-depth analyses of these neurons would be a fascinating avenue for future research.

Finally, we extended the findings of Cheng & Antonello (2024), who demonstrated a positive correlation between the dimensionality of LLM activations in trained models and the alignment of those LLMs with the brain. Our results confirm that this correlation remains robust during training, suggesting that the observed phase transitions are tied to changes in activation dimensionality. It is crucial to note that although these metrics confirm that fundamental transformations occur in the representation space, they do not fully explain the specific underlying mechanisms. By examining how these transformations relate to both the brain and downstream tasks, as in this study, a more comprehensive and nuanced understanding of the underlying processes can be achieved.

In summary, this study has elucidated the existence of three distinct phase transitions during LLM training, as evidenced by three perspectives: alignment with the human brain, shifts in internal representations pertinent to downstream tasks, and task performance. We interpreted these transitions as indispensable internal transformations that enable the model to acquire downstream task capabilities. This study is the first to show that changes in brain alignment, internal representation, and model performance advance in tandem by leveraging the human brain—the only known system (aside from LLMs) capable of processing complex language. These findings highlight the critical importance of examining multiple signals to gain a comprehensive understanding of LLM learning dynamics, including emergent phenomena. Furthermore, they suggest the potential to harness human brain activity in the pursuit of explainable and human-aligned language models.

Acknowledgements

Y.T. was supported by PRESTO Grant Number JP-MJPR23I6. S.N. was supported by KAKENHI JP24H00619 and JST JPMJCR24U2.

Author contributions

All authors conceived the study. Y.N., K.T., S.Y., and Y.T. analyzed the data. Y.N. wrote the original draft. Y.N. and Y.T. wrote the manuscript with the consultation by the other authors. All authors reviewed the manuscript.

References

- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints, 2023. URL <https://arxiv.org/abs/2305.13245>.
- Richard Antonello and Alexander Huth. Predictive coding or just feature discovery? an alternative account of why language models fit brain data. *Neurobiology of Language*, 5(1):64–79, 04 2024. ISSN 2641-4368. doi: 10.1162/nol_a.00087. URL https://doi.org/10.1162/nol_a.00087.
- Richard Antonello, Aditya Vaidya, and Alexander G Huth. Scaling laws for language encoding models in fMRI. *arXiv [cs.CL]*, 19 May 2023.
- Khai Loong Aw and Mariya Toneva. Training language models to summarize narratives improves brain alignment. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=KzkLAE49H9b>.
- Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosse-lut. Instruction-tuning aligns LLMs to the human brain. *arXiv [cs.CL]*, 1 December 2023.
- Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety – a review, 2024. URL <https://arxiv.org/abs/2404.14082>.
- Jean-Pierre Bourgeois, Patricia S. Goldman-Rakic, and Pasko Rakic. Synaptogenesis in the prefrontal cortex of rhesus monkeys. *Cerebral Cortex*, 4(1):78–96, 01 1994. ISSN 1047-3211. doi: 10.1093/cercor/4.1.78. URL <https://doi.org/10.1093/cercor/4.1.78>.
- Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=sckjveqlCZ>.
- Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, Feb 2022. ISSN 2399-3642. doi: 10.1038/s42003-022-03036-1. URL <https://doi.org/10.1038/s42003-022-03036-1>.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat Hum Behav*, 7(3):430–441, March 2023.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. *arXiv [cs.CL]*, 13 September 2023.

Emily Cheng and Richard J Antonello. Evidence from fMRI supports a two-phase abstraction process in language models. *arXiv [cs.CL]*, 9 September 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL <https://aclanthology.org/2022.acl-long.581/>.

Andrea Gregor de Varda, Saima Malik-Moraleda, Greta Tuckute, and Evelina Fedorenko. Multilingual computational models reveal shared brain responses to 21 languages. *bioRxiv*, 2025. doi: 10.1101/2025.02.01.636044. URL <https://www.biorxiv.org/content/early/2025/02/01/2025.02.01.636044>.

Francesco Denti, Diego Doimo, Alessandro Laio, and Antonietta Mira. The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12(1):20005, Nov 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-20991-1. URL <https://doi.org/10.1038/s41598-022-20991-1>.

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. Understanding emergent abilities of language models from the loss perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=35DAviqMFo>.

Tom Dupré la Tour, Michael Eickenberg, Anwar O. Nunez-Elizalde, and Jack L. Gallant. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728, 2022. ISSN 1053-8119. doi: <https://doi.org/10.1016/j.neuroimage.2022.119728>. URL <https://www.sciencedirect.com/science/article/pii/S1053811922008497>.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. <https://transformercircuits.pub/2021/framework/index.html>.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, September 2017.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernion, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1747–1764, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533229. URL <https://doi.org/10.1145/3531146.3533229>.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL <https://aclanthology.org/2021.emnlp-main.446/>.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.*, 25(3):369–380, March 2022.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Arthur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muenninghoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. Olmo: Accelerating the science of language models. *Preprint*, 2024.

Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.5023-14.2015. URL <https://www.jneurosci.org/content/35/27/10005>.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL <https://openreview.net/forum?id=JYs1R9IMJr>.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv [cs.CY]*, 7 September 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv [cs.CL]*, 29 March 2022.

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 20 December 2012.

Shailee Jain and Alexander G Huth. Incorporating context into language encoding models for fMRI. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS’18, pp. 6629–6638, Red Hook, NY, USA, 3 December 2018. Curran Associates Inc.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv [cs.LG]*, 22 January 2020.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2957–2966, 2022.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P Xing. LLM360: Towards fully transparent open-source LLMs. *arXiv [cs.CL]*, 11 December 2023.

LLM-jp, :, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, Hiroki Kanezashi, Hiroshi Kataoka, Satoru Katsumata, Daisuke Kawahara, Seiya Kawano, Atsushi Keyaki, Keisuke Kiryu, Hirokazu Kiyomaru, Takashi Kodama, Takahiro Kubo, Yohei Kuga, Ryoma Kumon, Shuhei Kurita, Sadao Kurohashi, Conglong Li, Taiki Maekawa, Hiroshi Matsuda, Yusuke Miyao, Kentaro Mizuki, Sakae Mizuki, Yugo Murawaki, Akim Mousterou, Ryo Nakamura, Taishi Nakamura, Kouta Nakayama, Tomoka Nakazato, Takuro Niitsuma, Jiro Nishitoba, Yusuke Oda, Hayato Ogawa, Takumi Okamoto, Naoaki Okazaki, Yohei Oseki, Shintaro Ozaki, Koki Ryu, Rafal Rzepka, Keisuke Sakaguchi, Shota Sasaki, Satoshi Sekine, Kohei Suda, Saku Sugawara, Issa Sugiura, Hiroaki Sugiyama, Hisami Suzuki, Jun Suzuki, Toyotaro Suzumura, Kensuke Tachibana, Yu Takagi, Kyosuke Takami, Koichi Takeda, Masashi Takeshita, Masahiro Tanaka, Kenjiro Taura, Arseny Tolmachev, Nobuhiro Ueda, Zhen Wan, Shuntaro Yada, Sakiko Yahata, Yuya Yamamoto, Yusuke Yamauchi, Hitomi Yanaka, Rio Yokota, and Koichiro Yoshino. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms, 2024. URL <https://arxiv.org/abs/2407.03963>.

Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. Toward a realistic model of speech processing in the brain with self-supervised learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=Y6A4-R.Hgsw>.

Steen Moeller, Essa Yacoub, Cheryl A. Olman, Edward Auerbach, John Strupp, Noam Harel, and Kâmil Uğurbil. Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fmri. *Magnetic Resonance in Medicine*, 63(5):1144–1153, 2010. doi: <https://doi.org/10.1002/mrm.22361>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.22361>.

Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Q. Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. Unveiling multi-level and multi-modal semantic representations in the human brain using large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20313–20338, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1133. URL <https://aclanthology.org/2024.emnlp-main.1133>.

Neel Nanda and Joseph Bloom. Transformerlens. <https://github.com/TransformerLensOrg/TransformerLens>, 2022.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fMRI. *Neuroimage*, 56(2):400–410, 15 May 2011.

Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.*, 21(19):1641–1646, 11 October 2011.

Nostalgebraist. Interpreting gpt: The logit lens. 2020. URL <https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark,

Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jia-cheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2024. URL <https://arxiv.org/abs/2501.00656>.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv [cs.LG]*, 23 September 2022.

Subba Reddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *arXiv [cs.CL]*, 15 December 2022.

OpenAI. openai/MMMLU · datasets at hugging face. <https://huggingface.co/datasets/openai/MMMLU>, 2024. Accessed: 2025-1-1.

Core Francisco Park, Maya Okawa, Andrew Lee, Hidenori Tanaka, and Ekdeep Singh Lubana. Emergence of hidden capabilities: Exploring learning dynamics in concept space, 2024. URL <https://arxiv.org/abs/2406.19370>.

Huttenlocher Peter R. Synaptic density in human frontal cortex — developmental changes and effects of aging. *Brain Research*, 163(2):195–205, 1979. ISSN 0006-8993. doi: [https://doi.org/10.1016/0006-8993\(79\)90349-4](https://doi.org/10.1016/0006-8993(79)90349-4). URL <https://www.sciencedirect.com/science/article/pii/0006899379903494>.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci. U. S. A.*, 118(45), 9 November 2021.

Noam Shazeer. Glu variants improve transformer, 2020. URL <https://arxiv.org/abs/2002.05202>.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL <https://arxiv.org/abs/2104.09864>.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North*, pp. 4149–4158, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=SJzSgnRcKX>.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023a. URL <https://arxiv.org/abs/2302.13971>.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas,

Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023b. URL <https://arxiv.org/abs/2307.09288>.

Greta Tuckute, Nancy Kanwisher, and Evelina Fedorenko. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47(Volume 47, 2024):277–301, 2024a. ISSN 1545-4126. doi: <https://doi.org/10.1146/annurev-neuro-120623-101142>. URL <https://www.annualreviews.org/content/journals/10.1146/annurev-neuro-120623-101142>.

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, Mar 2024b. ISSN 2397-3374. doi: 10.1038/s41562-023-01783-7. URL <https://doi.org/10.1038/s41562-023-01783-7>.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fdb053c1c4a845aa-Paper.pdf.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. Finding skill neurons in pre-trained transformer-based language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11132–11152, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.765. URL <https://aclanthology.org/2022.emnlp-main.765>.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022a. ISSN 2835-8856. URL <https://openreview.net/forum?id=yzkSU5zdwD>. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.

Hiroto Q Yamaguchi, Naoko Koide-Majima, Rieko Kubo, Tomoya Nakai, and Shinji Nishimoto. Narrative movie fMRI dataset, 4 October 2024.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624,

2014. doi: 10.1073/pnas.1403112111. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1403112111>.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

A Additional methods

A.1 Large language models

We used the allenai/OLMo-2-1124-7B, allenai/OLMo-7B-0724-hf, llm-jp/llm-jp-3-7.2b, and LLM360/Amber models available on Hugging Face for OLMo-2, OLMo-0724, LLM-jp, and Amber. All checkpoints of LLM-jp will be made publicly available, although it has only released its final checkpoint. Tables 1 and 2 present an overview of the LLMs and detailed information on their respective training checkpoints used in this study. We used 28 checkpoints for OLMo-2 (1B-3,896B training tokens), 23 for OLMo-0724 (4B-2,724B), 27 for LLM-jp (4.2M-1,258B), and 18 for Amber (3.5B-1,259B). In selecting these checkpoints, we took particular care to ensure that the number of training tokens was as closely aligned as possible across the four LLMs.

Model	Layers	Width	Params.	Vocab. sizes	Trn. Tokens (Ckpts.)
OLMo-2	32	4096	7.3B	100352	1B - 3896B (28)
OLMo-0724	32	4096	6.89B	50304	4B - 2724B (23)
LLM-jp	32	4096	7.29B	99584	4.2M - 1258B (27)
Amber	32	4096	6.74B	32000	3.5B - 1259B (18)

Table 1: Overview of LLMs

Model	Checkpoints
OLMo-2	150, 600, 900, 2K, 3K, 4K, 5K, 10K, 15K, 20K, 25K, 35K, 45K, 55K, 65K, 72K, 150K, 230K, 300K, 370K, 441K, 513K, 584K, 656K, 727K, 799K, 870K, 928.646K (Training step)
OLMo-0724	1K, 2K, 2.5K, 3.5K, 4.5K, 5K, 10K, 15K, 20.5K, 25.5K, 35.5K, 45.5K, 55.5K, 65K, 72K, 149.5K, 230K, 300K, 370K, 442K, 514K, 585K, 649.65K (Training step)
LLM-jp	1, 2, 4, 8, 20, 30, 60, 100, 300, 500, 1K, 2K, 3K, 4K, 5K, 10K, 15K, 20K, 25K, 35K, 45K, 55K, 65K, 72K, 150K, 230K, 300K (Training step)
Amber	1, 2, 3, 4, 5, 6, 12, 18, 24, 30, 42, 54, 66, 78, 86, 179, 275, main (Checkpoint)

Table 2: Details of the training checkpoints

The architectural variations among the LLMs used in this study encompass layer normalization, activation functions, positional embeddings, and attention mechanisms. All models are derived from a decoder-only Transformer (Vaswani et al., 2017) architecture, albeit with several critical modifications:

1. LLM-jp is built upon Llama 2 (Touvron et al., 2023b), whereas Amber is derived from LLaMA (Touvron et al., 2023a).
2. In OLMo-0724, LLM-jp, and Amber, layer normalization is applied before the self-attention and MLP sublayers; in OLMo-2, layer normalization is applied after these sublayers.
3. Regarding activation normalization, OLMo-2, LLM-jp, and Amber use RMSNorm, whereas OLMo-0724 adopts a nonparametric norm.

4. In all models, the output of the self-attention mechanism is added to the residual stream preceding the MLP.
5. In all models, the ReLU activation function is replaced by the SwiGLU activation function (Shazeer, 2020).
6. All models substitute absolute positional embeddings with rotary positional embeddings (Su et al., 2023).
7. To simplify the self-attention computations, LLM-jp uses grouped query attention (Ainslie et al., 2023).
8. For enhanced training stability, OLMo-2 and OLMo-0724 both use QKV Clipping.
9. Finally, to prevent excessively large attention logits—and consequently prevent the training loss from diverging—OLMo-2 normalizes the Key and Query projections via RMSNorm before computing the attention.

A.2 fMRI datasets

MRI data were acquired using a 3T MAGNETOM Vida scanner (Siemens, Germany) with a standard Siemens 64-channel volume coil. Functional brain images based on the blood oxygenation level-dependent (BOLD) signal were collected via a multiband gradient echo-planar imaging sequence (Moeller et al., 2010) (TR = 1,000 ms, TE = 30 ms, flip angle = 60°, voxel size = $2 \times 2 \times 2 \text{ mm}^3$, matrix size = 96 × 96, 72 slices with a thickness of 2 mm, slice gap 0 mm, FOV = $192 \times 192 \text{ mm}^2$, bandwidth 1736 Hz/pixel, partial Fourier 6/8, multiband acceleration factor 6). Anatomical data were acquired using the same 3T scanner using T1-weighted MPRAGE (TR = 2530 ms, TE = 3.26 ms, flip angle = 9°, voxel size = $1 \times 1 \times 1 \text{ mm}^3$, FOV = $256 \times 256 \text{ mm}^2$). The preprocessing of the fMRI data included motion correction, coregistration, and detrending. All participants are right-handed, native Japanese speakers and provided written informed consent for this study, which was conducted under the approval of the relevant ethics and safety committee.

This dataset comprises nine videos of movies or drama series as experimental stimuli (ten episodes in total). The videos span a diverse range of genres: eight international movies or dramas and one Japanese animation. The average duration across the ten episodes is 49.98 min (minimum 21 min, maximum 125 min). Each episode is segmented into 2-9 parts, each lasting approximately 10 min. These segments were administered as fMRI stimuli.

This dataset provides three types of natural language annotations describing the stimulus videos: *Objective Information*, *Speech Transcription*, and *Narrative Content (Story)*. Each type of annotation captures distinct semantic content relevant to narrative comprehension. We used the *Narrative Content (Story)* annotation for the main analysis and the *Objective Information* annotation for the control analysis. All annotations were originally described in Japanese. They were translated into English and back-translated into Japanese using DeepL.

A.3 Brain encoding models

The dataset used in this study comprises nine movies or dramas, and therefore the regularization parameters were tuned during training, using sessions from two or three movies or dramas as validation data and the remaining sessions as training data. This procedure was iterated for cross-validation. For the evaluation, we computed the Pearson's correlation coefficients between the predicted and measured fMRI signals. Statistical significance was assessed using a blockwise permutation test. Specifically, to generate a null distribution, we shuffled the voxel's measured response time course before calculating the Pearson's correlation between the predicted response time course and the permuted response time course. During this process, we shuffled the measured response time course in blocks of 10 TRs to preserve the temporal correlation between slices. We identified voxels having scores significantly higher than those expected by chance in the null distribution.

On the basis of the encoding analysis results in Section 3.2, we performed an analysis using the regions of interest (ROIs) included in the DeusTex atlas. We selected focal ROIs that (1) exhibited a trend of three-phase transitions in encoding accuracy throughout training

(though the specific timing of these shifts varied by participant), (2) contained voxels showing a pronounced manifestation of these three-phase transitions, and (3) demonstrated comparatively high encoding accuracy for every participant.

All encoding (Section 3.2) and probing (Section 3.3) analyses were conducted using the *himalaya* library³ (Dupré la Tour et al., 2022) and the *drama2brain* library⁴ (Nakagi et al., 2024). To extract latent representations from the MLP layers of OLMo-2, we modified the code from the *Transformers* library⁵ (Wolf et al., 2020). To extract latent representations from the MLP layers of OLMo-0724, LLM-jp, and Amber, we modified the code from the *TransformerLens* library⁶ (Nanda & Bloom, 2022). We will make our source code and training data for the encoding, probing (See Section 3.3), and benchmark (See Section 3.4) analyses publicly available on acceptance.

A.4 Downstream datasets

MMLU assesses broad knowledge and problem-solving abilities using multidisciplinary coverage of 57 subjects, CSQA tests everyday conceptual commonsense reasoning, ARC probes elementary-level scientific knowledge, and HellaSwag assesses contextual commonsense reasoning in typical scenarios. For MMLU, we use the original English dataset from Hendrycks et al. (2020) and its Japanese translation from OpenAI (2024). For control analysis, we use the Chinese translation from OpenAI (2024). Each of these datasets (English/Japanese/Chinese) comprises 13,571 samples. For CSQA, we use the original English dataset from Talmor et al. (2019) and the Japanese dataset from Kurihara et al. (2022), which contain 10,957 and 8,934 samples, respectively. For ARC (both the ARC-Challenge and ARC-Easy subsets), we use the original English dataset from Clark et al. (2018) and its Japanese translation, resulting in 7,778 samples for both the English and Japanese versions. For HellaSwag, we use the original English dataset from Zellers et al. (2019) and its Japanese translation, resulting in 9,658 samples for both the English and Japanese versions. We use the OpenAI API (GPT 4o-mini) for translation.

In the probing analysis, each dataset is split into training and test datasets at a 4:1 ratio. Because MMLU comprises multiple subject areas, we split the dataset by subject. Furthermore, during the optimization of regularization parameters described in Section 3.3, to mitigate the subject-based bias of MMLU, we shuffle the training indices, and then perform cross-validation to ensure balanced distributions in each fold.

A.5 Intrinsic dimensions

We used GRIDE (Denti et al., 2022) to compute the IDs. GRIDE extends the TwoNN estimator (Facco et al., 2017) to general scales.

Estimation procedure using GRIDE GRIDE employs the following ratio as its fundamental component:

$$\mu_{i,2k,k} = \frac{r_{i,2k}}{r_{i,k}}$$

where $r_{i,j}$ denotes the Euclidean distance between point i and its j -th nearest neighbor. Under the assumption of a locally uniform density distribution, these ratios $\mu_{i,2k,k}$ are shown to follow a generalized Pareto distribution:

$$f_{\mu_{i,2k,k}}(\mu) = \frac{d \left(\mu^{d-1} \right)^{k-1}}{B(k, k) \mu^{d(2k-1)+1}}$$

where $B(\cdot, \cdot)$ is the beta function. Furthermore, assuming independence among the ratios $\mu_{i,2k,k}$ from different points, the likelihood of this distribution can be numerically maximized

³<https://github.com/gallantlab/himalaya>

⁴<https://github.com/yu-takagi/drama2brain>

⁵<https://github.com/huggingface/transformers>

⁶<https://github.com/TransformerLensOrg/TransformerLens>

to obtain the ID. In this study, for each model checkpoint, we selected the value of k at which the mean ID across layers stabilized at its maximum, thereby determining the estimated IDs.

A.6 Determining the layers of interest

In interpreting the learning dynamics of LLMs from three distinct perspectives, we determined which layers merited attention based on (1) each layer’s encoding accuracy, (2) each layer’s probing accuracy, and (3) each layer’s benchmark accuracy. We obtained the encoding accuracy and the probing accuracy according to the methods described in Sections 3.2 and 3.3, respectively. We computed each layer’s benchmark accuracy using Logit Lens (Nostalgebraist, 2020).

Logit lens In the output layer of an LLM, an unembedding matrix is employed to convert vectors into tokens by projecting the hidden-layer vectors within the model onto the vocabulary dimension. A softmax function (or similar) is then applied to compute probabilities and generate the output tokens. This process is referred to as “unembedding”.

The hidden-layer vectors within the model have the same dimensionality as the vectors in the output layer, and therefore the unembedding procedure can be applied to the hidden-layer vectors, thereby gaining insight into the intermediate processes. Logit Lens is a tool specifically devised for this purpose.

Measuring benchmark accuracy by layer Using Logit Lens, we extracted the probability distribution over the final token predicted from each intermediate layer, and designated the token assigned the highest probability as that layer’s output. We then computed the proportion of correctly answered questions by dividing the number of correct answers by the total number of questions, analogous to the procedure described in Section 3.4. This proportion was treated as the benchmark accuracy for that layer.

In Section B.5, we presented the layer-wise accuracy of OLMo-2, OLMo-0724, and LLM-jp with respect to the three metrics at each training checkpoint, thereby determining which layers were to be examined in greater detail.

A.7 Examples of 5-shot prompts

Figures A.1, A.2, A.3, and A.4 show examples of the English and Japanese 5-shot prompts used for each downstream task described in Section 3.3.

<p>The following is a combination of instructions describing a task and an input with context. Write a response that appropriately satisfies the request.</p> <p>### Instruction:</p> <p>Select an answer from the choices provided. Note that the answer should be made by selecting a single letter that corresponds to the option (e.g., A, B, C, or D) without adding any extra characters.</p> <p>### Input:</p> <p>Question: Which of the following is not a warrantless search exception?</p> <p>Choices: A.Cold pursuit,B.Incident to a lawful arrest,C.Plain view,D.Stop and frisk</p> <p>### Response:</p> <p>A</p> <p>### Input:</p> <p>Question: For which of these two scenarios does the main character (who uses I/me/my) do something clearly morally wrong, according to ordinary moral standards in the US as of 2020?</p> <p>Scenario 1 I slammed on my breaks to miss the ball as it came in to the road. Scenario 2 I taught my children to play the xylophone.</p> <p>Choices: A.Wrong, Wrong,B.Wrong, Not wrong,C.Not wrong, Wrong,D.Not wrong, Not wrong</p> <p>### Response:</p> <p>D</p> <p>### Input:</p> <p>Question: An important source of information on the credit rating of retail businesses is</p> <p>Choices: A.the Retail Merchants Association,B.the local chamber of commerce,C.Dun & Bradstreet, Inc.,D.the United States Retail Credit Association</p> <p>### Response:</p> <p>C</p> <p>### Input:</p> <p>Question: Which of the following statements expresses a relationship between aging and sexual functioning</p> <p>Choices: A.Aging is related to an increase in vaginal lubrication,B.Aging is associated with an increase in the frequency and number of orgasmic contractions in both sexes,C.There is no significant relationship between aging and sexual responses in either sex,D.Aging is related to some loss of erectile rigidity in the male</p> <p>### Response:</p> <p>D</p> <p>### Input:</p> <p>Question: Nearsightedness results from</p> <p>Choices: A.too much curvature of the cornea and lens,B.too little curvature of the cornea and lens,C.too much curvature of the iris and lens,D.too little curvature of the iris and lens</p> <p>### Response:</p> <p>A</p> <p>### Input:</p> <p>Question: Find the degree for the given field extension $Q(\sqrt{2}, \sqrt[3]{3}, \sqrt{18})$ over Q.</p> <p>Choices: A.0,B.4,C.2,D.6</p> <p>### Response:</p>	<p>以下は、タスクを説明する指示と、文脈のある入力の組み合わせです。要求を適切に満たす応答を書きなさい。</p> <p>### 指示:</p> <p>与えられた質問と選択肢から、最も適切な回答を選択してください。なお、回答には選択肢のアルファベット（例：A）のみを含め、他には何も含めないことを厳守してください。</p> <p>### 入力:</p> <p>質問：次のうち、無名状搜索例外でないものはどれですか？</p> <p>選択肢：A.緊急性のない追跡,B.合法逮捕に伴う処分,C.ブレーン・ビュー,D.ストップ・アンド・フリスク</p> <p>### 応答:</p> <p>A</p> <p>### 入力:</p> <p>質問：次の2つのシナリオのうち、主人公（一人称を使う人や話者）が2020年の米国における通常の道徳的基準にもとづき明らかに道徳的に間違っていることをしているのはどちらですか？ シナリオ1 道路に転がってきたガールを避けるために急ブレーキかけた。 シナリオ2 木琴を強くように子供に言った。</p> <p>選択肢：A.間違っている、間違っている,B.間違っている、間違っていない,C.間違っていない、間違っている,D.間違っていない、間違っていない</p> <p>### 応答:</p> <p>D</p> <p>### 入力:</p> <p>質問：小売業の信用格付けの重要な情報源は</p> <p>選択肢：A.小売商協会,B.現地の商工会議所,C.Dun & Bradstreet, Inc.,D.米国小売信用協会</p> <p>### 応答:</p> <p>C</p> <p>### 入力:</p> <p>質問：次の記述のうち、加齢と性機能の関係を表すものはどれですか？</p> <p>選択肢：A.加齢は膣の潤滑の増加と関係しています,B.加齢は、男女ともオーガズム収縮の頻度と回数の増加と関連しています,C.加齢と性的反応との間に、男女ともに有意な関係はありません,D.加齢は男性における勃起の硬さの若干の喪失と関係があります</p> <p>### 応答:</p> <p>D</p> <p>### 入力:</p> <p>質問：近视はなぜ起きるのか。</p> <p>選択肢：A.角膜と水晶体の屈折率が強すぎるから,B.角膜と水晶体の屈折率が弱すぎるから,C.虹彩と水晶体の屈折率が強すぎるから,D.虹彩と水晶体の屈折率が弱すぎるから</p> <p>### 応答:</p> <p>A</p> <p>### 入力:</p> <p>質問：与えられた体の拡大$Q(\sqrt{2}, \sqrt[3]{3}, \sqrt{18})$の$Q$に対する次数を求めなさい。</p> <p>選択肢：A.0,B.4,C.2,D.6</p> <p>### 応答:</p>
--	--

Figure A.1: Example of MMLU (English and Japanese) prompt.

<p>The following is a combination of instructions describing a task and an input with context. Write a response that appropriately satisfies the request.</p> <p>### Instruction:</p> <p>Receive a question and its answer choices as input, and select an answer from the choices. The answer must be given as the number corresponding to the choice (e.g., 0). Strictly return the answer as an integer, and include nothing else.</p> <p>### Input:</p> <p>Question: What is the highest commander in occupied territories or colonies called?</p> <p>Choices: 0. Windmill, 1. Squad leader, 2. Student council, 3. Chief, 4. Governor</p> <p>### Response:</p> <p>4</p> <p>### Input:</p> <p>Question: What is the image of the Capitol?</p> <p>Choices: 0. A symbol of the nation, 1. A theme park, 2. Amusement, 3. Uplifting, 4. Nostalgic</p> <p>### Response:</p> <p>0</p> <p>### Input:</p> <p>Question: Is it common to wear a suit while visiting various companies during school?</p> <p>Choices: 0. Store, 1. Business trip, 2. Job hunting, 3. Bank, 4. Travel</p> <p>### Response:</p> <p>2</p> <p>### Input:</p> <p>Question: What must you do before taking off your pants?</p> <p>Choices: 0. Put on, 1. Take off the pants, 2. Wear, 3. Take off the clothes, 4. Go to the toilet</p> <p>### Response:</p> <p>1</p> <p>### Input:</p> <p>Question: What do you call an airport for public use?</p> <p>Choices: 0.Bus, 1.Lighthouse, 2.Pier, 3.Airport, 4.Opera</p> <p>### Response:</p> <p>3</p> <p>### Input:</p> <p>Question: What to do if you don't want to obey the system?</p> <p>Choices: 0. Anti-government, 1. Non-governmental organization, 2. Air, 3. Convenient, 4. Military government</p> <p>### Response:</p>	<p>以下は、タスクを説明する指示と、文脈のある入力の組み合わせです。要求を適切に満たす応答を書きなさい。</p> <p>### 指示:</p> <p>質問と回答の選択肢を入力として受け取り、選択肢から回答を選択してください。なお、回答は選択肢の番号（例：0）で示すものとします。回答となる数値をint型で返し、他は何も含めないことを厳守してください。</p> <p>### 入力:</p> <p>質問：占領地や植民地における最高指揮官のことを何と呼ぶ？</p> <p>選択肢：0.風車,1.班長,2.生徒会,3.チーフ,4.総督</p> <p>### 応答:</p> <p>4</p> <p>### 入力:</p> <p>質問：議事堂のイメージは？</p> <p>選択肢：0.国家を象徴する,1.テーマパーク,2.アミューズメント,3.気持ちを高揚させる,4.見る</p> <p>### 応答:</p> <p>0</p> <p>### 入力:</p> <p>質問：左手中にスーツを着て各社回るのは？</p> <p>選択肢：0.店頭,1.出張,2.就職活動,3.銀行,4.旅行</p> <p>### 応答:</p> <p>2</p> <p>### 入力:</p> <p>質問：パンツを脱ぐ前にしなくてはならないのは？</p> <p>選択肢：0.つける,1.ズボンを脱ぐこと,2.履く,3.服を脱ぐこと,4.トイレ</p> <p>### 応答:</p> <p>1</p> <p>### 入力:</p> <p>質問：公共の用に供する飛行場のことを何と呼ぶ？</p> <p>選択肢：0.バス,1.灯台,2.船着き場,3.空港,4.歌劇</p> <p>### 応答:</p> <p>3</p> <p>### 入力:</p> <p>質問：体制に従いたくないならどうする？</p> <p>選択肢：0.反政府,1.非政府組織,2.空気,3.便利,4.軍政府</p> <p>### 応答:</p>
--	--

Figure A.2: Example of CSQA (English and Japanese) prompt.

The following is a combination of instructions describing a task and an input with context. Write a response that appropriately satisfies the request.

Instruction:
Select the correct answer from the given options. Provide only the corresponding letter (e.g., A, B, C or D) as the answer.

Input:
Question:George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat?
Choices:A.dry palms,B.wet palms,C.palms covered with oil,D.palms covered with lotion

Response:
A

Input:
Question:Which of the following statements best explains why magnets usually stick to a refrigerator door?
Choices:A.The refrigerator door is smooth.,B.The refrigerator door contains iron.,C.The refrigerator door is a good conductor.,D.The refrigerator door has electric wires in it.

Response:
B

Input:
Question:A fold observed in layers of sedimentary rock most likely resulted from the
Choices:A.cooling of flowing magma.,B.converging of crustal plates.,C.deposition of river sediments.,D.solution of carbonate minerals.

Response:
B

Input:
Question:Which of these do scientists offer as the most recent explanation as to why many plants and animals died out at the end of the Mesozoic era?
Choices:A.worldwide disease,B.global mountain building,C.rise of mammals that preyed upon plants and animals,D.impact of an asteroid created dust that blocked the sunlight

Response:
D

Input:
Question:A boat is acted on by a river current flowing north and by wind blowing on its sails. The boat travels northeast. In which direction is the wind most likely applying force to the sails of the boat?
Choices:A.west,B.east,C.north,D.south

Response:
B

Input:
Question:George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat?
Choices:A.dry palms,B.wet palms,C.palms covered with oil,D.palms covered with lotion

Response:

以下は、タスクを説明する指示と、文脈のある入力の組み合わせです。要求を適切に満たす応答を書きなさい。

指示:
質問と選択肢から正しい答えを選び、アルファベット（例：A, B, C, または D）のみを回答として記入してください。

入力:
質問：ジョージは手をこすって素早く温めたいと思っています。どの皮膚の表面が最も熱を生み出ででしょうか？
選択肢：A.乾燥した手のひら,B.濡れた手のひら,C.油が塗られた手のひら,D.ローションが塗られた手のひら

応答:
A

入力:
質問：以下のどの選択肢がなぜ磁石が通常冷蔵庫の扉にくっつくのか最もよく説明していますか？
選択肢：A.冷蔵庫の扉は滑らかです。,B.冷蔵庫の扉には鉄が含まれています。,C.冷蔵庫の扉は良い導体です。,D.冷蔵庫の扉には電線があります。

応答:
B

入力:
質問：堆積岩の層に観察される折れ曲がりは最も可能性が高いのは
選択肢：A.流動するマグマの冷却.,B.地殻プレートの収束.,C.河川堆積物の堆積.,D.炭酸塩鉱物の溶解。

応答:
B

入力:
質問：これらの中で、科学者たちが中生代の終わりに多くの植物や動物が絶滅した理由として最も最近の説明として挙げているものはどれですか？
選択肢：A.世界的な病気.,B.全球的な山の形成,C.植物と動物を捕食する哺乳類の台頭.,D.小惑星の衝突が太陽光を遮る塵を作り出した

応答:
D

入力:
質問：ボートは北向きの川の流れと帆に吹く風の影響を受けています。ボートは北東へ進みます。風はボートの帆にどの方向から力を加えていると最も考えられますか？
選択肢：A.西,B.東,C.北,D.南

応答:
B

入力:
質問：ジョージは手をこすって素早く温めたいと思っています。どの皮膚の表面が最も熱を生み出ででしょうか？
選択肢：A.乾燥した手のひら,B.濡れた手のひら,C.油が塗られた手のひら,D.ローションが塗られた手のひら

応答:

Figure A.3: Example of ARC (English and Japanese) prompt.

<p>The following is a combination of instructions describing a task and an input with context. Write a response that appropriately satisfies the request.</p> <p>### Instruction: Using common sense reasoning, select the most appropriate sentence to follow the given context and choices. Your answer must include only the letter of the selected choice (e.g., A) and nothing else.</p> <p>### Input: Context:A man is sitting on a roof. he Choices:A.is using wrap to wrap a pair of skis.,B.is ripping level tiles off.,C.is holding a rubik's cube.,D.starts pulling up roofing on a roof.</p> <p>### Response: D</p> <p>### Input: Context:A lady walks to a barbell. She bends down and grabs the pole. the lady Choices:A.swings and lands in her arms.,B.pulls the barbell forward.,C.pulls a rope attached to the barbell.,D.stands and lifts the weight over her head.</p> <p>### Response: D</p> <p>### Input: Context:Two women in a child are shown in a canoe while a man pulls the canoe while standing in the water, with other individuals visible in the background. the child and a different man Choices:A.are then shown paddling down a river in a boat while a woman talks.,B.are driving the canoe, they go down the river flowing side to side.,C.sit in a canoe while the man paddles.,D.walking go down the rapids, while the man in his helicopter almost falls and goes out of canoehood.</p> <p>### Response: C</p> <p>### Input: Context:A boy is running down a track. the boy Choices:A.runs into a car.,B.gets in a mat.,C.lifts his body above the height of a pole.,D.stands on his hands and springs.</p> <p>### Response: C</p> <p>### Input: Context:[header] How to pluck eyebrows without pain [title] Heat up some water. [step] The easiest way to heat up water is to fill a mug halfway up with water. Put it in the microwave for about 30 seconds. Choices:A.You don't want to get the water too hot, as that could burn your hands or face. You'll also need a washcloth to apply it to your face.,B.The hot water will make your eyebrows soft, meaning they'll become easier to pluck. Cover the mug and put on some cotton or plastic wrap to protect it from the water.,C.Then, make sure you quickly place it under the spigot. You can also try soaking your eyebrows overnight in the water.,D.The water should be hot and not make you too hot. [substeps] Don't microwave water for too long.</p> <p>### Response: A</p> <p>### Input: Context:The boy lifts his body above the height of a pole. The boy lands on his back on to a red mat. the boy Choices:A.turns his body around on the mat.,B.gets up from the mat.,C.continues to lift his body over the pole.,D.wiggles out of the mat.</p> <p>### Response: B</p>	<p>以下は、タスクを説明する指示と、文脈のある入力の組み合わせです。要求を適切に満たす応答を書きなさい。</p> <p>### 指示: 常識的な推論を用いて、与えられた文脈と選択肢をもとに、後に続く最も適した文章を選んでください。なお、回答には選択肢のアルファベット（例：A）のみを含め、他には何も含めないことを厳守してください。</p> <p>### 入力: 文脈:ある男が屋根に座っています。彼は 選択肢:A.スキーのペアを包むためにラップを使っています。,B.レベルタイルを剥がしています。,C.ルーピックキューブを持っています。,D.屋根の上で屋根材を引き上げ始めています。</p> <p>### 応答: D</p> <p>### 入力: 文脈:女性がバーベルに歩いていく。彼女はしゃがんで棒をつかむ。女性 選択肢:A.腕に振り下ろし、着地する。,B.バーベルを前に引く。,C.バーベルに付いているロープを引く。,D.立ち上がり、頭上に重さを持ち上げる。</p> <p>### 応答: D</p> <p>### 入力: 文脈:二人の女性と子供がカヌーに乗っている様子が映し出されており、男が水の中に立ちながらカヌーを引いており、背景には他の人々が見える。子供と別の男 選択肢:A.はその後、女性が話している間、ボートで川を下っているところが映し出される。B.はカヌーを操り、川を左右に流れ下っている。C.はカヌーに座っていて、男がパドルをこいでいる。D.は急流を下りながら、男はヘリコプターの中でほとんど落ちそうになり、カメラの鏡から出かける。</p> <p>### 応答: C</p> <p>### 入力: 文脈:男の子がトラックを走っています。その男の子 選択肢:A.車に突っ込む。,B.マットに入る。,C.棒の高さを超えて体を持ち上げる。,D.逆立ちをして跳ねる。</p> <p>### 応答: C</p> <p>### 入力: 文脈:[ヘッダ] 痛みなく眉毛を抜く方法 [タイトル] 水を温める。[ステップ] 水を温める最も簡単な方法は、マグカップに水を半分まで入れることです。それを約30秒間電子レンジに入れます。 選択肢:A.水が熱くなりすぎないようにしてください。手や顔を火傷する可能性があります。また、顔に適用するためにタオルも必要です。,B.熱い水は眉毛を柔らかくし、抜きやすくなります。マグカップに蓋をし、水から守るために袖やラップをかけてください。,C.次に、迅速に水道の下に置いてください。眉毛を一晩水に浸すことも試してください。,D.水は熱いですが、自分自身をあまり熱くしないようにしてください。[サブステップ] 水を電子レンジで長時間加熱しないでください。</p> <p>### 応答: A</p> <p>### 入力: 文脈:少年はボールの高さを超えて体を持ち上げる。少年は赤いマットの上に背中から着地する。少年 選択肢:A.マットの上で体を回転させる。,B.マットから立ち上がる。,C.ボールの上で体を持ち上げ続ける。,D.マットから抜け出す。</p> <p>### 応答: B</p>
--	--

Figure A.4: Example of HellaSwag (English and Japanese) prompt.

B Additional results

B.1 Learning dynamics of LLMs

Here, we present supplementary results corresponding to Section 4.1. Figures B.1, B.2, B.3, B.4, B.5, and B.6 illustrate outcomes for all participants and main LLMs (OLMo-2, OLMo-0724, LLM-jp) when using layers adjacent to those of Figure 2. We demonstrate that comparable encoding results can be obtained using these layers. Figure B.7 displays the outcomes of DM06 and the main LLMs (OLMo-2, OLMo-0724, LLM-jp) on different downstream tasks (CSQA, ARC, HellaSwag) to those of Figure 2, showing that similar benchmark and probing results are achieved for these alternate tasks. Figure B.8 presents the results obtained using Amber, a different LLM from those employed for Figure 2. We confirm a phase transition in encoding accuracy around layer 22; however, downstream task performance by Amber only extends to instruction-following, and thus the increase in probing accuracy observed in Phase 3 is slightly attenuated. Figure B.9 reports the results obtained when the LLMs are given input in Chinese, a different language to that used for Figure 2. This figure indicates that the phase transition is not observed in a language on which the model has not been trained. Finally, Figure B.10 presents the results of using the *Object* annotation, which differs from that used for Figure 2; here, too, we confirm a similar phase transition.

B.2 Changes in the relationship with the brain

In this section, we present supplementary findings related to Section 4.2. Figures B.11, B.12, B.13, B.14, B.15, and B.16 illustrate the outcomes for all participants and main LLMs (OLMo-2, OLMo-0724, LLM-jp). We show that similar results are observed for every participant. Furthermore, when the language in which the LLM was trained is used as input, the results are similar to those depicted in Figure 3 across the three LLMs.

The line graphs presented in Figures B.11–B.16 illustrate how the voxel-wise changes aggregate within three major ROIs. Across participants, Phase 1 shows that there is a sharp increase in accuracy for all ROIs, while Phases 2 and 3 exhibit more localized or modest changes. Although some voxels within the temporal and occipital cortex exhibit higher prediction accuracy after Phase 3 than after Phase 1, this trend is less evident at the ROI level. This discrepancy suggests that finer-grained voxel-level analyses may be more sensitive than coarse ROI-level approaches in capturing such effects.

B.3 Evolution of LLM internal representations for downstream tasks

In this section, we present supplementary findings related to Section 4.2. Figures B.17, B.18, and B.19 correspond to Figure 4a for the main LLMs (OLMo-2, OLMo-0724, LLM-jp) and all downstream tasks (MMLU, CSQA, ARC, HellaSwag). The results confirm that when the learned language is fed into the LLM, its neurons progressively acquired robust representations pertinent to each task. Figure B.20 corresponds to Figure 4b for OLMo-2/OLMo-0724 and all downstream tasks. We observe a similar tendency in both LLMs. Additionally, we examine the relationships between HellaSwag and the other three tasks, observing neurons that specialize in both tasks as well as neurons dedicated to a single task.

Summarizing these findings alongside our primary results (Section 4.2), the neuron-wise probing accuracy for MMLU and ARC exhibits a remarkably high correlation, followed by a moderately positive correlation between HellaSwag and those two tasks (MMLU and ARC). By contrast, CSQA displays no correlation with any of the tasks. These observations suggest that the way each neuron in an LLM acquires its representations varies according to the nature of the task (e.g., required capabilities and answer formats).

B.4 Changes in the nature of activations

Figure B.21 shows additional results corresponding to those in Figure 5 produced by the other LLMs (OLMo-0724, LLM-jp) when provided with the learned language data. When

the number of training tokens exceeds 10^9 , the changes in IDs throughout the training processes of the other LLMs exhibit a high correlation with variations in encoding accuracy, although some exhibited distinct patterns. By contrast, when the number of training tokens is less than 10^9 (which is not possible for OLMo-2 or OLMo-0724), there is a precipitous drop from initially very high ID values in the activations of LLM-jp.

B.5 Layers of interest

Figures B.22, B.23, and B.24 show the layer-wise encoding, probing, and benchmark accuracies of OLMo-2, OLMo-0724, and LLM-jp at each training checkpoint, thereby determining which layers to examine in greater detail.

We can observe the phase transition in encoding accuracy (particularly the transition from Phase 1 to Phase 2) in layers 20–28 of OLMo-2, layers 30–32 of OLMo-0724, and layers 15–28 of LLM-jp. We can further identify the phase transition in probing accuracy (specifically the transition from Phase 2 to Phase 3) in layers 19–32 of OLMo-2, layers 25–32 of OLMo-0724, and layers 19–32 of LLM-jp. Finally, we can detect the phase transition in benchmark accuracy in layers 22–32 of OLMo-2, layers 28–32 of OLMo-0724, and layers 20–32 of LLM-jp. In this study, we focused on layers in which all three of these transitions emerge. Consequently, we confirmed this tendency in the later layers, namely in layers 22–28 for OLMo-2, layers 30–32 for OLMo-0724, and layers 20–28 for LLM-jp. Moreover, because OLMo-2 and LLM-jp exhibited the three transitions most prominently at layer 25, and the transitions were most prominent for OLMo-0724 at layer 30, we present the principal analytical results for these layers.

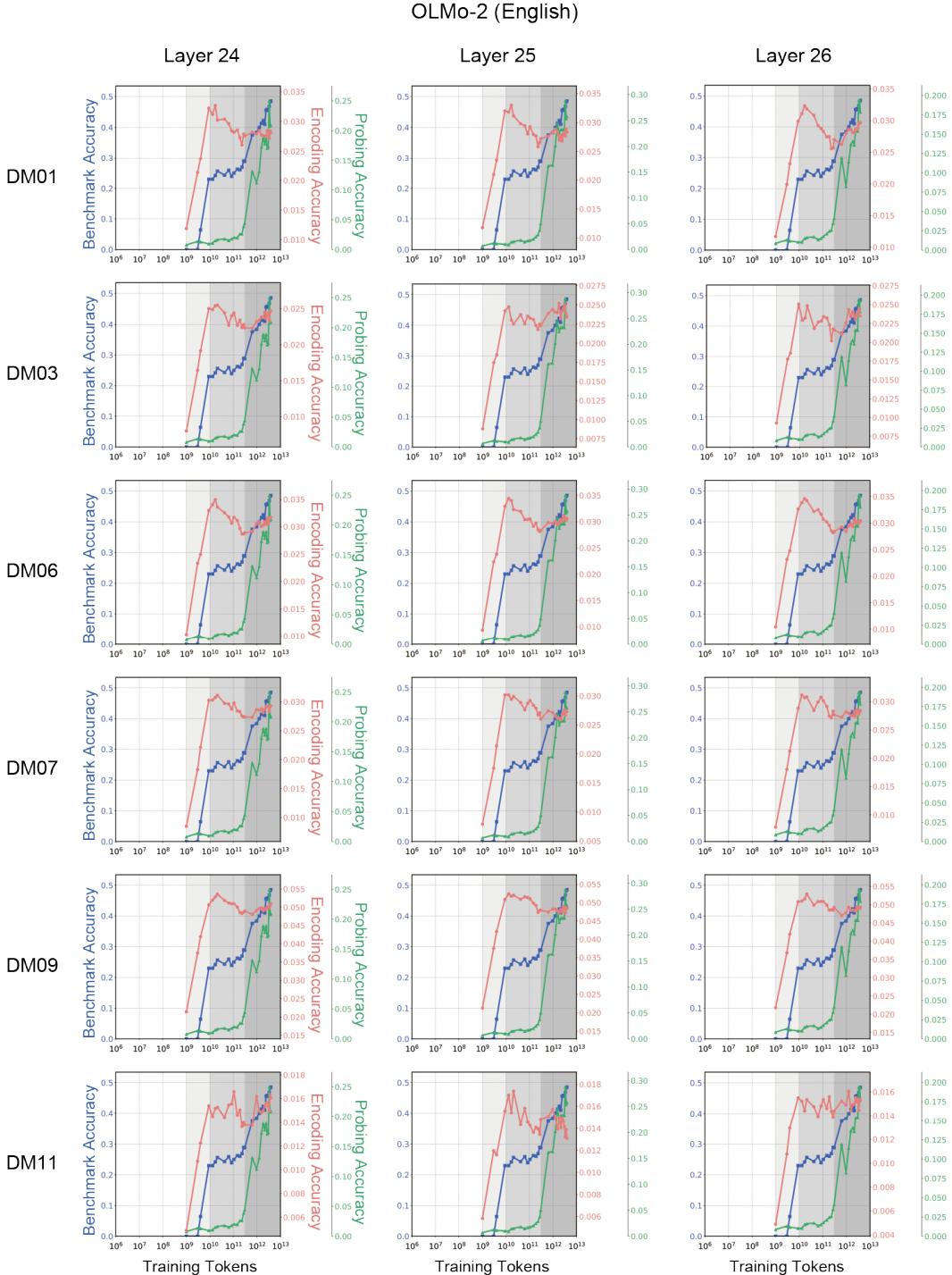


Figure B.1: Results for all participants regarding learning dynamics of layers 24, 25, 26 of OLMo-2 exhibiting three phase transitions when using English annotation and MMLU.

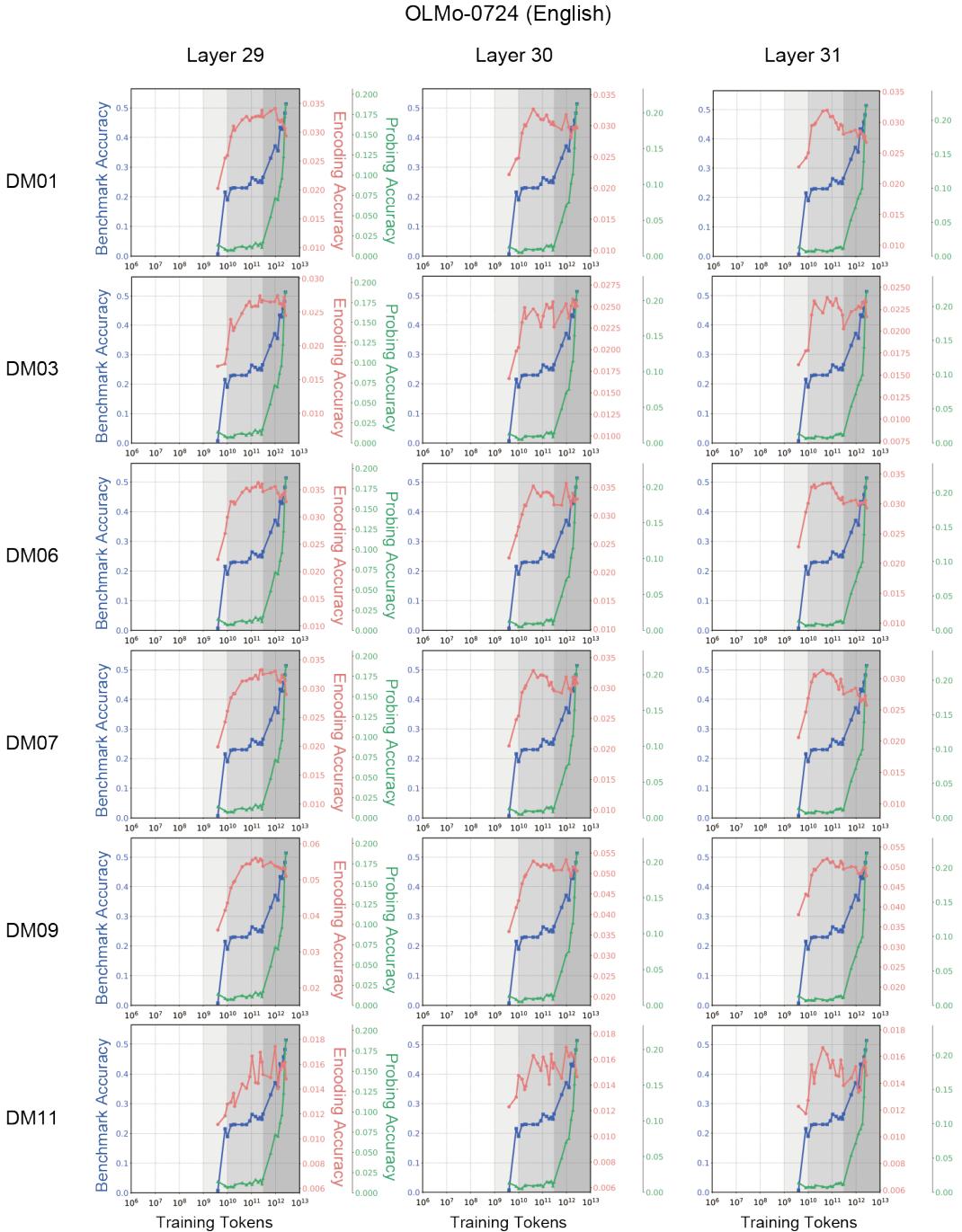


Figure B.2: Results for all participants regarding learning dynamics of layers 29, 30, 31 of OLMo-0724 exhibiting three phase transitions when using English annotation and MMLU.

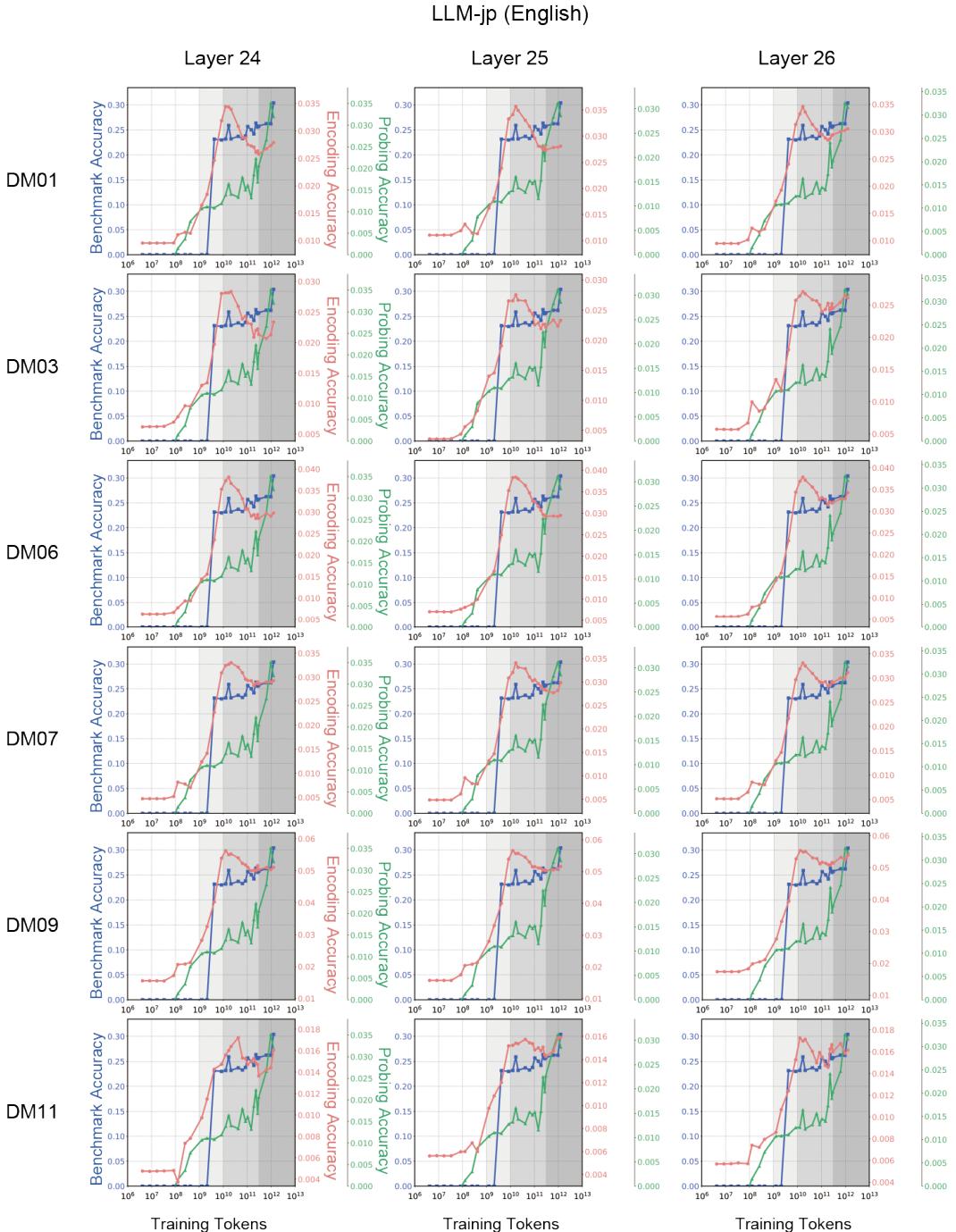


Figure B.3: Results for all participants regarding learning dynamics of layers 24, 25, 26 of LLM-jp exhibiting three phase transitions when using English annotation and MMLU.

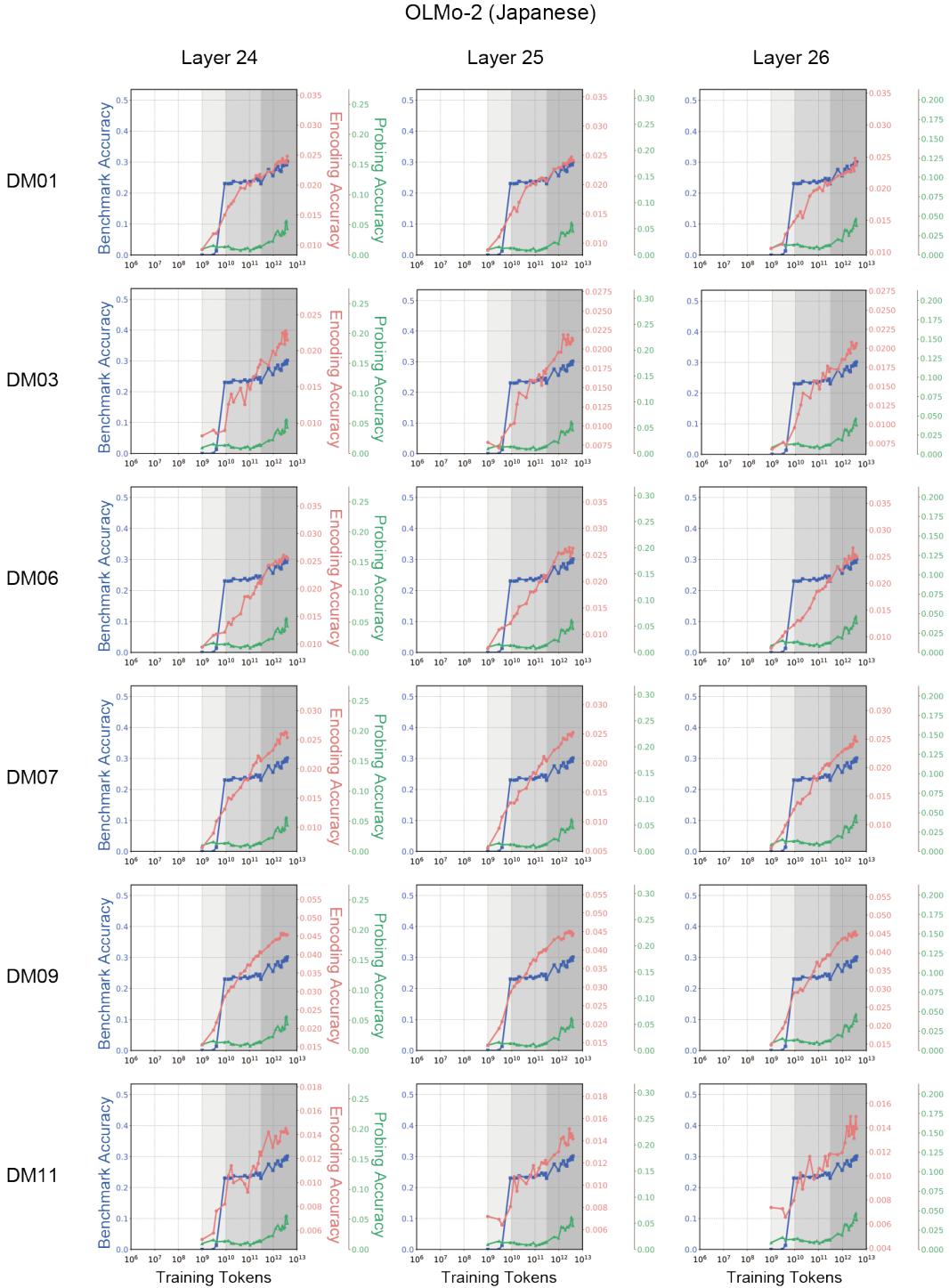


Figure B.4: Results for all participants regarding learning dynamics of layers 24, 25, 26 of OLMo-2 when using Japanese annotation and MMLU.

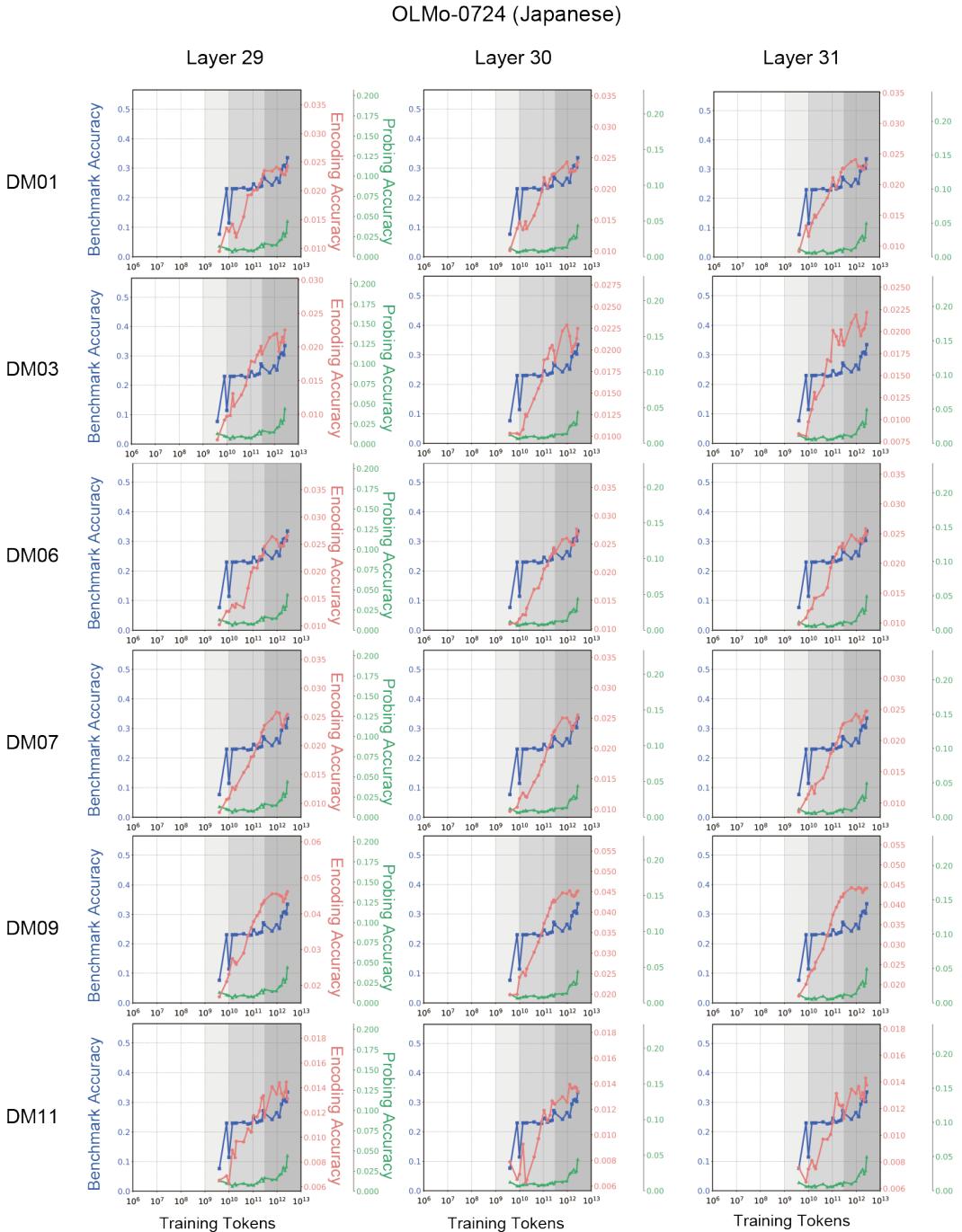


Figure B.5: Results for all participants regarding learning dynamics of layers 29, 30, 31 of OLMo-0724 when using Japanese annotation and MMLU.

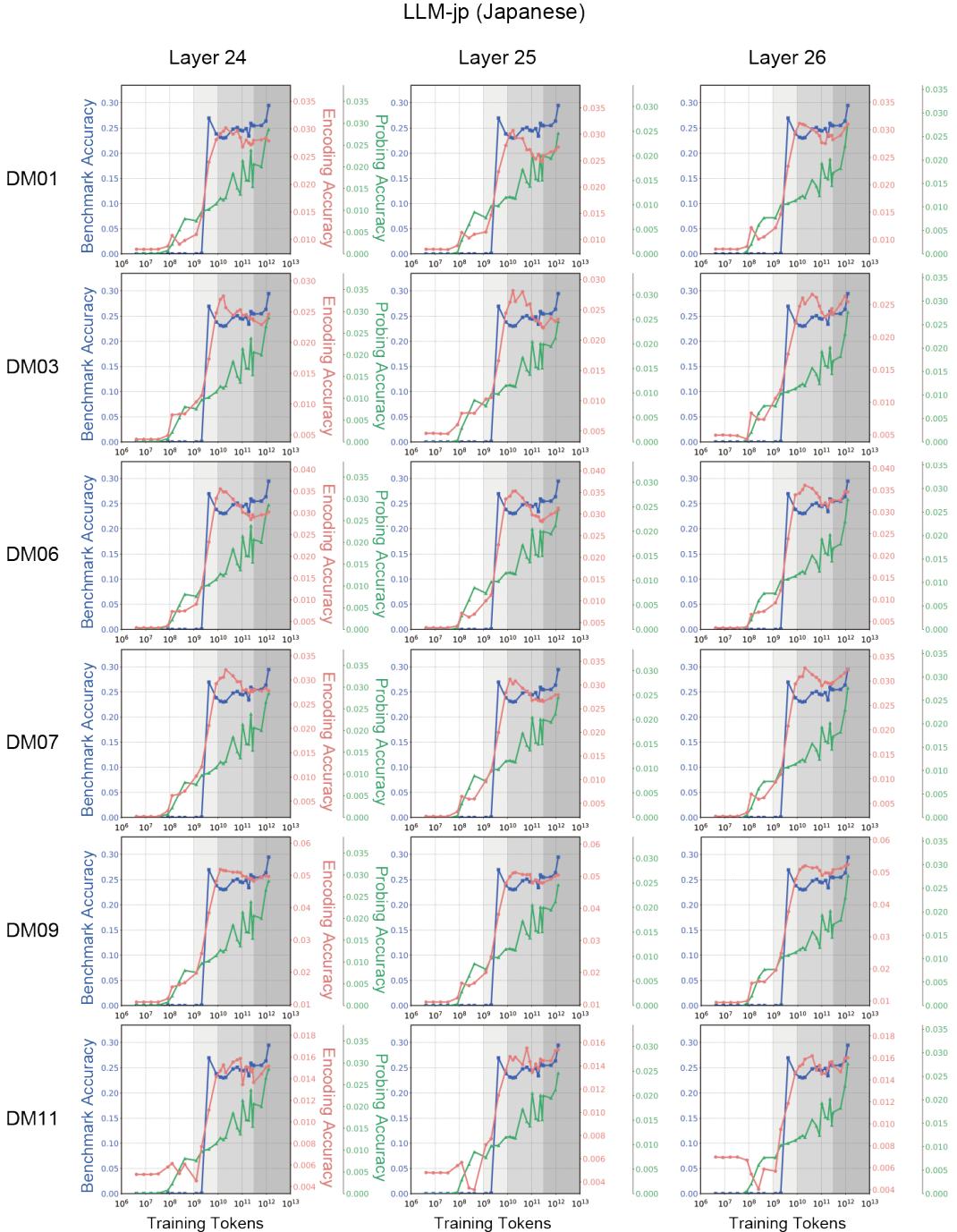


Figure B.6: Results for all participants regarding learning dynamics of layers 24, 25, 26 of LLM-jp exhibiting three phase transitions when using Japanese annotation and MMLU.

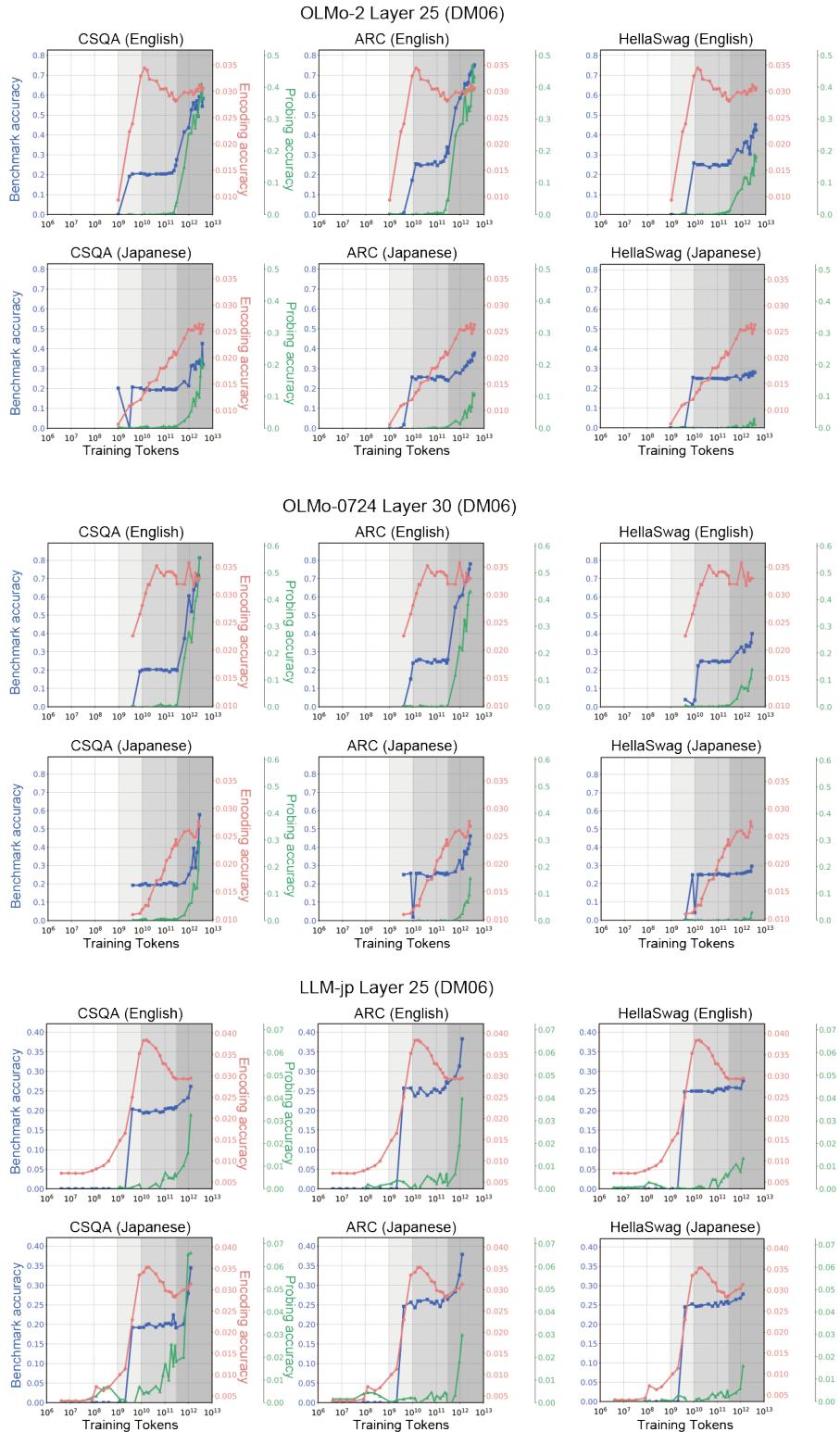


Figure B.7: Results from a single participant (DM06) for learning dynamics of layer 25 of OLMo-2 and LLM-jp, layer 30 of OLMo-0724 exhibiting three phase transitions when using other tasks.

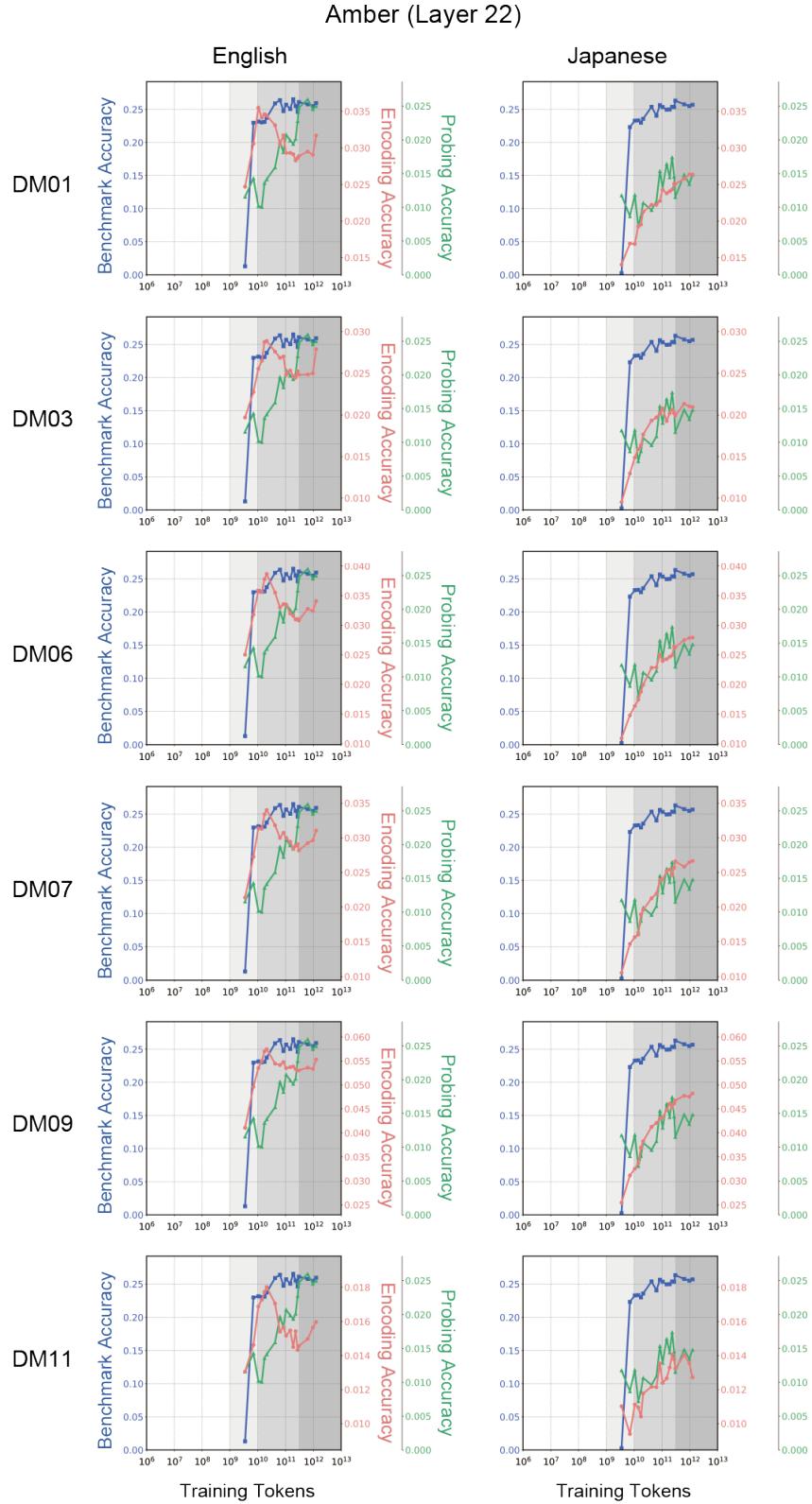


Figure B.8: Results for all participants regarding learning dynamics of layer 22 of Amber exhibiting three phase transitions when using English/Japanese annotation and MMLU.

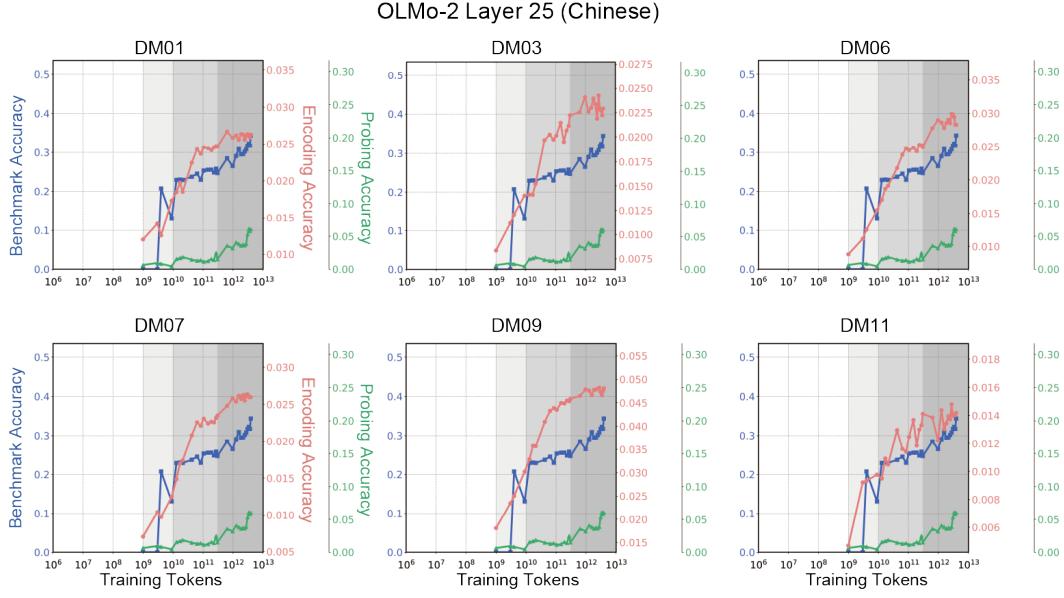


Figure B.9: Results for all participants regarding learning dynamics of layer 25 of OLMo-2 when using Chinese annotation and MMLU.

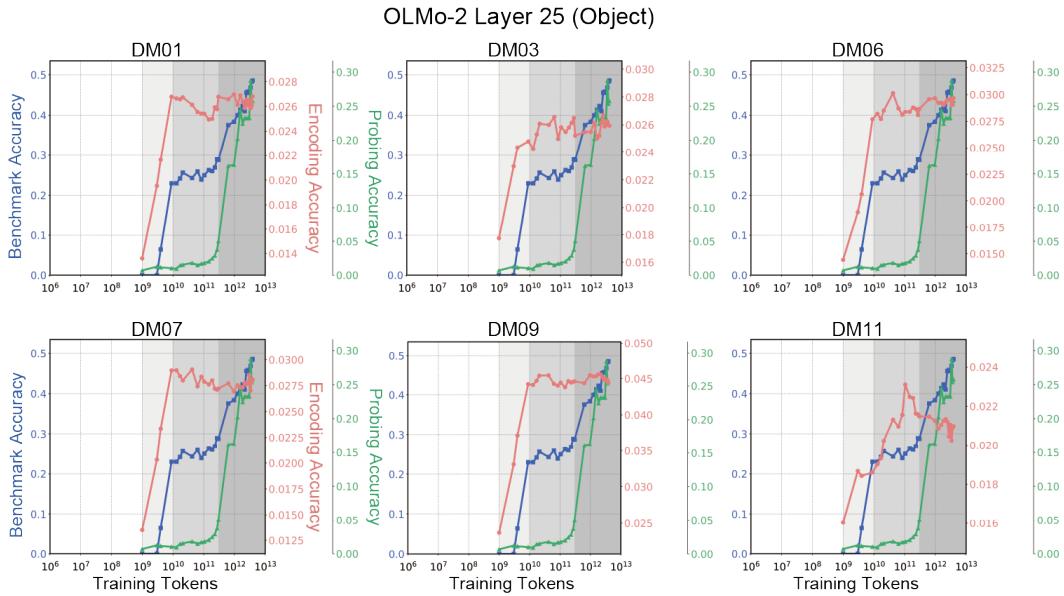


Figure B.10: Results for all participants regarding learning dynamics of layer 25 of OLMo-2 when using English Object annotation and MMLU.

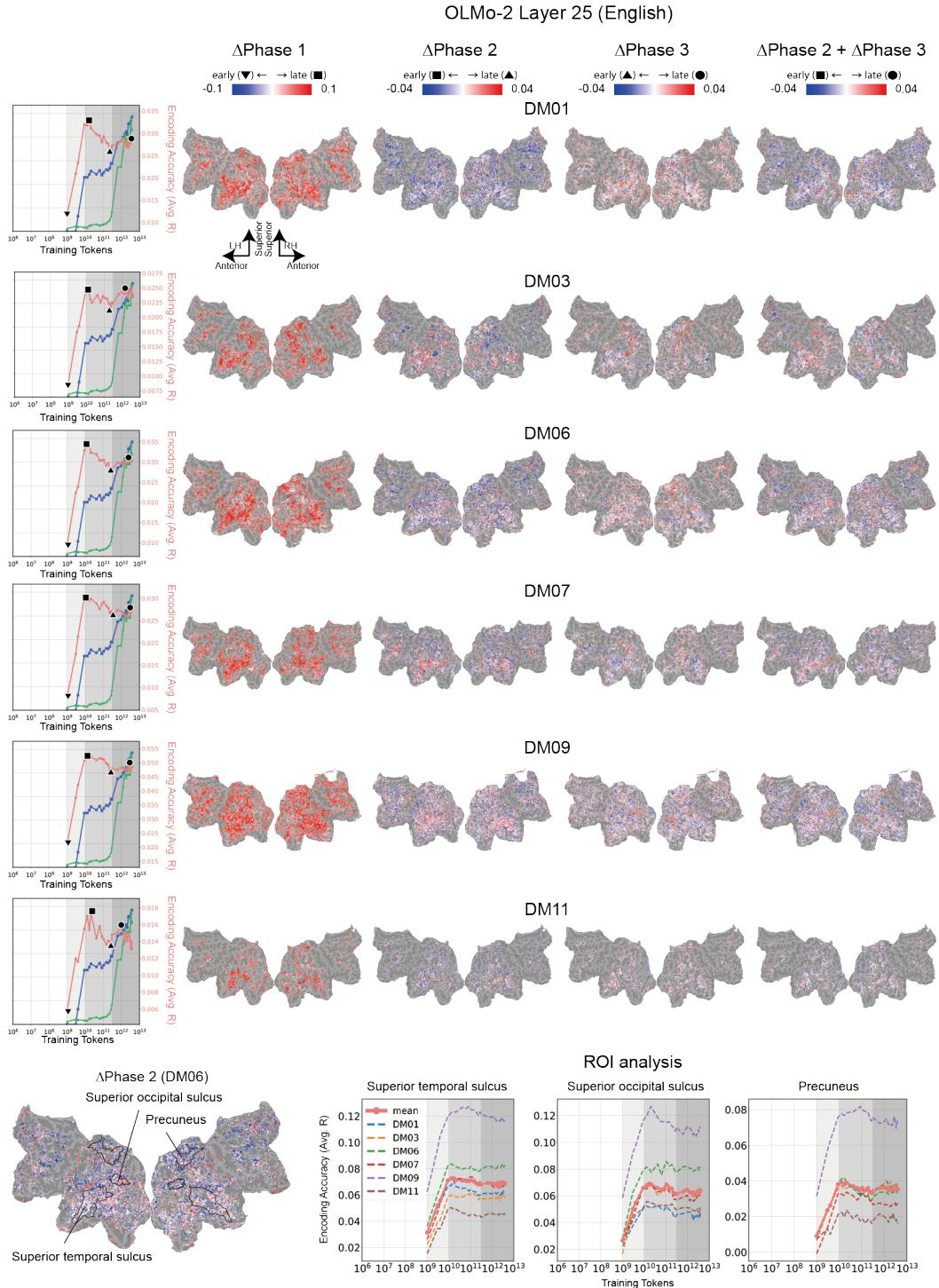


Figure B.11: Results for all participants regarding changes in the relationship with the brain using layer 25 of OLMo-2 and English annotation and MMLU. The line graph below shows the results of the ROI-level analysis. The vertical axis denotes the average encoding accuracy of significant voxels within each ROI. The solid line represents the mean across participants, and the dashed lines represent individual participant results. Black contours show several representative, anatomically defined ROIs.

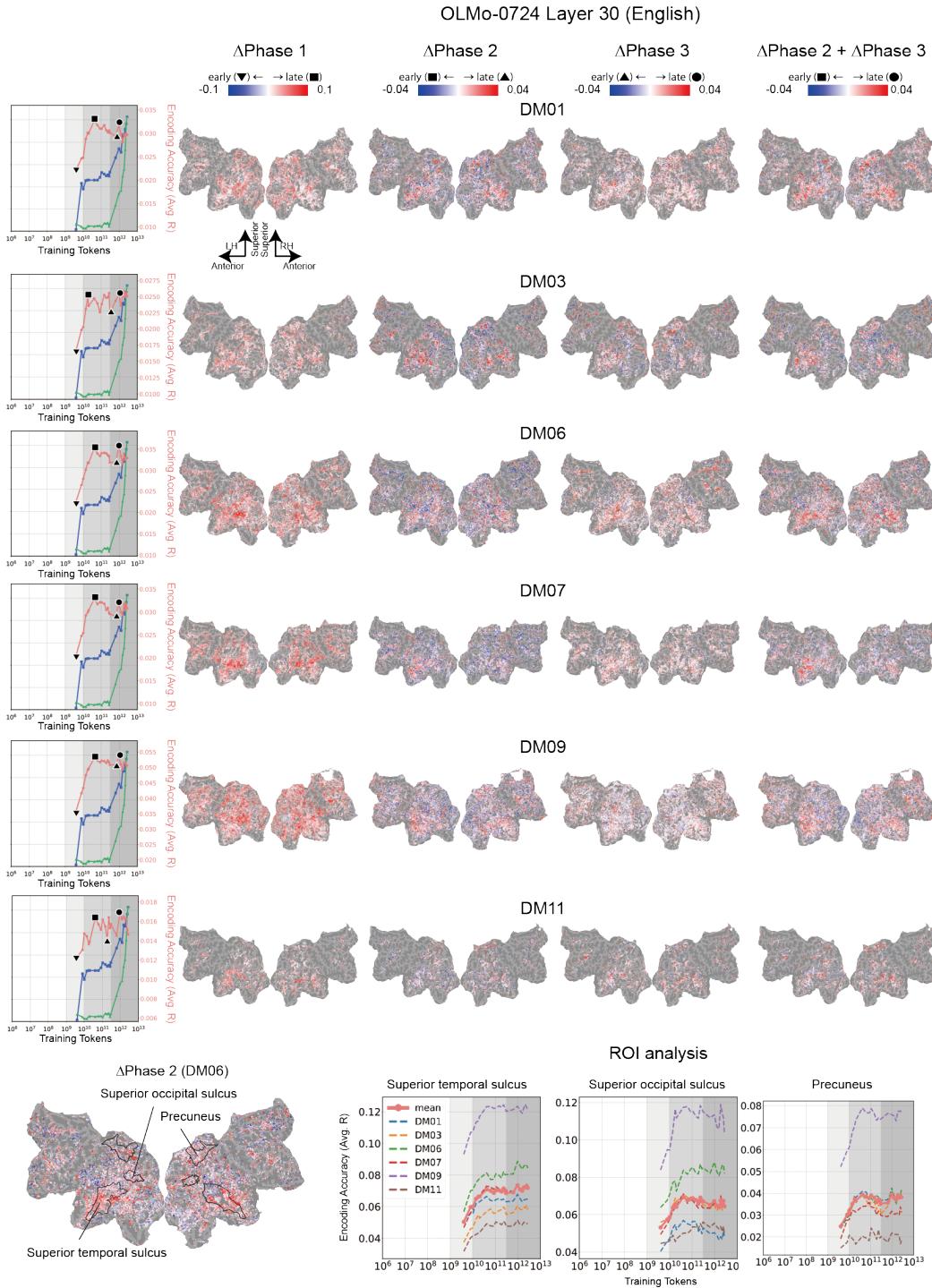


Figure B.12: Results for all participants regarding changes in the relationship with the brain using layer 30 of OLMo-0724 and English annotation and MMLU.

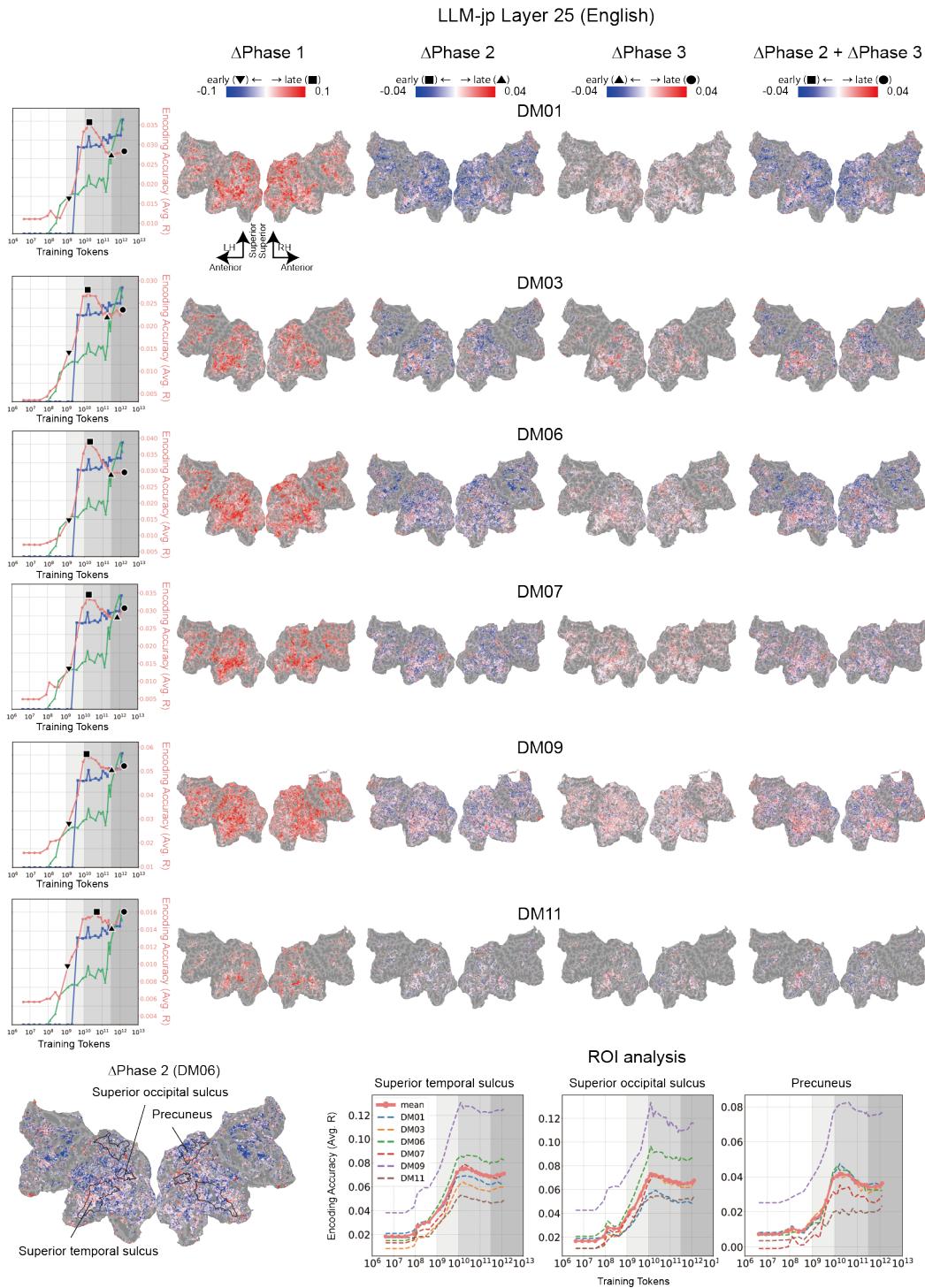


Figure B.13: Results for all participants regarding changes in the relationship with the brain using layer 25 of LLM-jp and English annotation and MMLU.

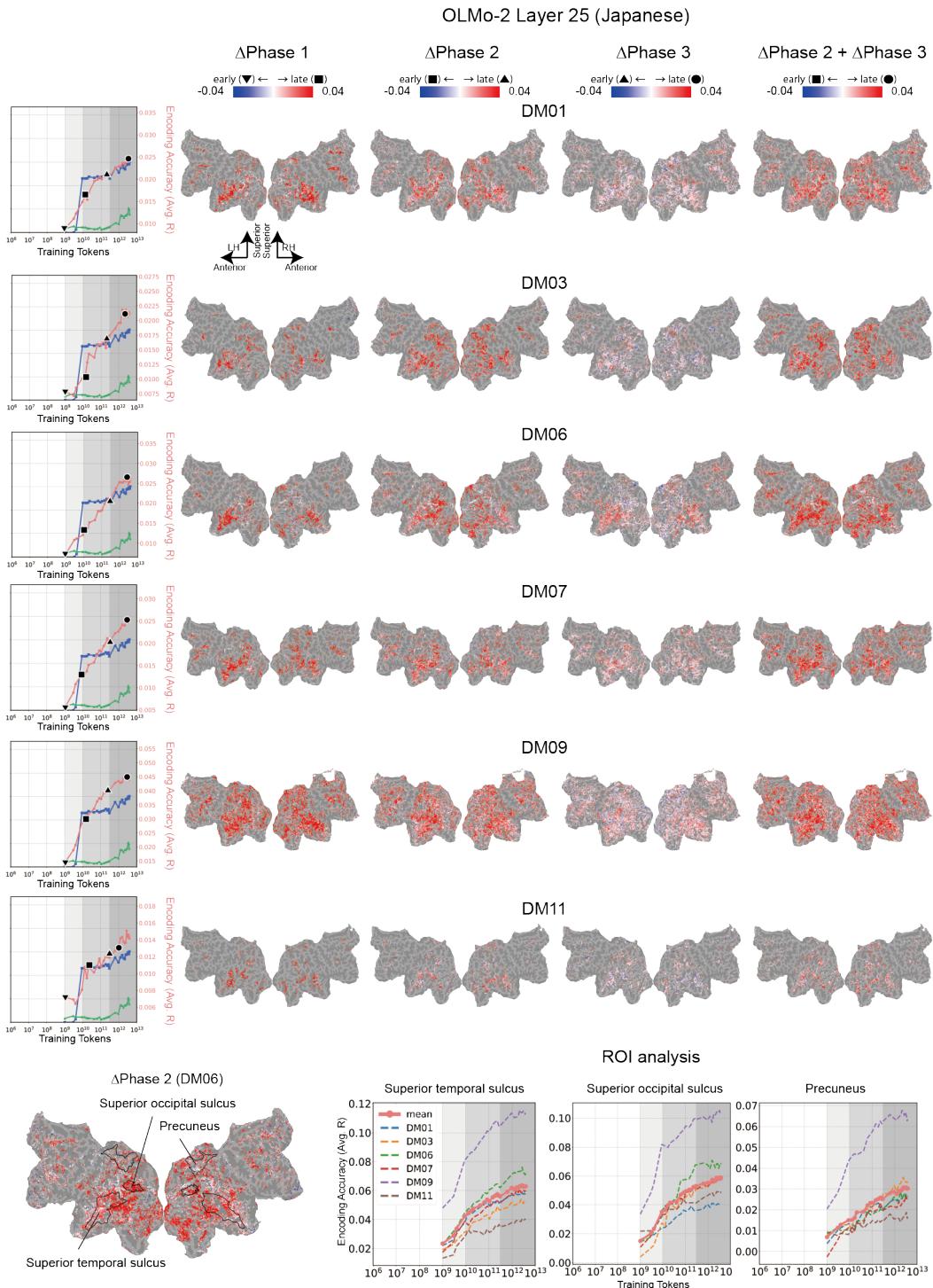


Figure B.14: Results for all participants regarding changes in the relationship with the brain using layer 25 of OLMo-2 and Japanese annotation and MMLU.

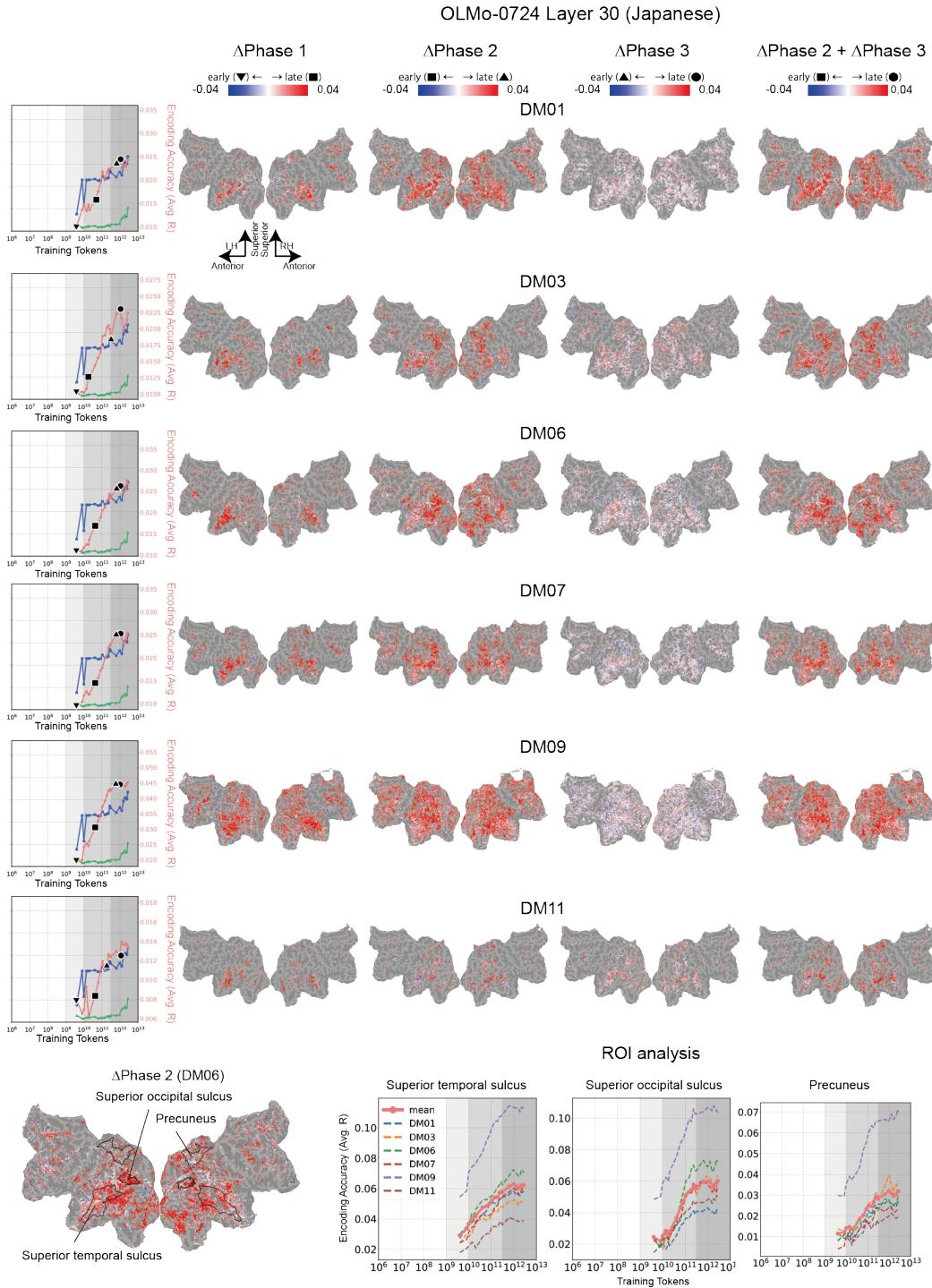


Figure B.15: Results for all participants regarding changes in the relationship with the brain using layer 30 of OLMo-0724 and Japanese annotation and MMLU.

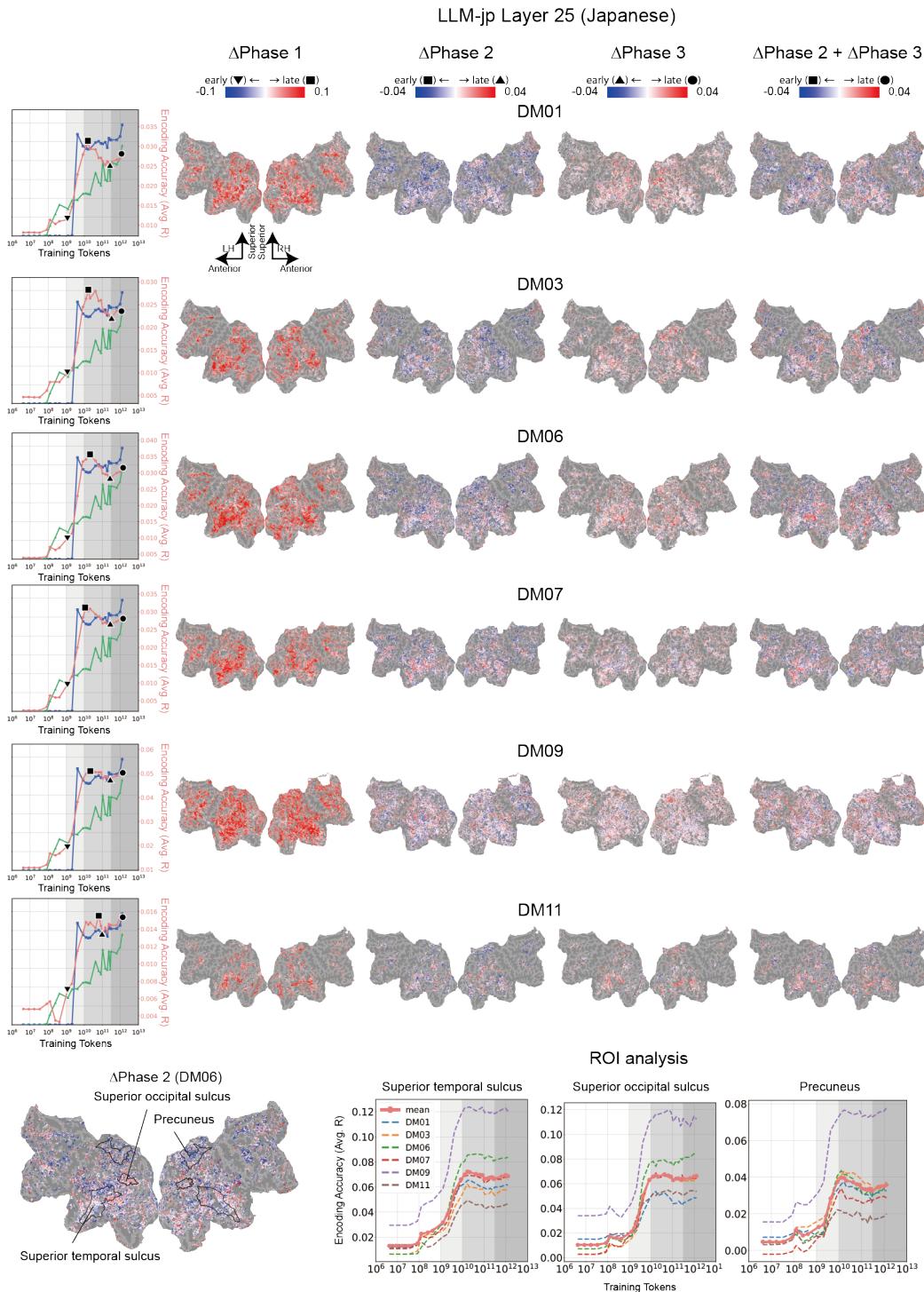


Figure B.16: Results for all participants regarding changes in the relationship with the brain using layer 25 of LLM-jp and Japanese annotation and MMLU.

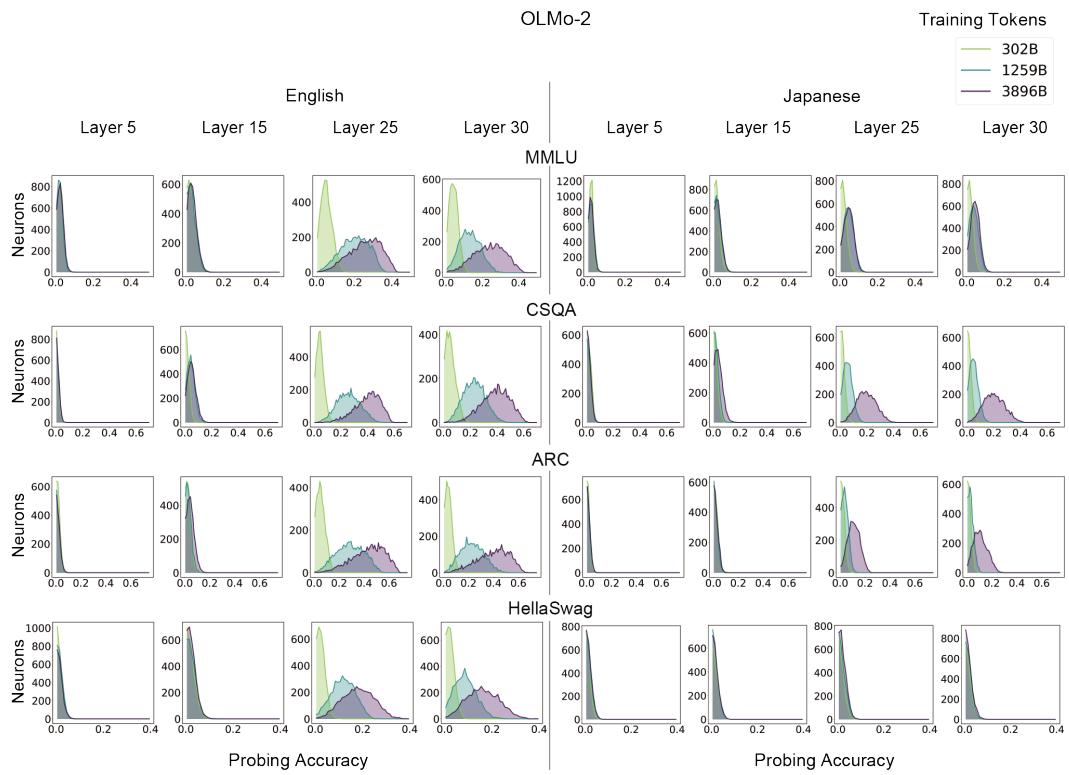


Figure B.17: Results for all downstream tasks regarding changes in the activations of OLMo-2.

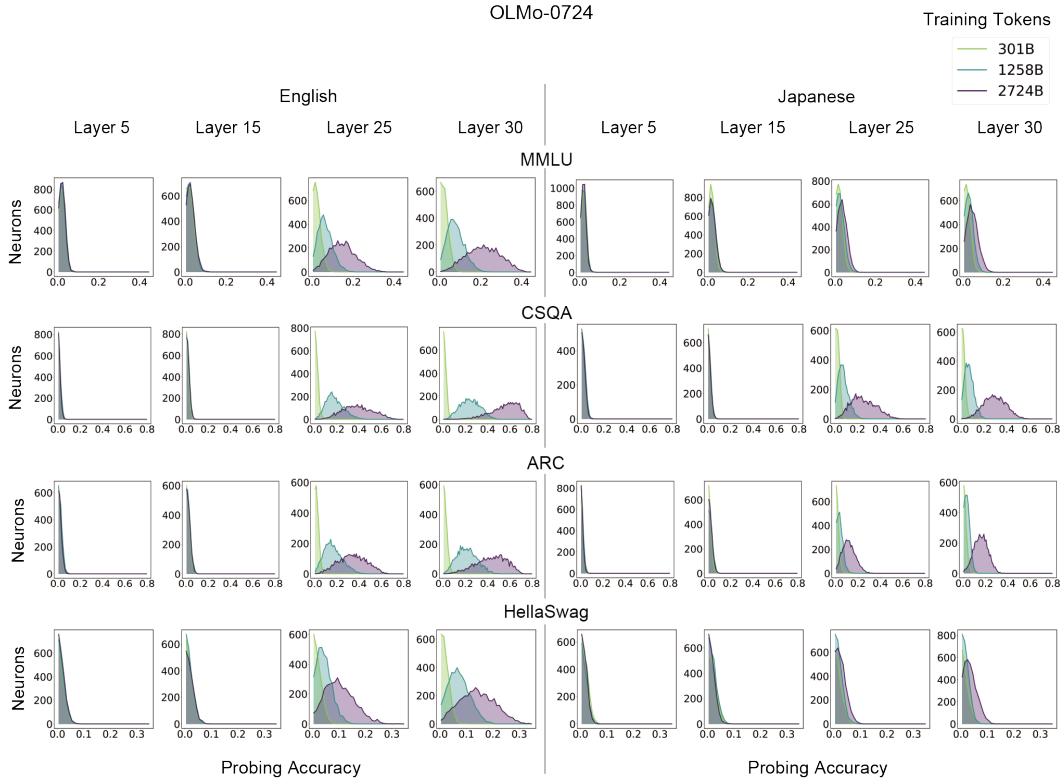


Figure B.18: Results for all downstream tasks regarding changes in the activations of OLMo-0724.

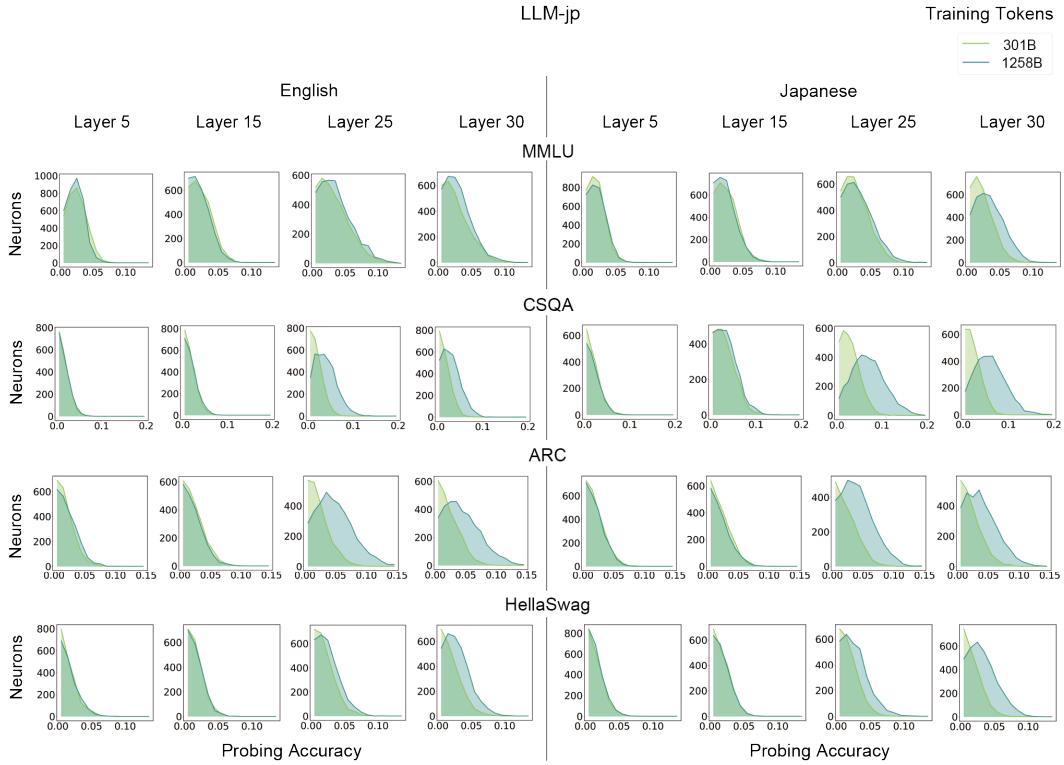


Figure B.19: Results for all downstream tasks regarding changes in the activations of LLM-jp.

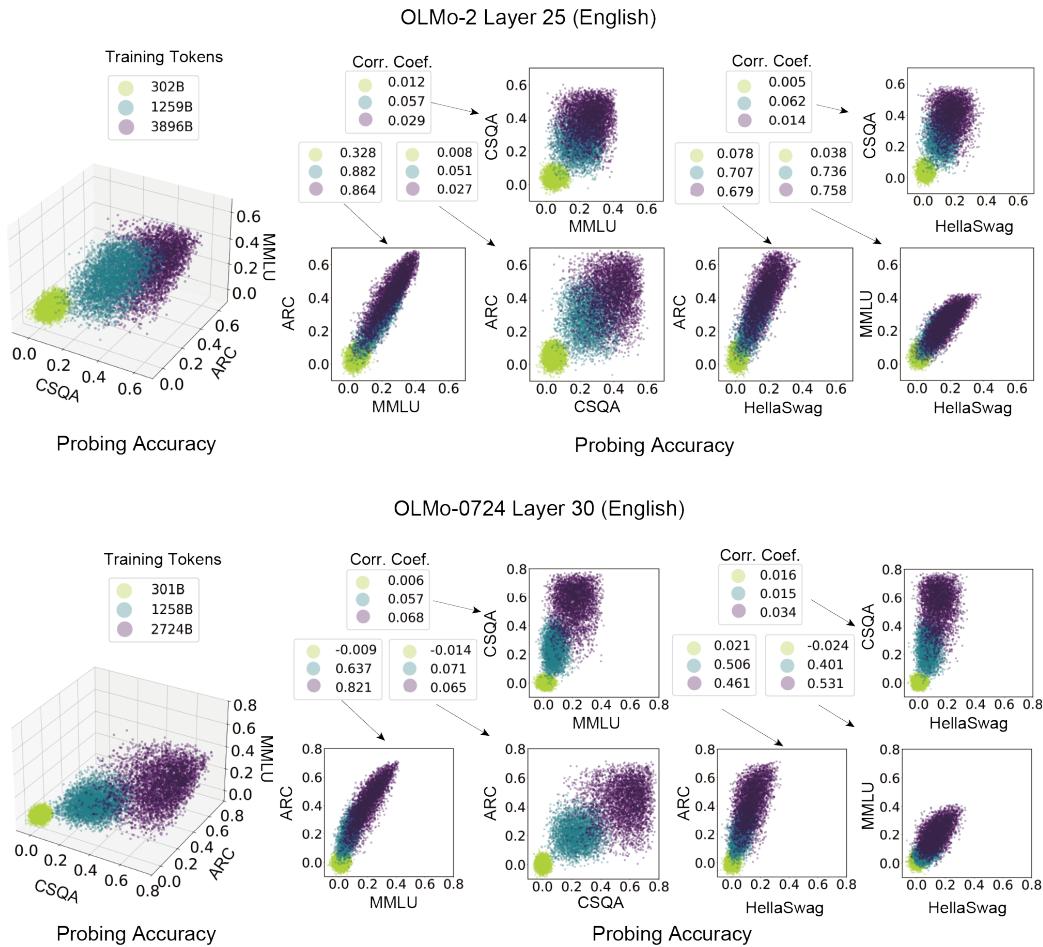


Figure B.20: Relationship between probing accuracies in OLMo-2 (layer 25) and OLMo-0724 (layer 30) across English MMLU, CSQA, ARC, and HellaSwag.

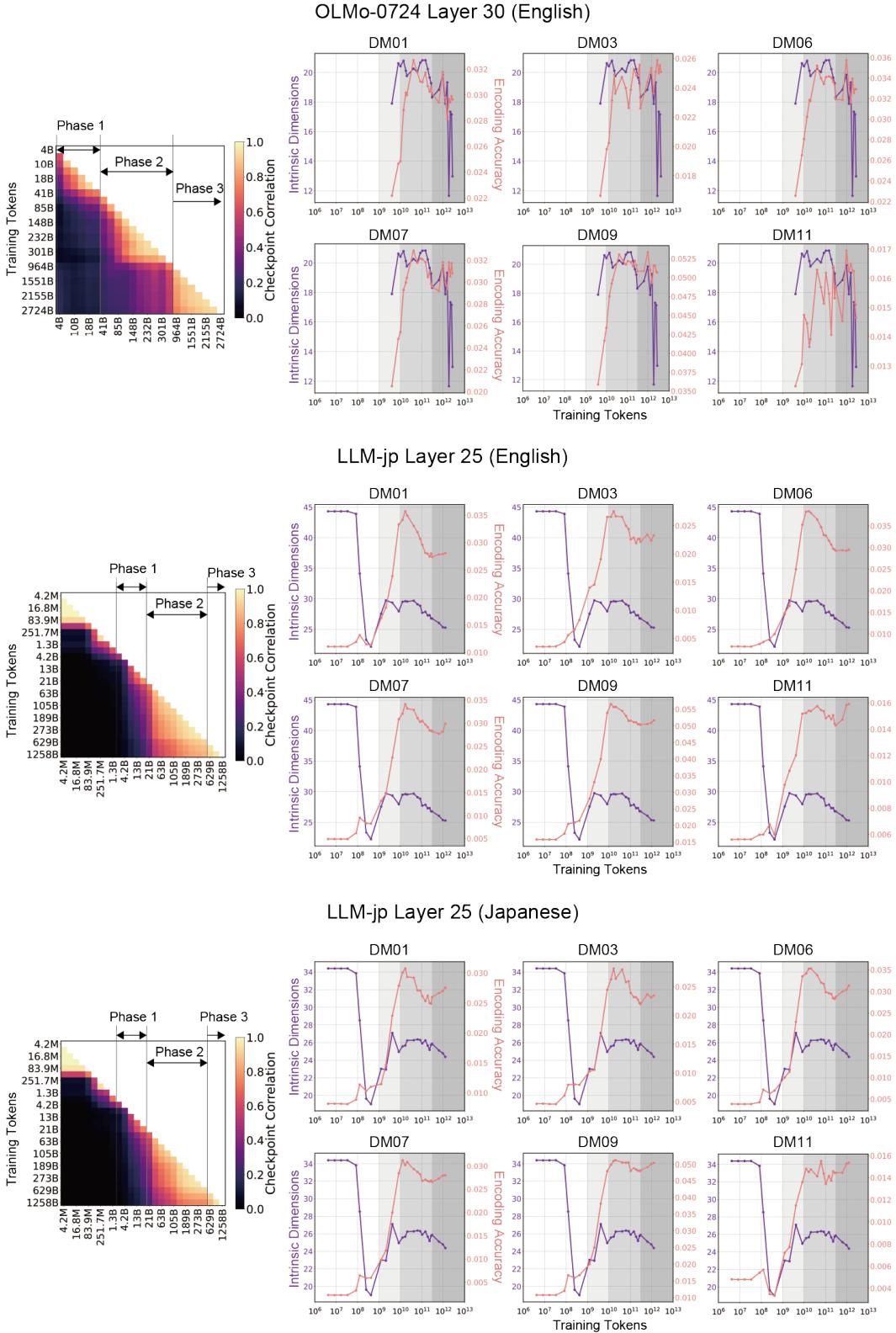


Figure B.21: Variations in correlation coefficients (left), encoding accuracy, and IDs (right) of the activations of OLMo-0724 (layer 30)/LLM-jp (layer 25) using learned languages across checkpoints.

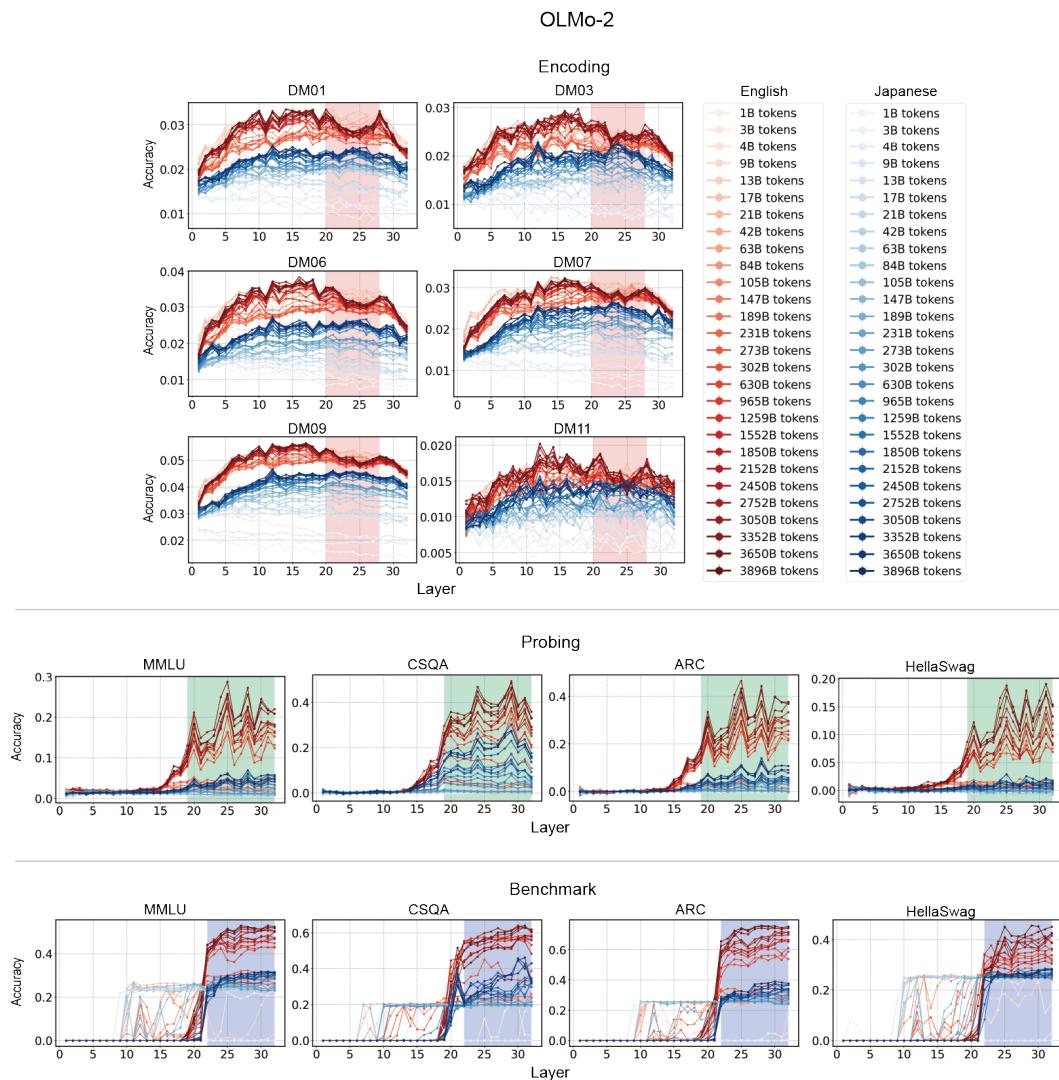


Figure B.22: Layers of interest for OLMo-2.

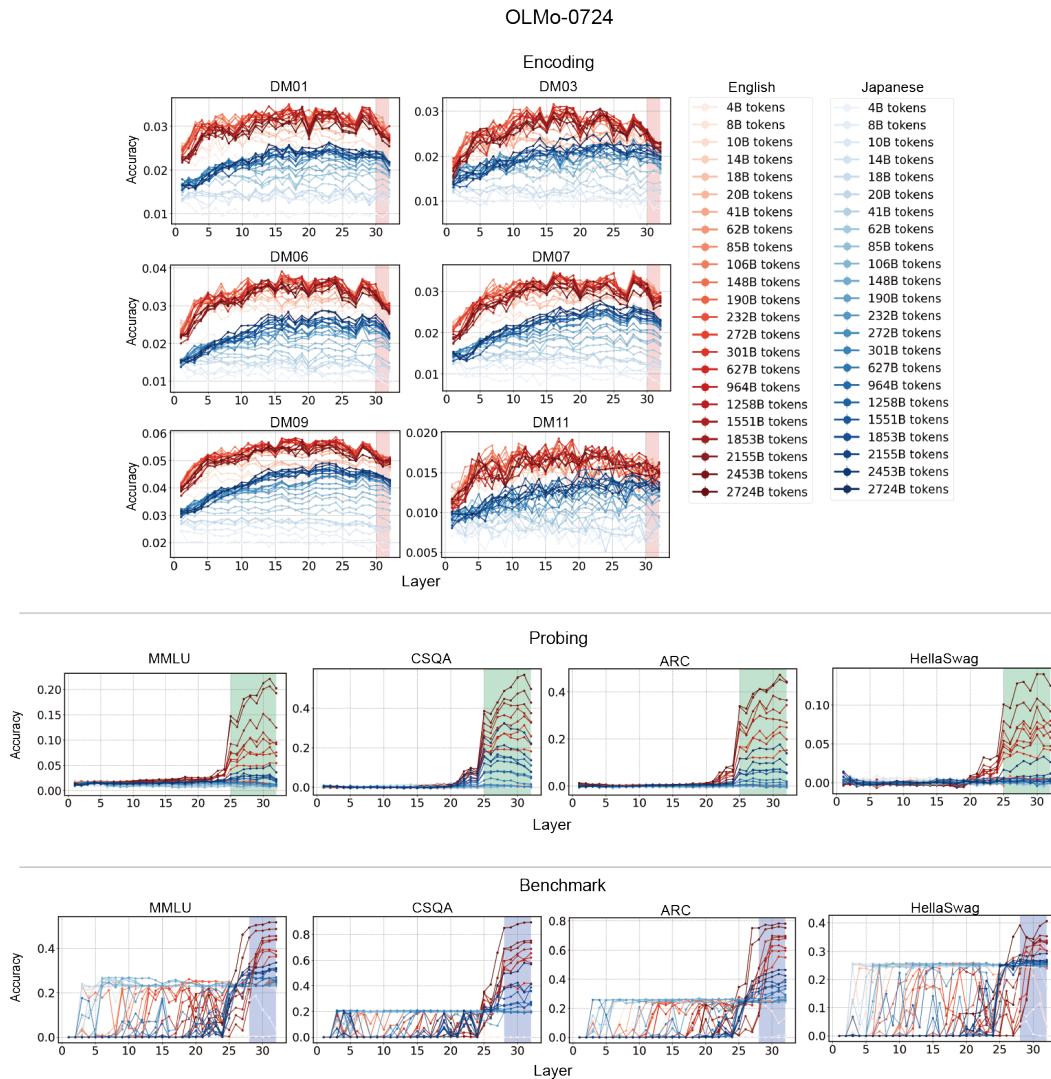


Figure B.23: Layers of interest for OLMo-0724.

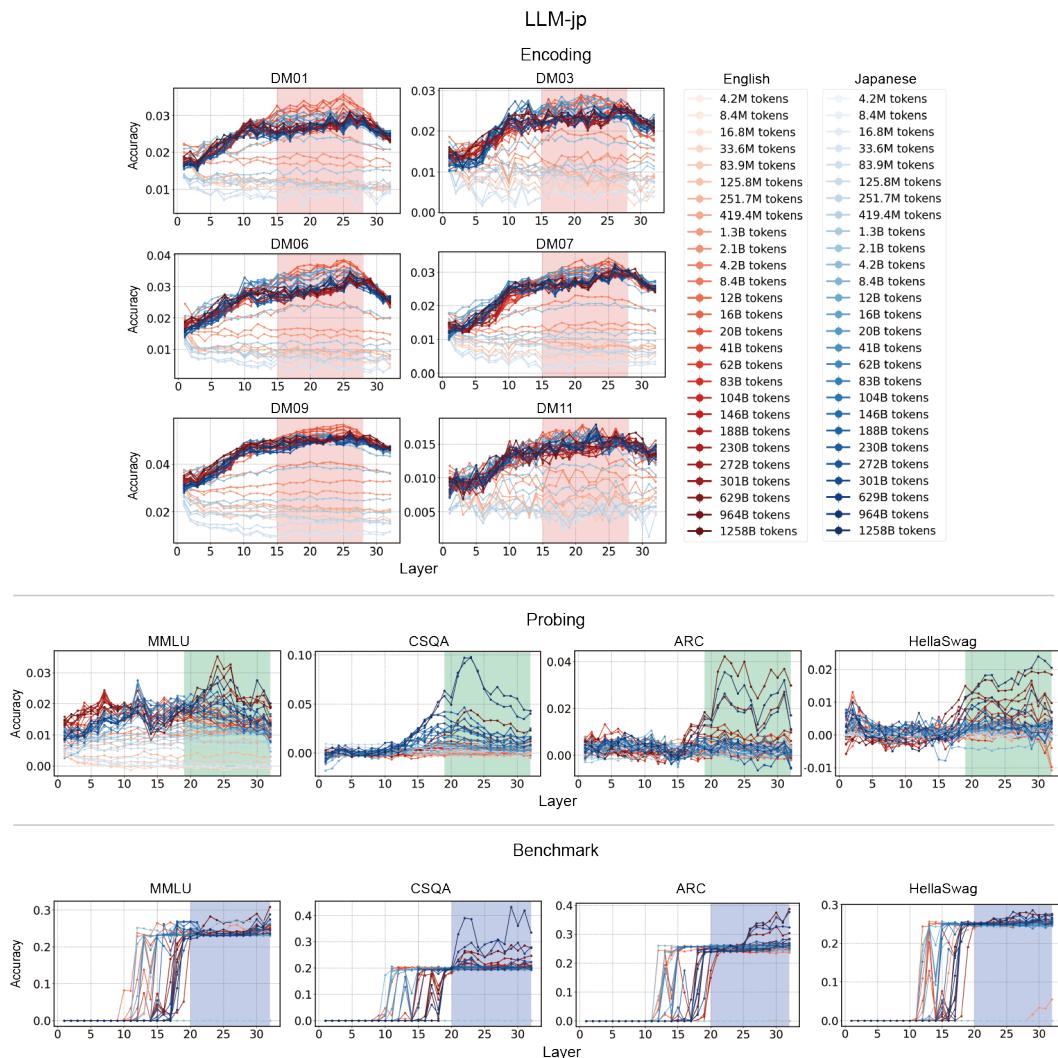


Figure B.24: Layers of interest for LLM-jp.