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

Anonymous ACL submission

Abstract

While large language models (LLMs) acquire diverse capabilities during training, internal learning dynamics remain poorly understood. To address this, we incorporate a neuroscientific perspective and analyze three interrelated dimensions: the similarity between LLMs and the human brain, the internal states of LLMs, and downstream task performance. Across models varying in data and architecture, we identify three phase transitions 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 LLM training, while opening new avenues for interdisciplinary research bridging AI and neuroscience.

1 Introduction

Large language models (LLMs) acquire distinct capabilities as model scale, training data, or training steps increase. Some capabilities improve smoothly with scale (Kaplan et al., 2020; Hoffmann et al., 2022; Tirumala et al., 2022; Xia et al., 2023; Zhang et al., 2024; Ren and Sutherland, 2025; Gadre et al., 2025; Qi et al., 2025; Bonnaire et al., 2025), whereas others emerge abruptly (Ganguli et al., 2022; Wei et al., 2022a; Du et al., 2024; Snell et al., 2024), or exhibit more complex, non-monotonic behavior (Huang et al., 2024). These behaviors have predominantly been discovered through evaluations of model outputs (e.g., downstream task performance), leading to numerous breakthroughs and strengthening our understanding of LLM training. Furthermore, recent efforts in mechanistic interpretability have begun to uncover the underlying internal changes that occur during training process (Olsson et al., 2022; Chen et al., 2023; Gopalani et al., 2024; Gopalani and Hu, 2025; Cui et al., 2024). Together, these lines of work examine capability acquisition from external and internal perspectives.

Neuroscience offers a complementary framework for understanding deep neural networks (DNNs), which were originally inspired by the human brain. By comparing internal representations of DNNs with human brain activity, prior studies have provided human-centered insights into model representations (Yamins et al., 2014; Güçlü and van Gerven, 2015), an approach recently extended to LLMs (Jain and Huth, 2018; Schrimpf et al., 2021; Goldstein et al., 2022; Oota et al., 2022; Tuckute et al., 2024a). For example, Antonello et al. (2023) revealed that brain activity aligns more strongly with larger-, higher-performance models. These findings highlight the potential of the human brain, an intricate system with diverse functions and the only system other than LLMs known to support rich linguistic abilities, to serve as a tool for model evaluation. A detailed review of related work is provided in Appendix A.

Despite this progress, key gaps remain. Learning dynamics in LLMs have rarely been studied from a neuroscience perspective, leaving the relationship between capability acquisition and brain alignment largely unexplored. In addition, prior studies that have examined learning dynamics have largely focused on isolated perspectives, making it unclear how these dynamics, and other relevant factors, jointly interact. Therefore, it remains unknown whether advances in LLM capabilities lead to increased alignment with the brain, decreased alignment, or more complex, non-monotonic trajectories (Figure 1a).

Here, we trace LLM learning dynamics by integrating three complementary analyses: *brain en-*

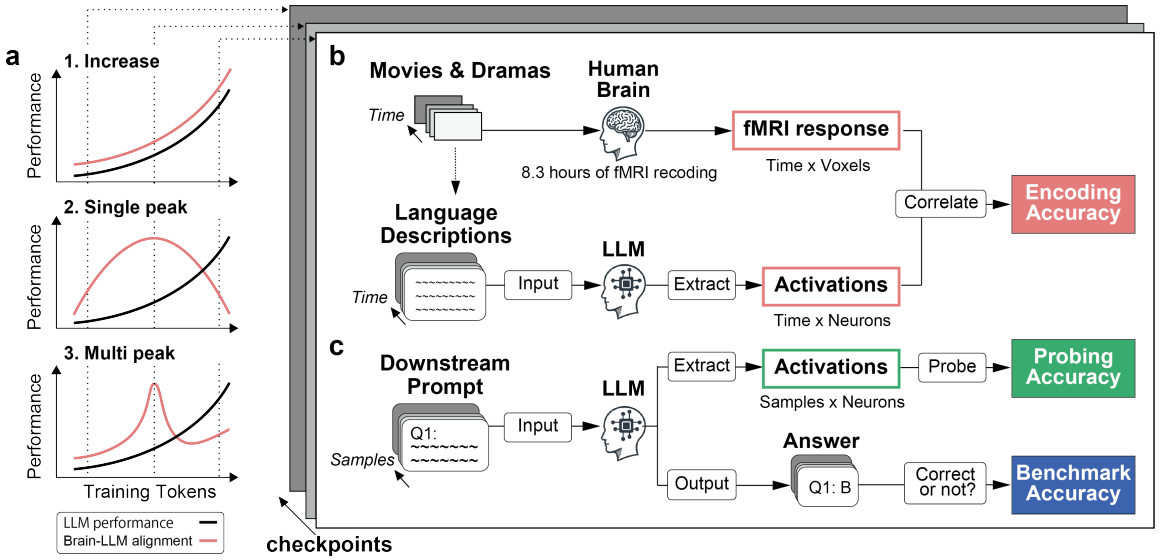

Figure 1: **Overview of the study. a** Schematic illustration of several hypothesized relationships between capability acquisition and brain alignment during LLM training (e.g., increasing, decreasing, non-monotonic). **b** *Brain encoding* analysis **c** *Probing* analysis (top) and *benchmark* analysis (bottom).

coding analysis, which assesses alignment with human brain activity; *probing* analysis, which detects shifts in internal representations; and *benchmark* analysis, which measures downstream task performance[1] (Figure 1b,c). 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***). These results show that LLM learning trajectories are more complex than previously suggested by monotonic scaling relationships.

2 Methods

2.1 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 additional 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 B.1 for further details on the models used. Because multilayer perceptron (MLP) layers account for most model parameters and are thought to encode essential features (Bereska and Gavves, 2024; Geva et al., 2021), we use MLP activations as internal representations.

2.2 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

[1]All experimental resources will be publicly available. See Sections B.1 and B.5 for the details.

activity during the training process. Specifically, for each checkpoint described in Sections 2.1 and B.1, we perform a **brain encoding analysis** (Naselaris et al., 2011; Nishimoto et al., 2011; Huth et al., 2012) by evaluating the prediction accuracy of a learned linear mapping from each layer's activations to brain activity (Figure 1b). 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 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 stimuli comprise nine video clips—eight international titles and one Japanese title—all presented in Japanese, with international titles dubbed accordingly. The dataset provides three types of natural-language annotations. 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 additional analysis, both in English and Japanese. See Section B.2 for additional dataset details. For evaluation, we use the final segment of each movie or series as test data (7,737 s in total), and the remaining sessions as training data (22,262 s). All prediction performance results are computed using the test dataset.

Model construction We extract LLM activations for each annotation from all hidden layers. Each time point comprises multiple tokens, and we average their activations. Because multiple annotations exist for each second, we extract the activations for each annotator and then average them across all annotators. We train L2-regularized linear models to predict voxel-wise fMRI responses from the corresponding LLM activations. Model weights are estimated on the training data and applied to the test data, with regularization parameters selected for each voxel via cross-validation. 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 B.3 for additional details about model construction. In the main analysis, we focus on later layers of each LLM: layer 25 for OLMo-2 and LLM-jp, and layer 30 for OLMo-0724. These layers exhibit the largest checkpoint-wise changes and are most strongly associated with phase-transition phenomena across all analyses (encoding, probing, and benchmark analyses; see Section B.7 for details). We confirm that adjacent layers show qualitatively similar patterns (see Figure C.5).

2.3 Probing and benchmark

As a complement to the brain encoding models described in Section 2.2, we use two additional families of measures to track learning dynamics: (i) probing analyses that quantify how downstream task–relevant information is represented in the LLMs' internal states, and (ii) benchmark performance on the same downstream tasks (Figure 1c). Both probing and benchmark scores are computed separately for each LLM checkpoint.

2.3.1 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 probing and benchmark analyses, we use 5-shot prompts in English and Japanese, excluding prompts that exceed each model's maximum context length. Additional details about the datasets are provided in Section B.4, and examples of the prompts are shown in Section B.8. For probing, each dataset is split into training and test sets in a 4:1 ratio, and all reported prediction accuracies are computed on the test set.

2.3.2 Probing for MLP activations

To assess how downstream task–relevant representations are acquired during training, we perform probing analyses on MLP activations across all hidden layers (Figure 1c, top). We feed the downstream task prompts into each LLM and extract the final-token activations from every layer. For each sample, we construct an answer matrix (samples $\times$

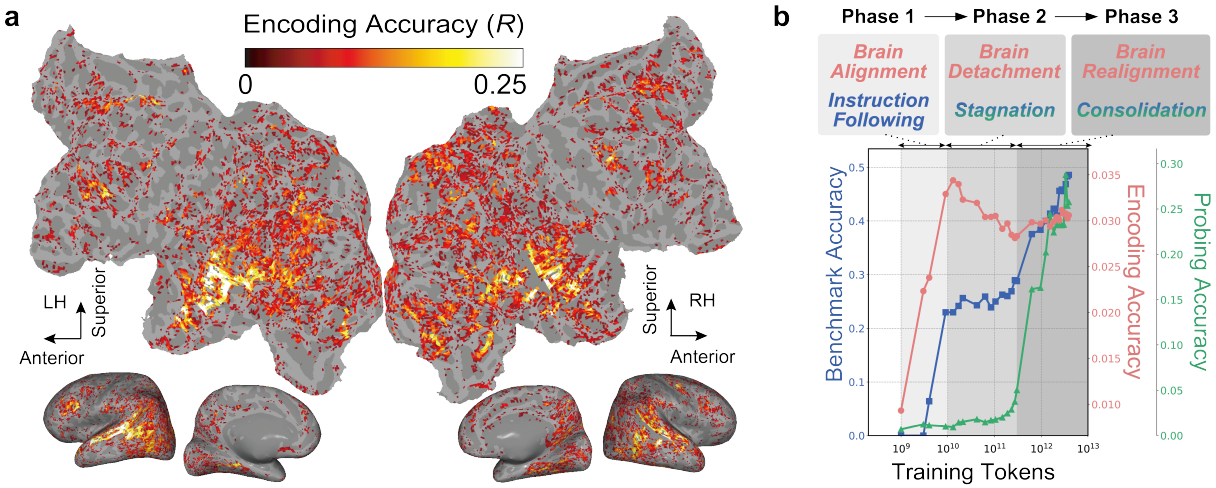

Figure 2: **Learning dynamics of LLMs exhibiting three phase transitions. a** Encoding accuracy for a single participant (DM06) when OLMo-2's final checkpoint are used, projected onto the inflated (both sides, lateral, and medial views) and flattened cortical surface (center, occipital areas are at the center), for both the left and right hemispheres. Brain regions with significant accuracy are colored (all colored voxels $P < 0.05$, FDR corrected). **b** Triple phase-transition phenomena during the learning process of OLMo-2, 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 checkpoints.

choices), with correct choices labeled as 1 and incorrect choices as 0. Our goal is to learn a mapping from this answer matrix to the observed activations across all layers. To this end, we fit L2-regularized linear regression models that transform the answer matrix for the training data into the corresponding activations, and then apply the learned weights to the test data. The regularization parameter is optimized via 4-fold cross-validation for each neuron. Probing accuracy is quantified as the Pearson correlation between predicted and observed activations, providing a neuron-level measure of task-relevant information at each checkpoint. We focus on the same set of layers as in the encoding analysis.

2.3.3 Benchmark performance

Finally, we directly evaluate how the ability to solve downstream tasks emerges during training (Figure 1c, bottom). For each LLM checkpoint, we measure performance on the downstream datasets described in Section 2.3.1. The evaluation metric is the fraction of correctly answered items, i.e., the proportion of samples for which the model's output exactly matches the correct answer.

2.4 Principal component analysis of estimated brain activations

To gain deeper insight into the observed phenomenon, we analyze how the encoding model weights evolve across checkpoints. For each participant and checkpoint, we first compute voxel-wise estimated brain activations for every sentence by multiplying the LLM activations for each of the 2,174 unique sentences by the learned encoding model weights. We then concatenate the estimated activations from all participants and checkpoints along the voxel dimension and apply PCA to this matrix along the sentence dimension. This procedure yields principal component (PC) scores for every voxel under each condition (participant $\times$ checkpoint), which we subsequently use to interpret the encoding model weights across checkpoints.

2.5 Direct analysis of MLP activations

To characterize how internal representations in LLMs evolve during training, we analyze changes in the correlation structure and dimensionality of MLP activations across checkpoints. Using the same dataset and layers as in the encoding analysis, we extract activations at each checkpoint and compute Pearson correlation coefficients between checkpoints. We further quantify representational dimensionality at each checkpoint using the intrinsic dimension (ID), a measure that has recently gained attention for characterizing LLM representations. IDs are estimated by applying the Generalized Ratios Intrinsic Dimension Estimator (GRIDE) (Denti et al., 2022) to layer-wise activations from the same dataset (see Section B.6 for

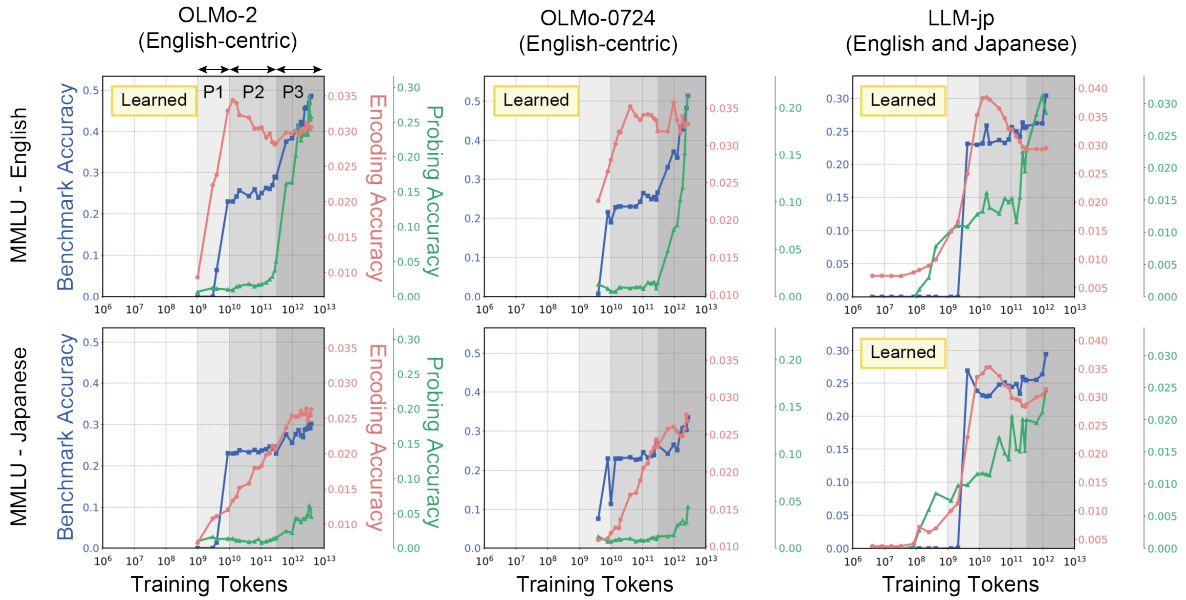

Figure 3: **Learning dynamics of LLMs exhibiting three phase transitions.** The horizontal axis denotes the number of training tokens. The vertical axis denotes the average encoding accuracy for all voxels of a single participant (DM06) (red lines), the benchmark accuracy (blue lines), and the average probing accuracy for all LLM neurons calculated using MMLU (green lines). For each model, we select the layer that captures the transitions occurring 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.

details).

3 Results

3.1 Triple phase transitions in LLMs

First, we evaluate whether LLM activations used in this study can predict brain activity. Figure 2a shows whole-brain voxel-wise encoding performance for OLMo-2 at the final checkpoint. The model predicts activity across a broad set of regions spanning high-level cognitive and sensory areas, confirming that OLMo-2 activations significantly account for brain responses.

Next, we examine how prediction accuracy evolves over training using the three analytical approaches. Figure 2b shows results for OLMo-2 with MMLU used for probing and benchmark analyses. Brain encoding accuracy reflects the average voxel-wise prediction performance from LLM activations in the selected layer, whereas probing accuracy reflects the average prediction accuracy across LLM neurons. Together, these metrics reveal three distinct phase transitions in later layers of OLMo-2 during training. The first transition emerges after around 10^9–10^{10} training tokens, where both the encoding and benchmark accuracy suddenly surge. The rise in benchmark accuracy (blue) indicates improved downstream task performance and the onset of instruction following, while the simultaneous increase in encoding accuracy (red) reflects enhanced whole-brain alignment. We refer to this stage as the ***Brain Alignment and Instruction Following*** phase. The second transition arises after around 10^{10}–$3{\cdot}10^{11}$ tokens, where the benchmark accuracy stagnates. Strikingly, encoding accuracy **declines** during this period, indicating reduced brain–LLM alignment. We therefore label this stage the ***Brain Detachment and Stagnation*** phase. The third transition occurs beyond approximately $3 \cdot 10^{11}$ tokens, where benchmark and probing accuracy increase sharply, accompanied by modest recovery in encoding accuracy. At this point, the increase in the benchmark accuracy suggests that the LLM gradually acquire the capability to solve tasks, whereas the change in the brain encoding accuracy implies a renewed enhancement in alignment with the brain. We thus refer to this stage as the ***Brain Realignment and Consolidation*** phase.

We confirm the triple phase transition in encoding accuracy using two independent indicators. First, the number of significant voxels follows the same transition pattern (Figure B.1). Second, PCA applied to a voxel-by-checkpoint encoding accuracy matrix reveals that PC1 scores across

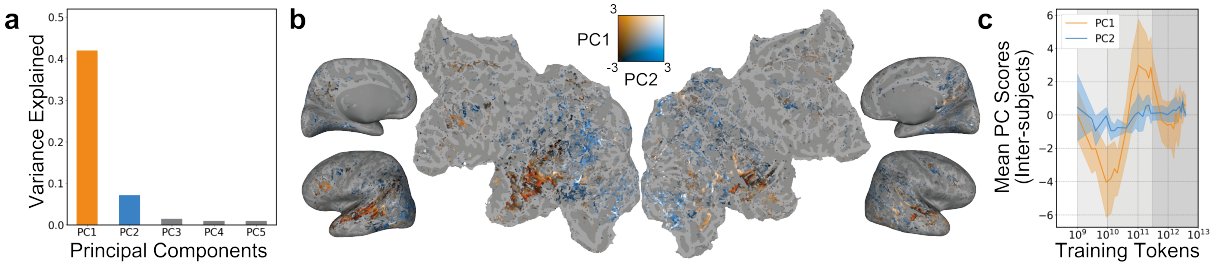

Figure 4: **Dynamics of brain regions involved in the three phases. a** Explained variance of the top five PCs of the estimated activations derived from OLMo-2's encoding model. **b** PC1 and PC2 scores at the final checkpoint mapped onto DM06's cerebral cortex; only voxels significant across all checkpoints are colored (PC1 in orange, PC2 in blue). **c** Mean PC1 and PC2 scores across checkpoints for voxels significant at all checkpoints. Scores were averaged within participants and then across participants; shaded areas indicate standard deviations.

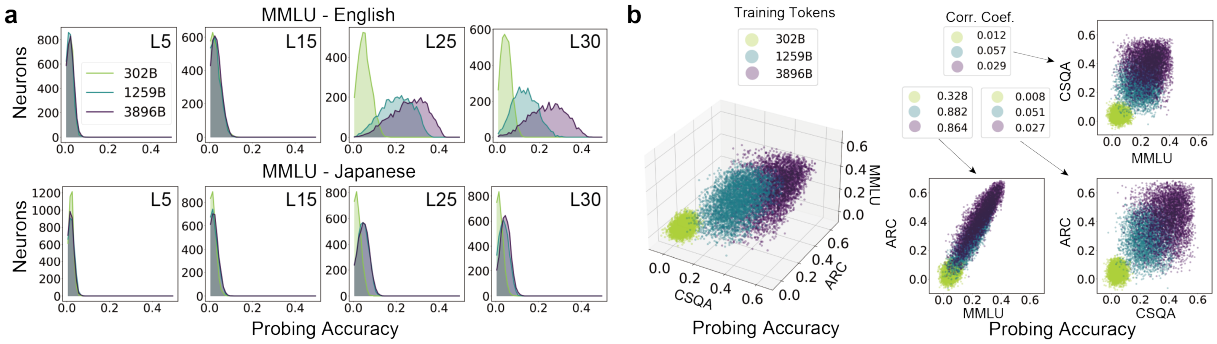

Figure 5: **Evolution of LLM internal representations. a** Evolution of probing accuracy for English and Japanese MMLU over training, assessed at layers 5, 15, 25, and 30 of OLMo-2. The horizontal axis shows probing accuracy, and the vertical axis shows the number of neurons in each 0.01 accuracy bin. The legend indicates the number of training tokens. **b** Relationship among per-neuron probing accuracies in OLMo-2 (layer 25) for English MMLU, CSQA, and ARC. Each axis denotes the probing accuracy for one task, and the color gradient reflects the number of training tokens. The legend shows the correlation coefficients for each task pair.

checkpoints exhibit consistent triple-phase dynamics (Figure B.2).

We examine whether this three-stage phase transition also emerges in other LLMs. Figure 3 presents the results for OLMo-2, OLMo-0724, and LLM-jp when using MMLU for the probing and benchmark analyses. 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. Section C.1 provides analogous results for adjacent layers, other downstream tasks (CSQA, ARC, HellaSwag), other LLMs (Amber), other languages (Chinese), and other annotations.

3.2 What happens during each phase?

Section 3.1 characterized the three phase transitions that arise during LLM training using encoding, probing, and benchmark analyses. However, the internal changes underlying each phase remain unclear. To address this, we analyze voxel-level changes in brain alignment and neuron-level shifts in LLM representations across phases.

3.2.1 Neural functional structure associated with phase transitions

In the previous sections, our analyses focused solely on the brain-wide average behavior associated with the phase transitions. We now characterize how different brain areas contribute to this effect. To this end, we predicted voxel-wise responses to the text stimuli using the estimated weights for each voxel, participant, and checkpoint, and then applied PCA to extract components of temporal change, which were subsequently mapped back onto the brain. This analysis included only voxels that exhibited significant prediction accuracy for each participant at each checkpoint.

n the English setting of OLMo-2, PC1 and PC2 explain most of the variance (PC1: 0.417, PC2: 0.072; Figure 4a). Mapping voxel-wise PC scores at the final checkpoint (Figure 4b) shows that PC1 reflects regions associated with semantic

and higher-order cognition (e.g., auditory cortex, frontal lobe, precuneus), whereas PC2 reflects associative and higher visual regions (e.g., parietal lobe, higher visual cortex). Across checkpoints, PC2 remains stable, while PC1 closely tracks the triple phase transitions (Figure 4c). These results suggest that semantic and higher-order cognitive regions primarily drive the observed phase transitions. Further details for all participants and LLMs can be found in Figures B.13–B.15.

3.2.2 Evolution of LLM internal representations for downstream tasks

Having examined voxel-level brain alignment, we next investigate how the triple phase transitions relate to internal representations supporting downstream task performance. We apply probing analysis, a standard tool in mechanistic interpretability that uses simple readout models to assess what information neurons encode (Elhage et al., 2021; Dai et al., 2022; Bereska and Gavves, 2024). Specifically, we analyze the distributions of per-neuron probing accuracy at the early, middle, and final checkpoints of Phase 3 (302B, 1259B, 3896B training tokens in OLMo-2). Figure 5a presents the results for neurons in layers 5, 15, 25, and 30 of OLMo-2 on MMLU. In the English setting, probing accuracy increases progressively in later layers, indicating the emergence of specialized activations as downstream task proficiency improves. Consistent patterns are observed across all downstream tasks and models (OLMo-2, OLMo-0724, and LLM-jp; Figures B.16–B.18).

We further examine whether neurons specialize for distinct downstream tasks, focusing on the same checkpoints as in Figure 5a. Figure 5b illustrates the relationships among the per-neuron probing accuracies of OLMo-2 for the English MMLU, CSQA, and ARC. As training progresses, some neurons become highly informative for all three tasks, whereas others remain selective for only a subset. Per-neuron accuracies for MMLU and ARC are highly correlated (r = 0.864), whereas CSQA shows almost no correlation with either task (r = 0.029 with MMLU, r = 0.027 with ARC). OLMo-0724 exhibits a similar pattern (Figure B.19), indicating substantial task-specific representational changes in multiple models. Additional results for all downstream tasks, including HellaSwag, are reported in Section C.3.

Overall, sharply task-specific neuron-level representations emerge primarily during Phase 3—rather than Phase 1, when instruction-following first appears. This stage coincides with a shift in LLM representations back toward brain alignment, suggesting a close link between task-specific representational specialization and changes in brain alignment.

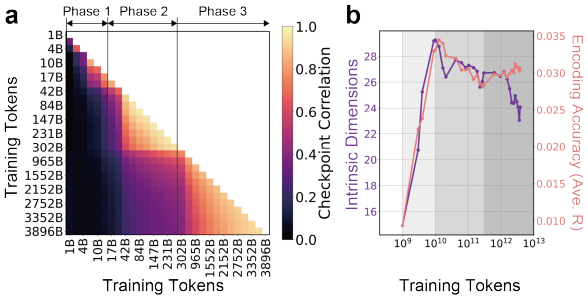

Figure 6: **The Nature of Activations. a** Variations in correlation coefficients of the activations of OLMo-2 across checkpoints. **b** IDs (purple line) and average encoding accuracy for all voxels of a single participant (DM06) (red line) across checkpoints.

3.3 Changes in the nature of activations

Based on previous 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 6a, 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 6b 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 (Antonello and Cheng, 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.

4 Discussion and conclusions

In this study, we quantitatively characterize LLM learning dynamics from three perspectives: alignment between LLMs and the brain, internal states

associated with downstream tasks, and downstream task performance. Integrating these analyses reveals a robust triple phase transition, with consistent timing across models despite substantial differences in training data and architecture. Our results show that human brain activity provides a biologically grounded reference for tracking how LLM capabilities emerge and consolidate during training, with implications for interpretability and safety.

While Antonello et al. (2023) reported a rise in alignment with increasing model size, our findings depict a a non-monotonic trajectory: an initial surge in brain alignment, a decline, and a later resurgence. This suggests that LLMs may adopt distinct computational strategies at various stages, rather than gradually acquiring brain-like language representations. For example, AlKhamissi et al. (2025b) tracked Pythia checkpoints up to 300B tokens and found that brain alignment is more tightly coupled with formal than functional linguistic competence. This is partially consistent with the alignment increase we observe in the *Brain Alignment and Instruction Following* phase, while our post-300B *Brain Realignment and Consolidation* phase captures a later shift not accessible in prior work and may be linked to the maturation of broader functional competence. Furthermore, Caucheteux and King (2022); Antonello et al. (2023); Antonello and Huth (2024); Antonello and Cheng (2024) have shown that, in trained models, the layers or representations with the highest next-word prediction performance do not necessarily align best with the brain; layers with somewhat lower prediction performance can show stronger brain alignment. If language-model performance is treated as a proxy for training progress, these observations correspond naturally to dynamics up to Phase 2 in our framework. Our results extend this picture by identifying an additional state in which brain alignment re-emerges alongside consolidated task competence.

By probing neuron-level activations, we also clarify how task-relevant representations emerge and are shared across tasks during training. Consistent with prior work identifying neurons that encode specific linguistic features (Tenney et al., 2019; Dai et al., 2022; Wang et al., 2022; Gurnee et al., 2023), we find that neurons making strong task-specific contributions appear abruptly in later layers and that patterns of neuron sharing vary systematically across tasks. These results extend earlier analyses focused on fully trained models by revealing how such structures develop over time.

The *Brain Alignment and Instruction Following* and *Brain Realignment and Consolidation* phases parallel previously reported emergent phenomena in LLMs, in which new capabilities arise abruptly once training surpasses critical thresholds Wei et al. (2022a,b); Caballero et al. (2023); Olsson et al. (2022). While prior work has primarily emphasized output behavior or isolated mechanisms, our results show that changes in brain alignment, internal representations, and task performance progress together. These phenomena are most pronounced for languages that are well represented in the training data, suggesting that deep internalization of their statistics can trigger large shifts in both brain alignment and behavior. Although the extent to which LLM capabilities are truly emergent remains debated (Schaeffer et al., 2023), our findings indicate that training dynamics exhibit substantial non-monotonic and phase-like structure.

We also extend the findings of Antonello and Cheng (2024), who showed a positive correlation between the dimensionality of LLM activations in trained models and their alignment with brain activity. Our results confirm that this correlation remains robust during training, suggesting that the observed phase transitions are tied to changes in activation dimensionality. At the same time, these metrics primarily capture large-scale transformations in representation space and do not by themselves pinpoint the underlying mechanisms. Relating them to both brain alignment and downstream task behavior, as in this study, provides a more integrated and nuanced picture of how these transformations unfold.

In summary, we identify a triple phase transition during LLM training that jointly governs brain alignment, internal task-related representations, and task performance. Brain–LLM alignment does not increase monotonically, but instead follows an early surge, a subsequent decline, and a later resurgence as downstream competence consolidates. Brain alignment, probing accuracy, and intrinsic dimensionality all track a model's position along this trajectory, suggesting their use as in-process diagnostics for adapting training schedules and interventions. More broadly, using human brain activity as a biologically grounded reference reveals how internal representations and behavior evolve together, pointing toward new strategies for developing more interpretable and human-aligned language models.

Limitations

In this study, we examined the learning dynamics of three LLMs and identified a three-stage phase transition. These models included the latest publicly available training checkpoints and revealed newly emergent behaviors. However, certain state-of-the-art LLMs with superior language capabilities remain closed-source, rendering them difficult to analyze. These state-of-the-art LLMs may potentially exhibiting additional phase transitions or significant shifts in linguistic acquisition following the dynamics we observed. Therefore, investigating the learning processes of these models is crucial to advancing LLM safety and reliability.

Our research further examined the learning dynamics of LLMs across four different downstream tasks, all of which demonstrated similar patterns. Nonetheless, we have yet to cover every task necessary for an exhaustively evaluation of LLM performance. To more precisely capture the latent changes underlying the acquisition of language ability in LLMs more precisely, future work should include analyses on a broader spectrum of tasks.

From neuroscience perspective, this study interpreted the learning dynamics of LLMs by drawing comparisons with the adult brain. Meanwhile, it is also important to compare the learning trajectories of LLMs with those of human development. Thus, one particularly exciting direction for future work would involve comparing LLM trajectories with human developmental data (Nakai et al., 2024), a step that may further illuminate these phase transitions and enhance model interpretability.

In addition, the present results are based on a single fMRI dataset acquired during naturalistic video viewing, and thus the generalizability of our findings to other experimental paradigms and datasets remains to be established. Future work should therefore examine whether the observed phase-transition patterns are consistently reproduced across multiple datasets and task settings. At the same time, it is important to note a distinctive characteristic of the dataset used in this study, namely the availability of detailed *Story* annotations. Notably, compared with the object-level representations reported in Section C.1, the phenomena of the phase transition tended to be more prominent when analyses focused on story-level representations. This observation is consistent with prior studies reporting that story-based stimuli yield the strongest predictability of brain activity among various linguistic and perceptual features (Nakagi et al., 2024). Taken together, these considerations suggest that the evolutionary dynamics of LLM–brain alignment may be particularly tied to higher-order semantic and narrative processing. Accordingly, datasets that emphasize such high-level information processing, rather than low-level sensory features alone, may be especially well suited for investigating the dynamics of LLM–brain interactions.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *Preprint*, arXiv:2305.13245.

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2025a. The LLM language network: A neuroscientific approach for identifying causally task-relevant units. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10887–10911.

Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Osama A Binhuraib, Antoine Bosselut, and Martin Schrimpf. 2025b. From language to cognition: How LLMs outgrow the human language network. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24332–24350.

Richard Antonello and Emily Cheng. 2024. Evidence from fMRI supports a two-phase abstraction process in language models. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*.

Richard Antonello and Alexander Huth. 2024. Predictive coding or just feature discovery? an alternative account of why language models fit brain data. *Neurobiology of Language*, 5(1):64–79.

Richard Antonello, Aditya Vaidya, and Alexander G Huth. 2023. Scaling laws for language encoding models in fMRI. *arXiv [cs.CL]*.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2024. Instruction-tuning aligns LLMs to the human brain. In *First Conference on Language Modeling*.

Khai Loong Aw and Mariya Toneva. 2023. Training language models to summarize narratives improves brain alignment. In *The Eleventh International Conference on Learning Representations*.

Leonard Bereska and Stratis Gavves. 2024. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mezard. 2025. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. 2023. Broken neural scaling laws. In *The Eleventh International Conference on Learning Representations*.

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

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

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

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

Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborova. 2024. A phase transition between positional and semantic learning in a solvable model of dot-product attention. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502. Association for Computational Linguistics.

Andrea Gregor de Varda, Saima Malik-Moraleda, Greta Tuckute, and Evelina Fedorenko. 2025. Multilingual computational models reveal shared brain responses to 21 languages. *bioRxiv*.

Francesco Denti, Diego Doimo, Alessandro Laio, and Antonietta Mira. 2022. The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12(1):20005.

Adrien Doerig, Tim C Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. 2025. High-level visual representations in the human brain are aligned with large language models. *Nat. Mach. Intell.*, 7(8):1220–1234.

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. 2024. Understanding emergent abilities of language models from the loss perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Tom Dupré la Tour, Michael Eickenberg, Anwar O. Nunez-Elizalde, and Jack L. Gallant. 2022. Feature-space selection with banded ridge regression. *NeuroImage*, 264:119728.

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, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

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

Samir Yitzhak Gadre, Georgios Smyrnis, Vaishaal Shankar, Suchin Gururangan, Mitchell Wortsman, Rulin Shao, Jean Mercat, Alex Fang, Jeffrey Li, Sedrick Keh, Rui Xin, Marianna Nezhurina, Igor Vasiljevic, Luca Soldaini, Jenia Jitsev, Alex Dimakis, Gabriel Ilharco, Pang Wei Koh, Shuran Song, and 6 others. 2025. Language models scale reliably with over-training and on downstream tasks. In *The Thirteenth International Conference on Learning Representations*.

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 Kernian, Shauna Kravec, and 11 others. 2022. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, page 1747–1764.

Changjiang Gao, Zhengwu Ma, Jiajun Chen, Ping Li, Shujian Huang, and Jixing Li. 2025. Increasing alignment of large language models with language processing in the human brain. *Nat. Comput. Sci.*, pages 1–11.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ariel Goldstein, Avigail Grinstein-Dabush, Mariano Schain, Haocheng Wang, Zhuoqiao Hong, Bobbi

Aubrey, Samuel A Nastase, Zaid Zada, Eric Ham, Amir Feder, Harshvardhan Gazula, Eliav Buchnik, Werner Doyle, Sasha Devore, Patricia Dugan, Roi Reichart, Daniel Friedman, Michael Brenner, Avinatan Hassidim, and 3 others. 2024. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nat. Commun.*, 15(1):2768.

Ariel Goldstein, Eric Ham, Mariano Schain, Samuel A Nastase, Bobbi Aubrey, Zaid Zada, Avigail Grinstein-Dabush, Harshvardhan Gazula, Amir Feder, Werner Doyle, Sasha Devore, Patricia Dugan, Daniel Friedman, Michael Brenner, Avinatan Hassidim, Yossi Matias, Orrin Devinsky, Noam Siegelman, Adeen Flinker, and 3 others. 2025. Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. *Nat. Commun.*, 16(1):10529.

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, and 13 others. 2022. Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.*, 25(3):369–380.

Pulkit Gopalani and Wei Hu. 2025. What happens during the loss plateau? understanding abrupt learning in transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Pulkit Gopalani, Ekdeep Singh Lubana, and Wei Hu. 2024. Abrupt learning in transformers: A case study on matrix completion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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 Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 22 others. 2024. Olmo: Accelerating the science of language models. *Preprint*.

Umut Güçlü and Marcel A. J. van Gerven. 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*.

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

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. 2020. Scaling laws for autoregressive generative modeling. *Preprint*, arXiv:2010.14701.

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, and 3 others. 2022. Training compute-optimal large language models. *arXiv [cs.CL]*.

Yufei Huang, Shengding Hu, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. Unified view of grokking, double descent and emergent abilities: A comprehensive study on algorithm task. In *First Conference on Language Modeling*.

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

Shailee Jain and Alexander G Huth. 2018. Incorporating context into language encoding models for fMRI. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 6629–6638, Red Hook, NY, USA. 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. 2020. Scaling laws for neural language models. *arXiv [cs.LG]*.

Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. 2024. Shared functional specialization in transformer-based language models and the human brain. *Nat. Commun.*, 15(1):5523.

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

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, and 9 others. 2023. LLM360: Towards fully transparent open-source LLMs. *arXiv [cs.CL]*.

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, and 64 others. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *Preprint*, arXiv:2407.03963.

Steen Moeller, Essa Yacoub, Cheryl A. Olman, Edward Auerbach, John Strupp, Noam Harel, and Kâmil Uğurbil. 2010. 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.

Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Q. Yamaguchi, Rieko Kubo, Shinji Nishimoto, and Yu Takagi. 2024. Unveiling multi-level and multi-modal semantic representations in the human brain using large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20313–20338. Association for Computational Linguistics.

Tomoya Nakai, Charlotte Constant-Varlet, and Jérôme Prado. 2024. Encoding models for developmental cognitive computational neuroscience: Promise, challenges, and potential. *Developmental Cognitive Neuroscience*, 70:101470.

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

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

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

Nostalgebraist. 2020. 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, and 21 others. 2024. 2 olmo 2 furious. *Preprint*, arXiv: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, and 7 others. 2022. In-context learning and induction heads. *arXiv [cs.LG]*.

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

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

Zhenting Qi, Fan Nie, Alexandre Alahi, James Zou, Himabindu Lakkaraju, Yilun Du, Eric P. Xing, Sham M. Kakade, and Hanlin Zhang. 2025. EvoLM: In search of lost language model training dynamics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yi Ren and Danica J. Sutherland. 2025. Learning dynamics of LLM finetuning. In *The Thirteenth International Conference on Learning Representations*.

Boyan Rong, Alessandro Thomas Gifford, Emrah Düzel, and Radoslaw Martin Cichy. 2025. The time course of visuo-semantic representations in the human brain is captured by combining vision and language models. *eLife*.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Curran Associates Inc.

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

Noam Shazeer. 2020. Glu variants improve transformer. *Preprint*, arXiv:2002.05202.

Charlie Victor Snell, Eric Wallace, Dan Klein, and Sergey Levine. 2024. Predicting emergent capabilities by finetuning. In *First Conference on Language Modeling*.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North*, pages 4149–4158, Stroudsburg, PA, USA. 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. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: analyzing the training dynamics of large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

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. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv: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, and 49 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Greta Tuckute, Nancy Kanwisher, and Evelina Fedorenko. 2024a. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47(Volume 47, 2024):277–301.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152. Association for Computational Linguistics.

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. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric 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, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. 2023. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13711–13738.

Hiroto Q. Yamaguchi, Naoko Koide-Majima, Rieko Kubo, Tomoya Nakai, and Shinji Nishimoto. 2024. Narrative movie fmri dataset. https://doi.org/10.18112/openneuro.ds005531.v1.0.0. OpenNeuro dataset ds005531, version 1.0.0.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.

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

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024. When scaling meets LLM finetuning: The effect of data, model and finetuning method. In *The Twelfth International Conference on Learning Representations*.

A Related work

A.1 Training Dynamics in LLMs

There are many studies quantifying LLM learning dynamics using different evaluation metrics and training stages. During pre-training, Kaplan et al. (2020); Henighan et al. (2020) showed that model loss follows predictable power-law relationships with respect to compute, data, and model size. In contrast, Wei et al. (2022a) has identified "emergent abilities", such as the onset of chain-of-thought prompting beyond a certain scale. Subsequent work has examined how memorization Tirumala et al. (2022) and perplexity Xia et al. (2023) change as a function of model size and the number of training steps. In addition, other studies have applied similar analyses to post-training stages (Ren and Sutherland, 2025; Zhang et al., 2024). More recently, Qi et al. (2025) investigated LLM learning dynamics across the entire training pipeline, including pre-training, continued pre-training, supervised fine-tuning, and reinforcement learning.

To understand LLMs internally, mechanistic interpretability seeks to unveil the internal computations learned by neural networks (Elhage et al., 2021; Dai et al., 2022; Bereska and Gavves, 2024). 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.

A.2 Interpretability from neuroscience insights

From a neuroscience perspective, related work has compared the internal representations of DNNs with human brain activity, demonstrating that the hierarchical structure of high-performing DNNs mirrors the hierarchical processing of the human visual cortex (Yamins et al., 2014; Güçlü and van Gerven, 2015). More recent efforts have employed linear mappings from LLM internal representations to brain activity to probe similarities in language processing between LLMs and the human brain (Jain and Huth, 2018; Schrimpf et al., 2021; Goldstein et al., 2022; Oota et al., 2022; Tuckute et al., 2024a; Goldstein et al., 2024), as well as similarities in vision–semantic processing (Doerig et al., 2025; Rong et al., 2025). Notable examples include investigations of LLM–brain alignment as a function of changes in model size and in the quality and quantity of training data (Caucheteux and King, 2022; Aw and Toneva, 2023; Antonello et al., 2023; Aw et al., 2024; Antonello and Huth, 2024; de Varda et al., 2025; Gao et al., 2025; AlKhamissi et al., 2025a,b), investigations of LLM–brain alignment with respect to the provided input context (Caucheteux et al., 2023; Tuckute et al., 2024b), investigations of LLM–brain alignment with respect to functional substructures within models (Kumar et al., 2024), and investigations of layer-wise LLM–brain alignment (Caucheteux and King, 2022; Antonello and Huth, 2024; Antonello and Cheng, 2024; Goldstein et al., 2025). In addition, brain activity tends to align more strongly with larger- and higher-performance models (Antonello et al., 2023), whereas the layers or representations with the highest next-word prediction accuracy do not necessarily yield the strongest brain alignment; indeed, layers or representations with moderately lower predictive performance can align better with the brain (Caucheteux and King, 2022; Antonello et al., 2023; Antonello and Huth, 2024; Antonello and Cheng, 2024). However, because most of these studies analyze fully trained models, the dynamics of LLM–brain alignment over the course of training remain largely unexplored. In this work, we directly address this gap by systematically probing how brain alignment evolves throughout LLM training.

B Additional methods

B.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.

The architectural variations among the LLMs used in this study encompass layer normalization,

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

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.

B.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 × 2 × 2 mm^3, matrix size = 96 × 96, 72 slices with a thickness of 2 mm, slice gap 0 mm, FOV = 192 × 192

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 × 1 × 1 mm^3, FOV = 256 × 256 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 additional analysis. All annotations were originally described in Japanese. They were translated into English and back-translated into Japanese using DeepL.

B.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.

B.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 additional 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 **??**, 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.

B.5 Code availability

All encoding (Section 2.2) and probing (Section 2.3.2) analyses were conducted using the *himalaya* library[2] (Dupré la Tour et al., 2022) and the *drama2brain* library[3] (Nakagi et al., 2024). To extract latent representations from the MLP layers of OLMo-2, we modified the code from the *Transformers* library[4] (Wolf et al., 2020). To ex-

[2] https://github.com/gallantlab/himalaya
[3] https://github.com/yu-takagi/drama2brain
[4] https://github.com/huggingface/transformers

tract latent representations from the MLP layers of OLMo-0724, LLM-jp, and Amber, we modified the code from the *TransformerLens* library[5] (Nanda and Bloom, 2022). We also use the *llm-jp-eval* library[6] for benchmark analysis. We will make our source code and training data for the encoding, probing and benchmark analyses publicly available on acceptance.

B.6 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 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.

B.7 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 2.2 and 2.3.2, 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 2.3.3. This proportion was treated as the benchmark accuracy for that layer.

In Section C.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.

B.8 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 2.3.1.

[5] https://github.com/TransformerLensOrg/TransformerLens

[6] https://github.com/llm-jp/llm-jp-eval

```
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 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.
### Input:
Question: Which of the following is not a warrantless search
exception?
Choices: A.Cold pursuit,B.Incident to a lawful arrest,C.Plain
view,D.Stop and frisk
### Response:
A
### Input:
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?
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.
Choices: A.Wrong, Wrong,B.Wrong, Not wrong,C.Not wrong,
Wrong,D.Not wrong, Not wrong
### Response:
D
### Input:
Question: An important source of information on the credit rating
of retail businesses is
Choices: A.the Retail Merchants Association,B.the local chamber
of commerce,C.Dun & Bradstreet, Inc.,D.the United States Retail
Credit Association
### Response:
C
### Input:
Question: Which of the following statements expresses a
relationship between aging and sexual functioning
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
### Response:
D
### Input:
Question: Nearsightedness results from
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
### Response:
A
### Input:
Question: Find the degree for the given field extension Q(sqrt(2),
sqrt(3), sqrt(18)) over Q.
Choices: A.0,B.4,C.2,D.6
### 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:
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。
### 入力:
質問：次のうち、無令状捜索例外でないものはどれですか？
選択肢：A.緊急性のない追跡,B.合法逮捕に伴う処分,C.プレーン・ビュー,D.ストッ
プ・アンド・フリスク
### 応答:
A
### 入力:
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ボールを避けるために急ブレーキか
けた。シナリオ2 | 木琴を弾くように子供に言った。
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
### 応答:
D
### 入力:
質問：小売業の信用格付けの重要な情報源は
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
### 応答:
C
### 入力:
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？
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は、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
### 応答:
D
### 入力:
質問：近視はなぜ起きるのか。
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
### 応答:
A
### 入力:
質問：与えられた体の拡大Q(sqrt(2), sqrtｂ(3), sqrt(18))のQに対する次数を
求めなさい。
選択肢：A.0,B.4,C.2,D.6
### 応答:
```

Figure A.1: Example of MMLU (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:
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.
### Input:
Question: What is the highest commander in occupied territories
or colonies called?
Choices: 0. Windmill, 1. Squad leader, 2. Student council, 3.
Chief, 4. Governor
### Response:
4
### Input:
Question: What is the image of the Capitol?
Choices: 0. A symbol of the nation, 1. A theme park, 2. Amusement,
3. Uplifting, 4. Nostalgic
### Response:
0
### Input:
Question: Is it common to wear a suit while visiting various
companies during school?
Choices: 0. Store, 1. Business trip, 2. Job hunting, 3. Bank, 4.
Travel
### Response:
2
### Input:
Question: What must you do before taking off your pants?
Choices: 0. Put on, 1. Take off the pants, 2. Wear, 3. Take off
the clothes, 4. Go to the toilet
### Response:
1
### Input:
Question: What do you call an airport for public use?
Choices: 0.Bus, 1.Lighthouse, 2.Pier, 3.Airport, 4.Opera
### Response:
3
### Input:
Question: What to do if you don't want to obey the system?
Choices: 0. Anti-government, 1. Non-governmental organization, 2.
Air, 3. Convenient, 4. Military government
### 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）でするものとします。 回答となる数値をint型で返し、
他には何も含めないことを厳守してください。
### 入力:
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？
選択肢：0.風車,1.班長,2.生徒会,3.チーフ,4.総督
### 応答:
4
### 入力:
質問：議事堂のイメージは？
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.ふる
### 応答:
0
### 入力:
質問：在学中にスーツを着て各社回るのは？
選択肢：0.店頭,1.出張,2.就職活動,3.銀行,4.旅行
### 応答:
2
### 入力:
質問：パンツを脱ぐ前にしなくてはならないのは？
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
### 応答:
1
### 入力:
質問：公共の用に供する飛行場のことを何と呼ぶ？
選択肢：0.バス,1.灯台,2.船着き場,3.空港,4.歌劇
### 応答:
3
### 入力:
質問：体制に従いたくないならどうする？
選択肢：0.反政府,1.非政府組織,2.空気,3.便利,4.軍政府
### 応答:
```

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
熱を生み出すでしょうか？
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
熱を生み出すでしょうか？
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.

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

### 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.

### 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.

### Response:
D

### 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.

### Response:
D

### 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.

### Response:
C

### 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.

### Response:
C

### 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.

### Response:
A

### 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.

### 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:
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。

### 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.ルービ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。

### 応答:
D

### 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。

### 応答:
D

### 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、カヌーの端から出かける。

### 応答:
C

### 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。

### 応答:
C

### 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、水から守るために綿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.水は熱いべきですが、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。

### 応答:
A

### 入力:
文脈:少年はポ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.ポールの上
で体を持ち上げ続ける。,D.マットから抜け出す。

### 応答:
```

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

C Additional results

C.1 Triple phase transitions in LLMs

Here, we present supplementary results corresponding to Section 3.1. Figures B.3, B.4, B.5, B.6, B.7, and B.8 illustrate outcomes for all participants and main LLMs (OLMo-2, OLMo-0724, LLM-jp) when using layers adjacent to those of Figure 3. We demonstrate that comparable encoding results can be obtained using these layers. Figure B.9 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 3, showing that similar benchmark and probing results are achieved for these alternate tasks. Figure B.10 presents the results obtained using Amber, a different LLM from those employed for Figure 3. 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.11 reports the results obtained when the LLMs are given input in Chinese, a different language to that used for Figure 3. This figure indicates that the phase transition is not observed in a language on which the model has not been trained. Finally, Figure B.12 presents the results of using the *Object* annotation, which differs from that used for Figure 3. Here, we confirm a weaker tendency of phase transition than that observed with the *Story* annotation, suggesting that the phase transition we identified tends to emerge in longer and more complex contexts.

C.2 Neural functional structure associated with phase transitions

In this section, we present supplementary findings related to Figure 4 in Section 3.2.1. Because the goal of this analysis is to quantitatively characterize the neural functional structure associated with the observed phase transitions, in this section we forcus on results for each model in its training language. Figures B.13 and B.15 correspond to Figures 4a and 4c for the main LLMs (OLMo-2, OLMo-0724, LLM-jp). Figure B.14 correspond to Figure 4b for all participants and the main LLMs. The results confirm that, across all participants and all models, we obtain patterns comparable to those shown in Figure 4.

C.3 Evolution of LLM internal representations for downstream tasks

In this section, we present supplementary findings related to Section 3.2.2. Figures B.16, B.17, and B.18 correspond to Figure 5a 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.19 corresponds to Figure 5b 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 3.2.2), 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).

C.4 Changes in the nature of activations

Figure B.20 shows additional results corresponding to those in Figure 6 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.

C.5 Layers of interest

Figures B.21, B.22, and B.23 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 encod-

ing 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.

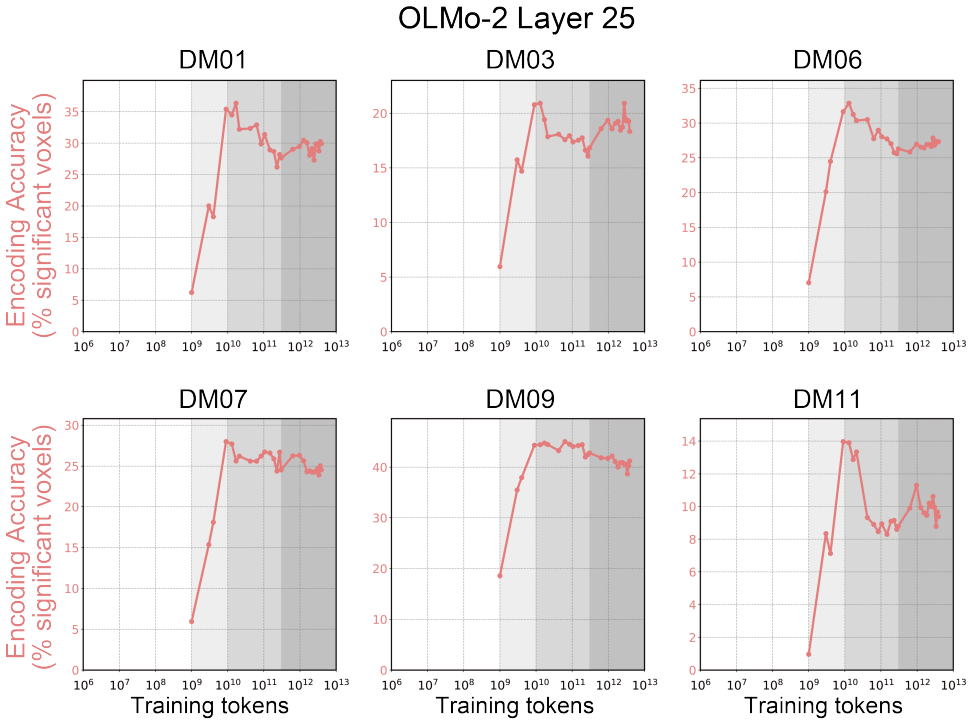

Figure B.1: Results for all participants on the learning dynamics characterized by three phase transition in the number of voxels showing significant encoding accuracy for layer 25 of OLMo-2 when using English annotations.

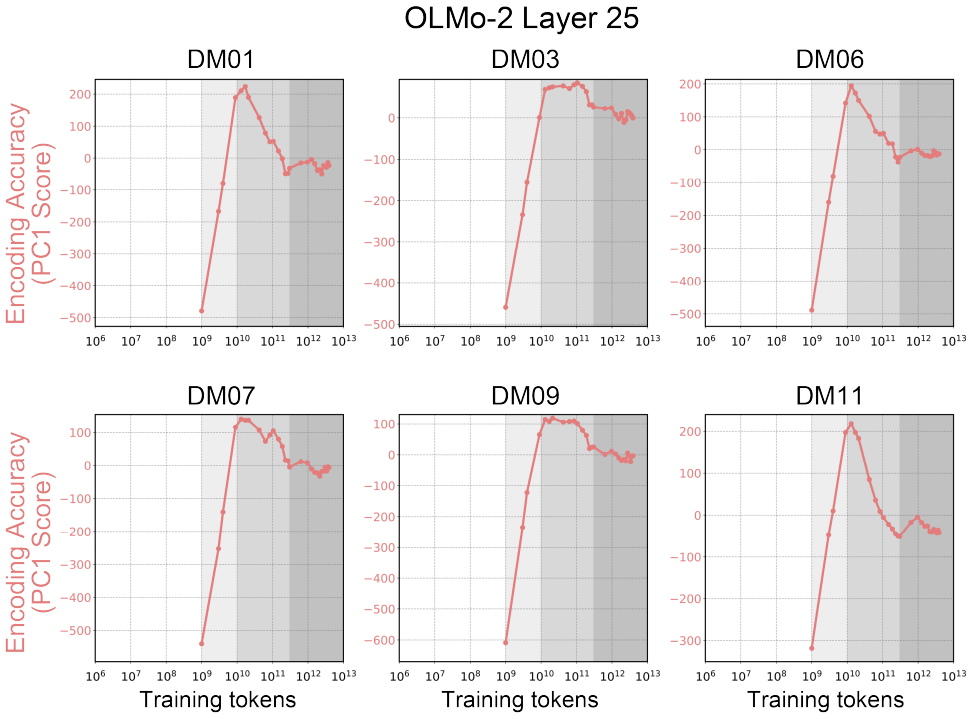

Figure B.2: Results for all participants on the learning dynamics characterized by three phase transition in the PC1 scores of encoding accuracy for layer 25 of OLMo-2 when using English annotations.

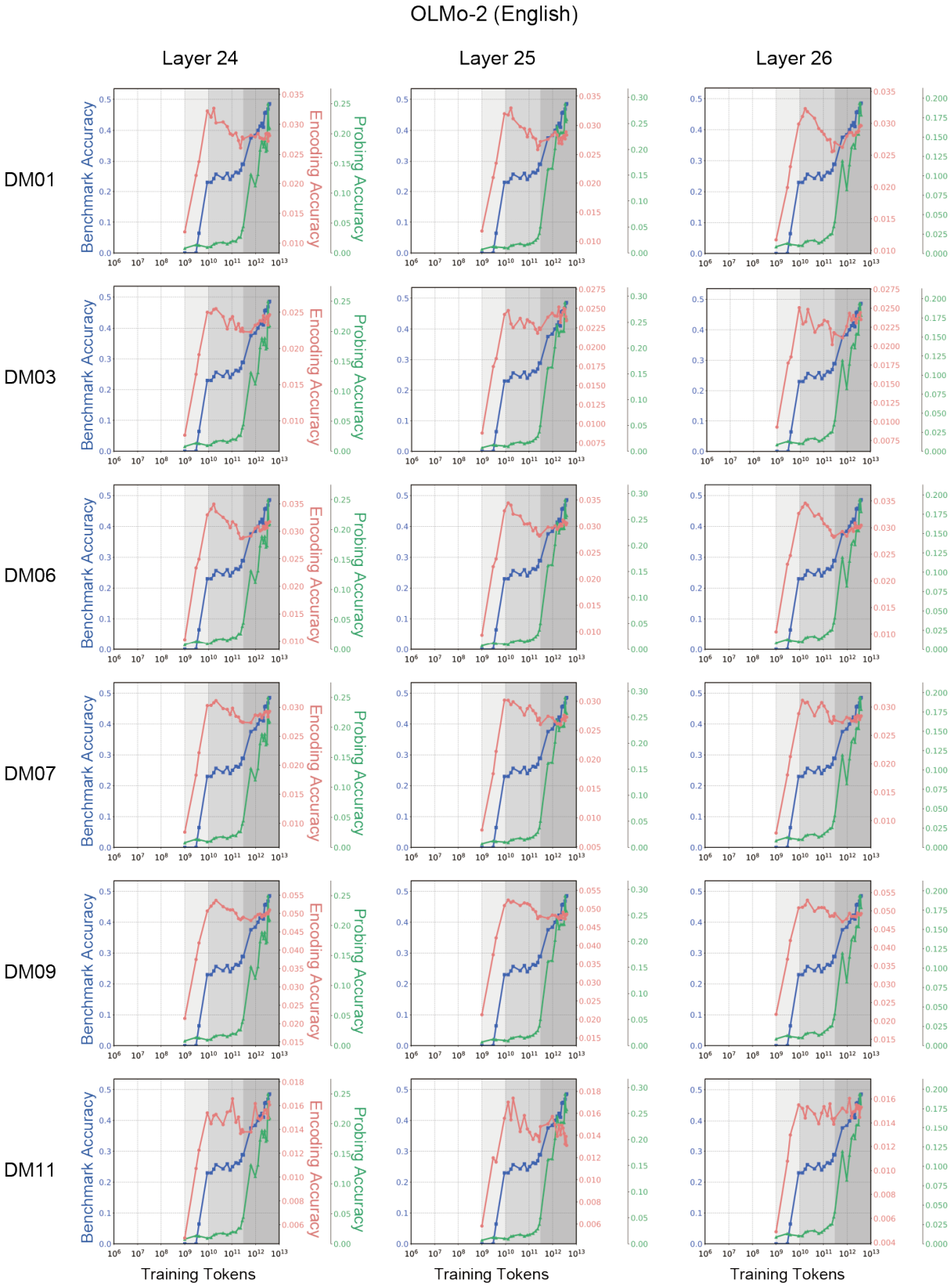

Figure B.3: Results for all participants regarding learning dynamics of layers 24, 25, 26 of OLMo-2 exhibiting three phase transitions when using English annotation and MMLU.

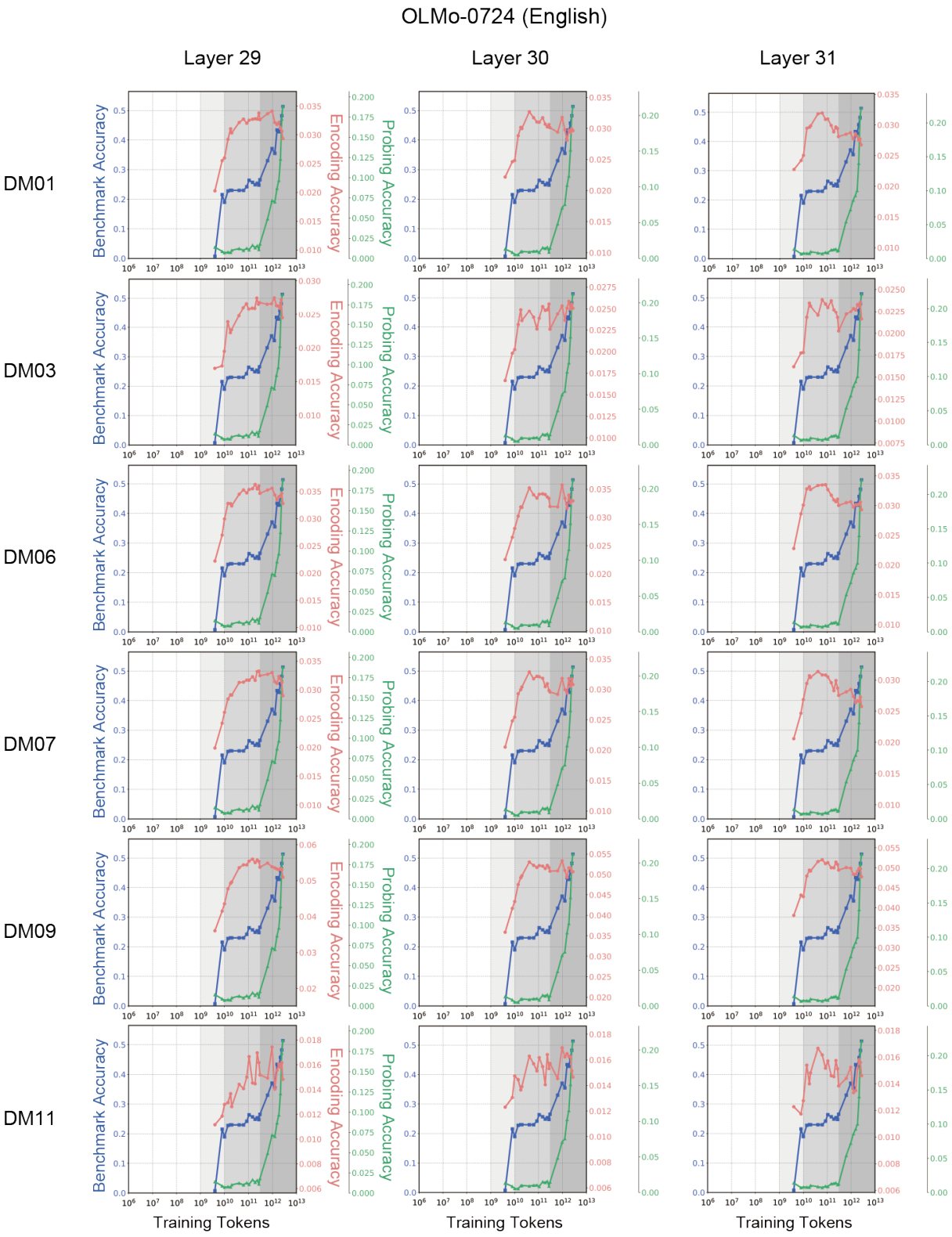

Figure B.4: Results for all participants regarding learning dynamics of layers 29, 30, 31 of OLMo-0724 exhibiting three phase transitions when using English annotation and MMLU.

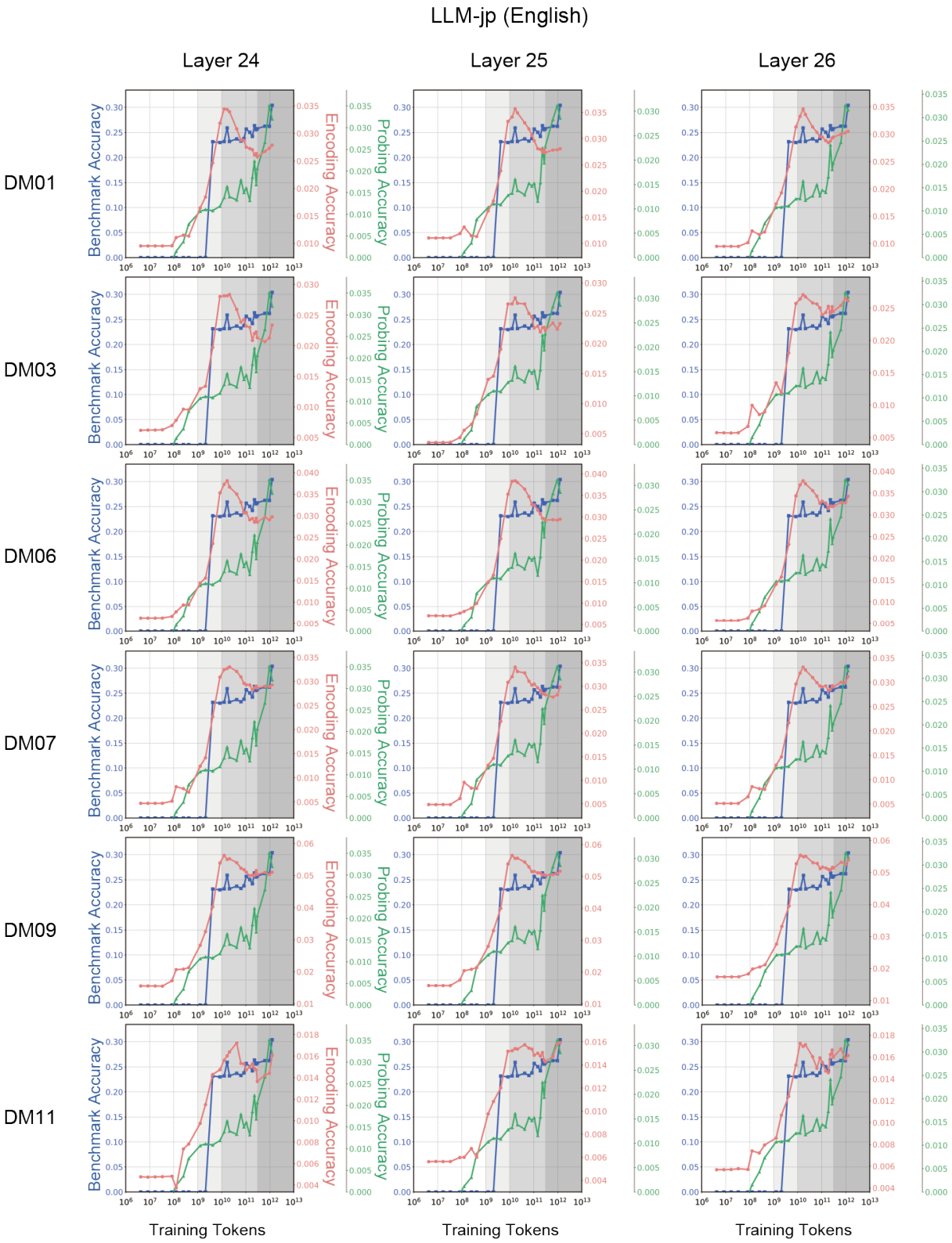

Figure B.5: Results for all participants regarding learning dynamics of layers 24, 25, 26 of LLM-jp exhibiting three phase transitions when using English annotation and MMLU.

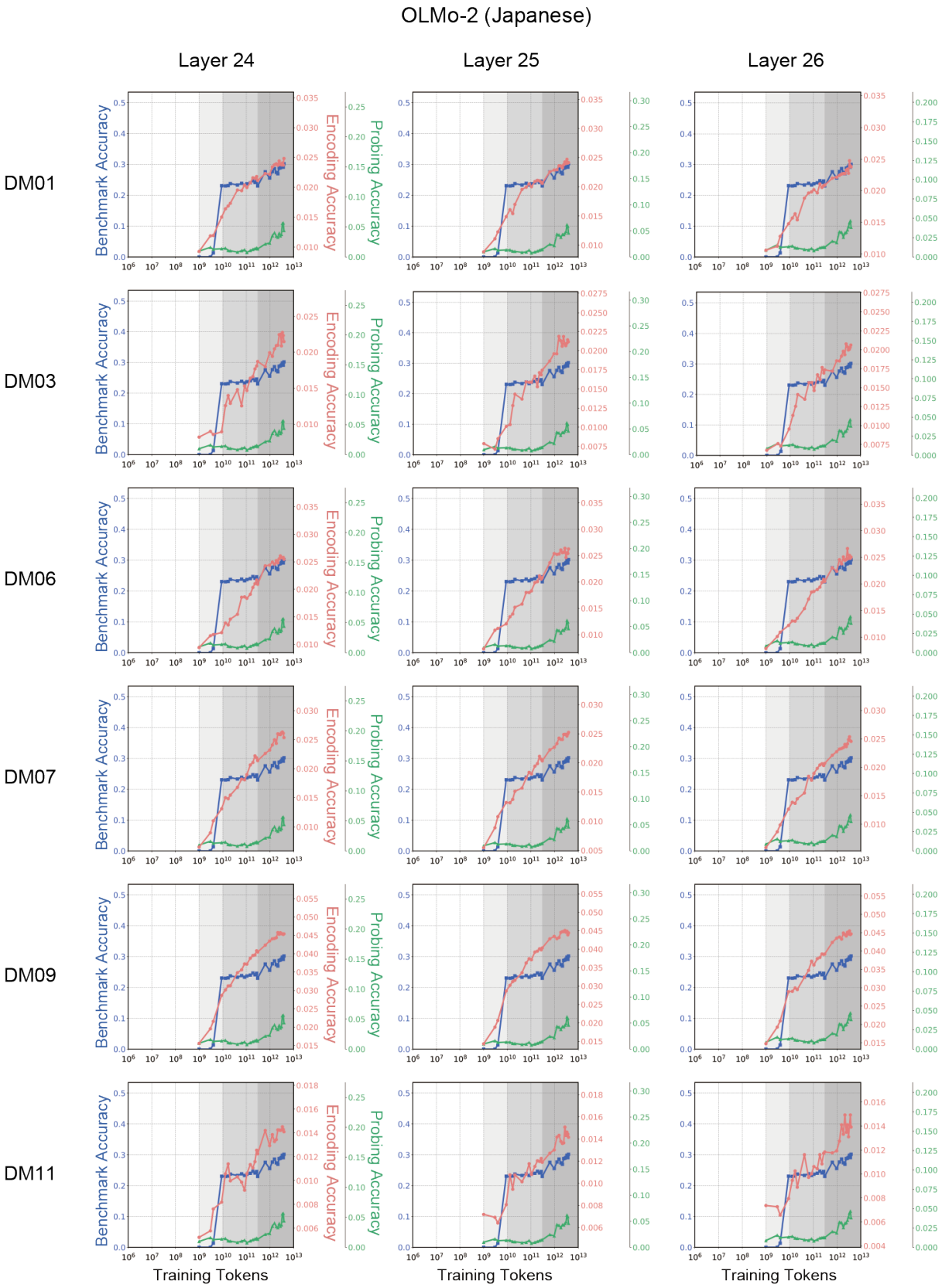

Figure B.6: Results for all participants regarding learning dynamics of layers 24, 25, 26 of OLMo-2 when using Japanese annotation and MMLU.

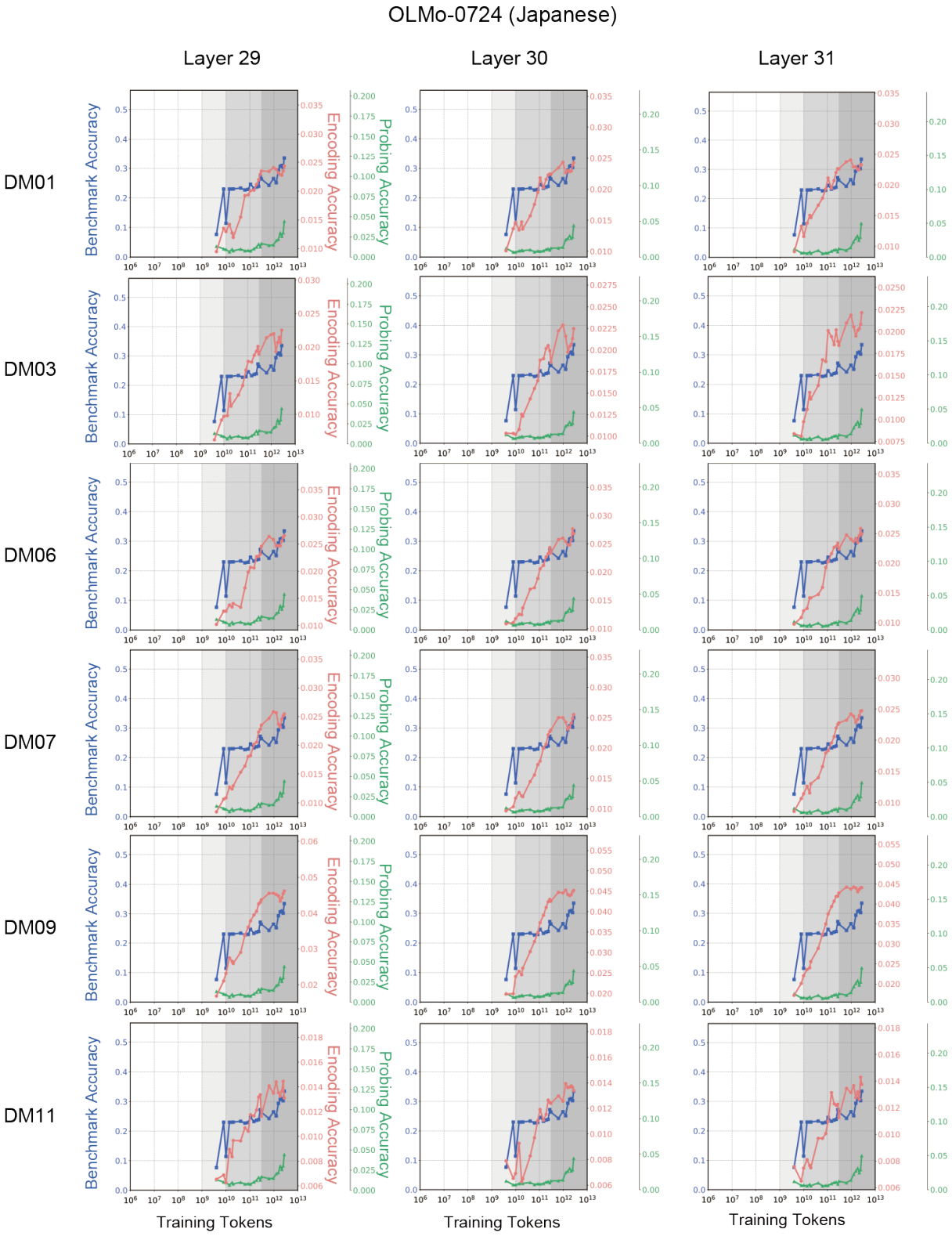

Figure B.7: Results for all participants regarding learning dynamics of layers 29, 30, 31 of OLMo-0724 when using Japanese annotation and MMLU.

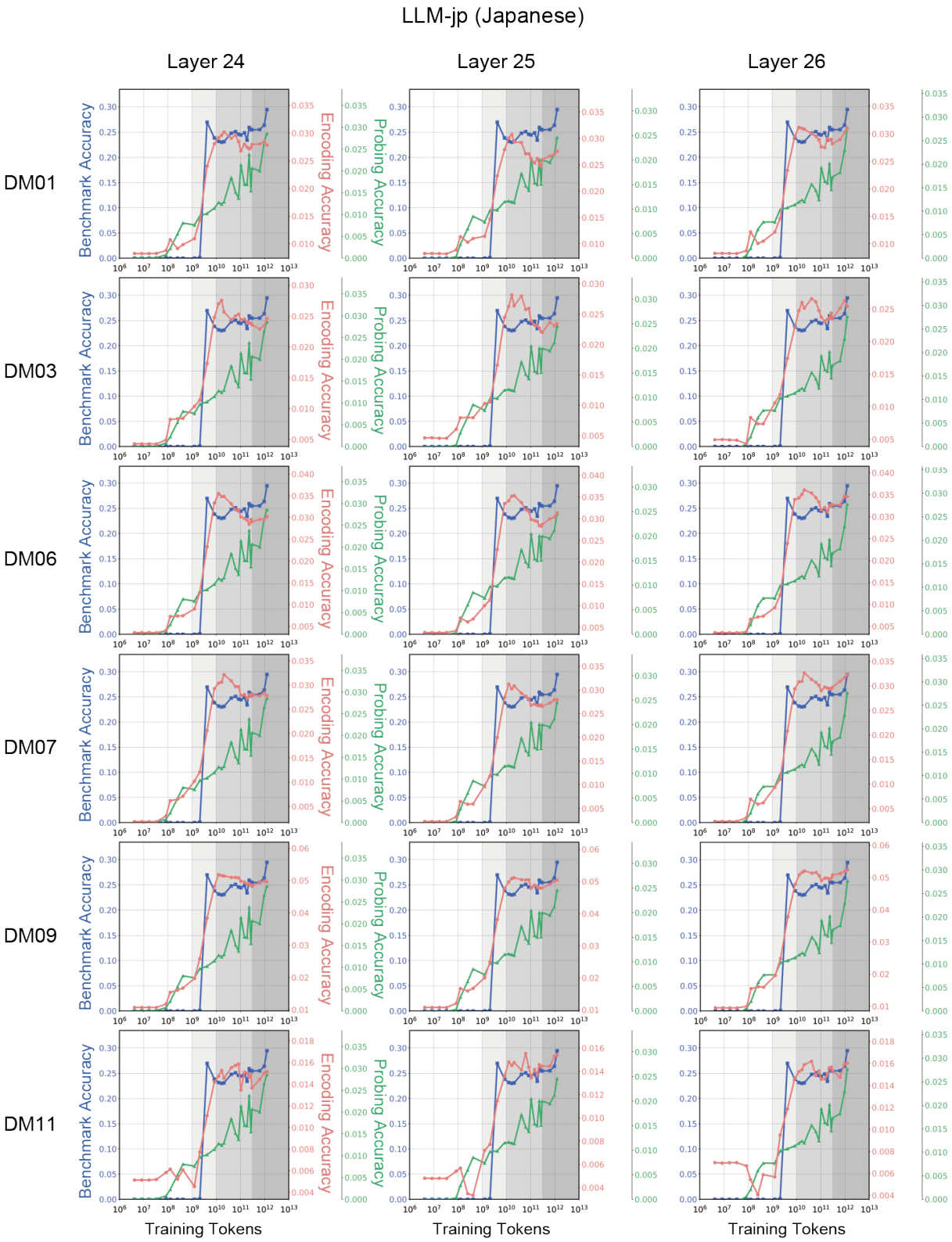

Figure B.8: Results for all participants regarding learning dynamics of layers 24, 25, 26 of LLM-jp exhibiting three phase transitions when using Japanese annotation and MMLU.

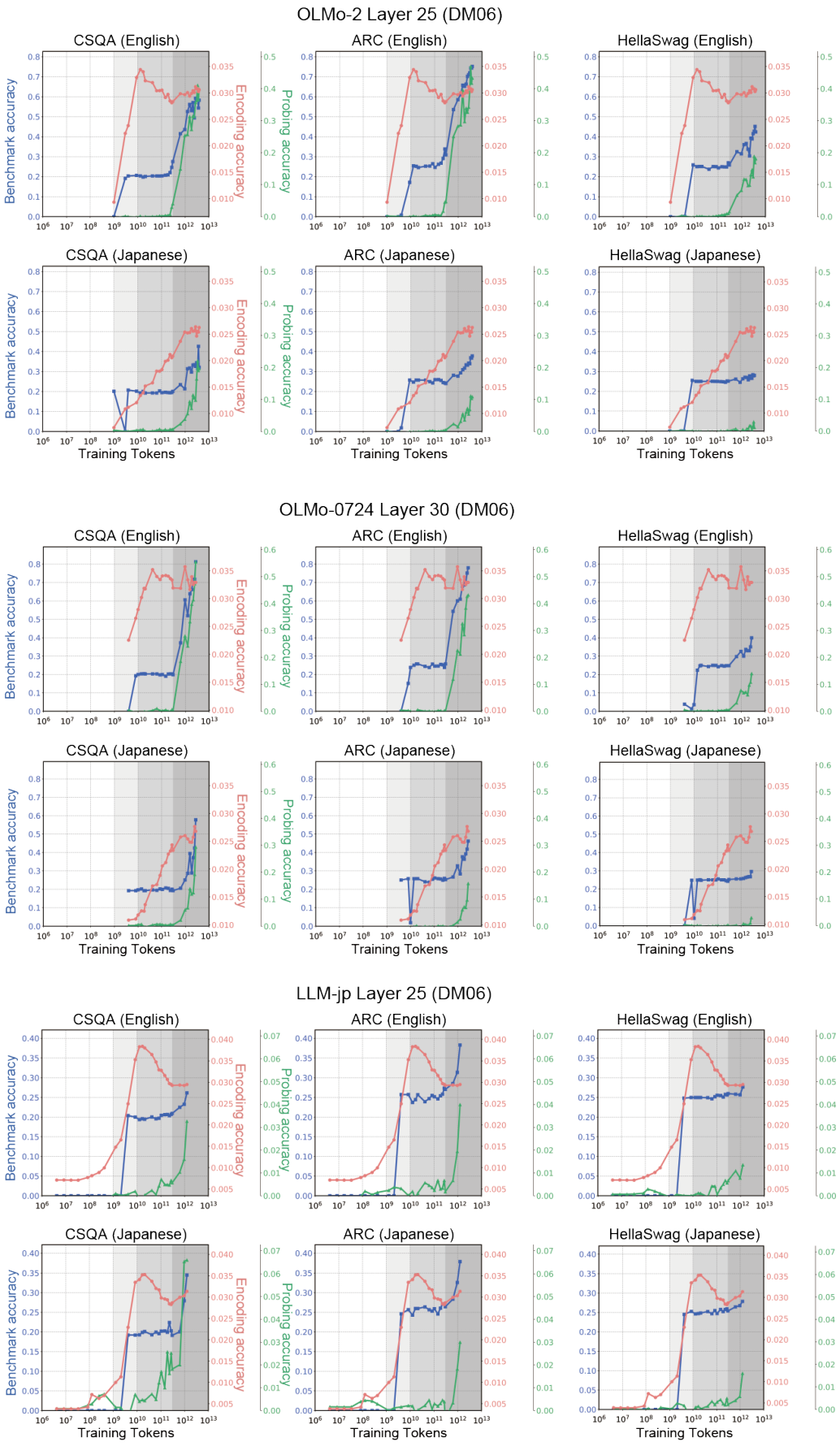

Figure B.9: Results from a single participant (DM06) for learning dynamics of layer 25 of OLMo-2 and LLM-jp, layer 30 of OLMo-0724 exhibiting three phase transitions when using other tasks.

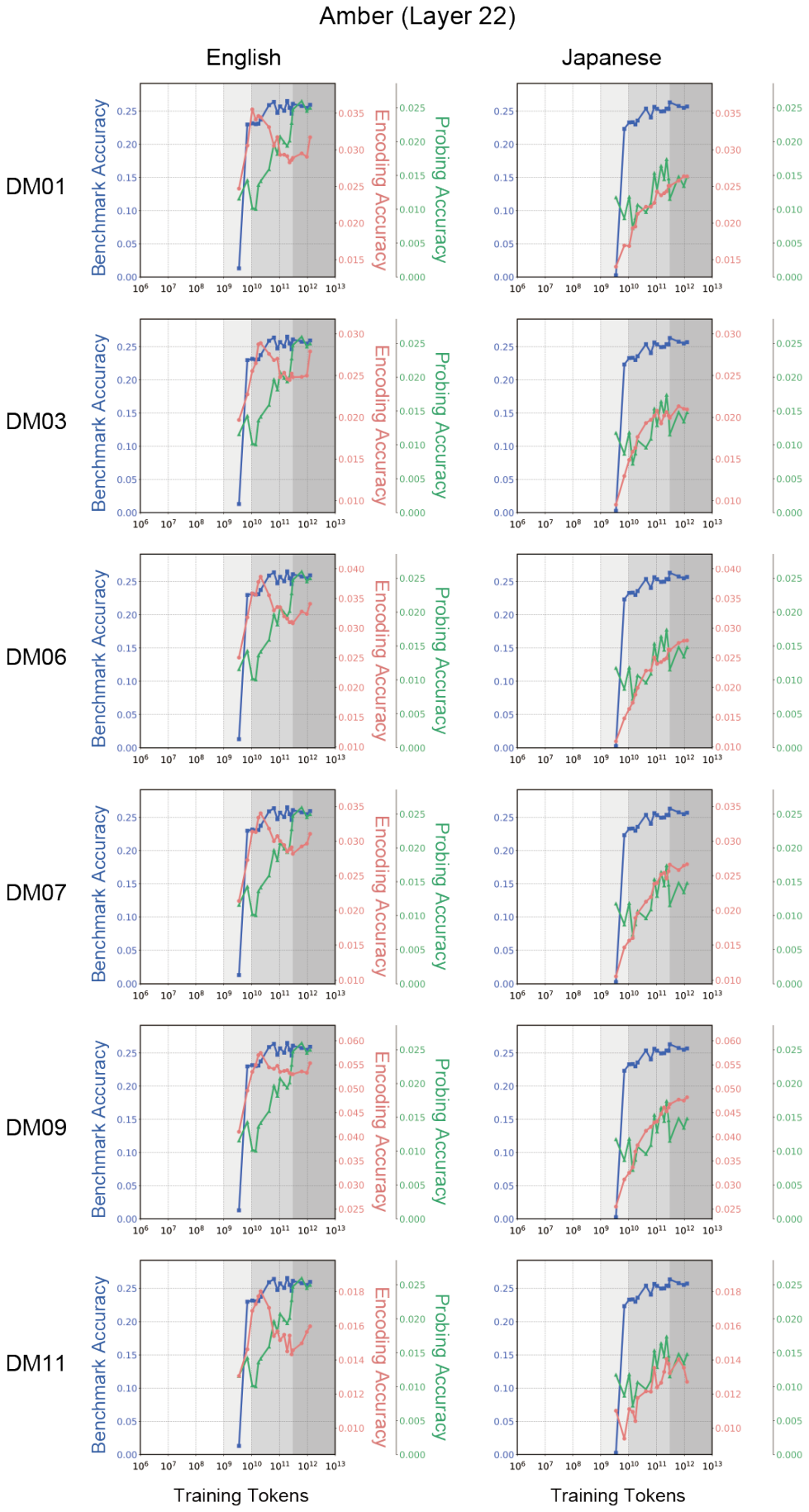

Figure B.10: Results for all participants regarding learning dynamics of layer 22 of Amber exhibiting three phase transitions when using English/Japanese annotation and MMLU.

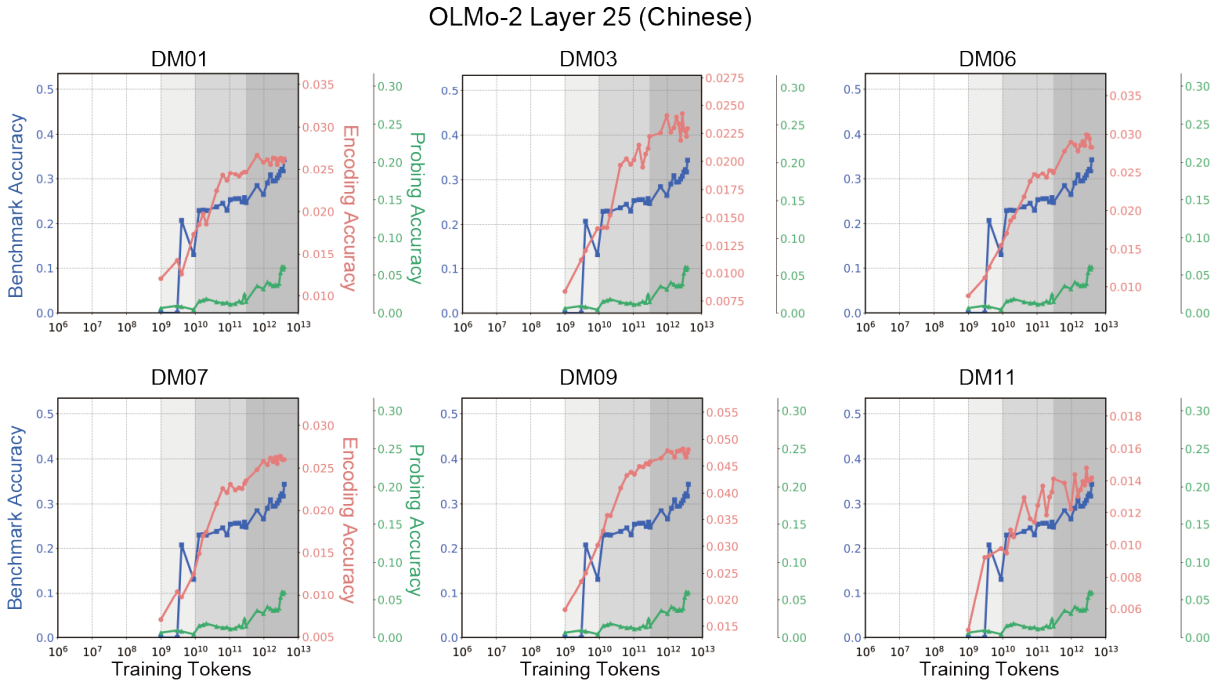

Figure B.11: Results for all participants regarding learning dynamics of layer 25 of OLMo-2 when using Chinese annotation and MMLU.

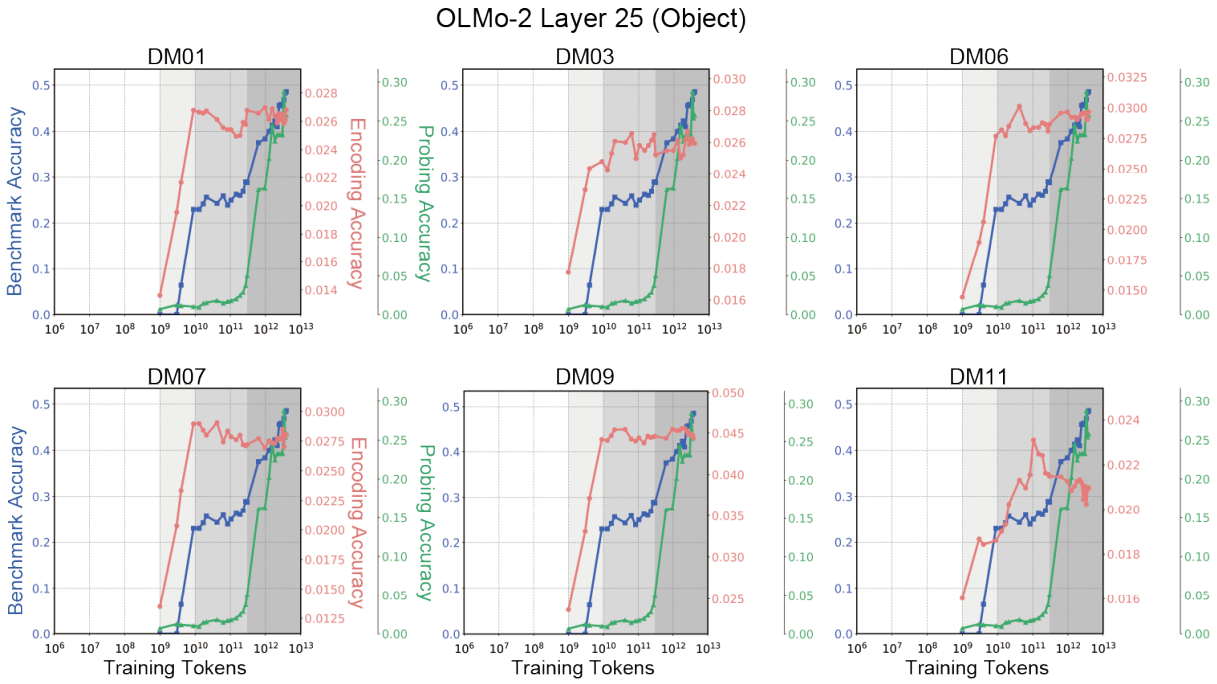

Figure B.12: Results for all participants regarding learning dynamics of layer 25 of OLMo-2 when using English *Object* annotation and MMLU.

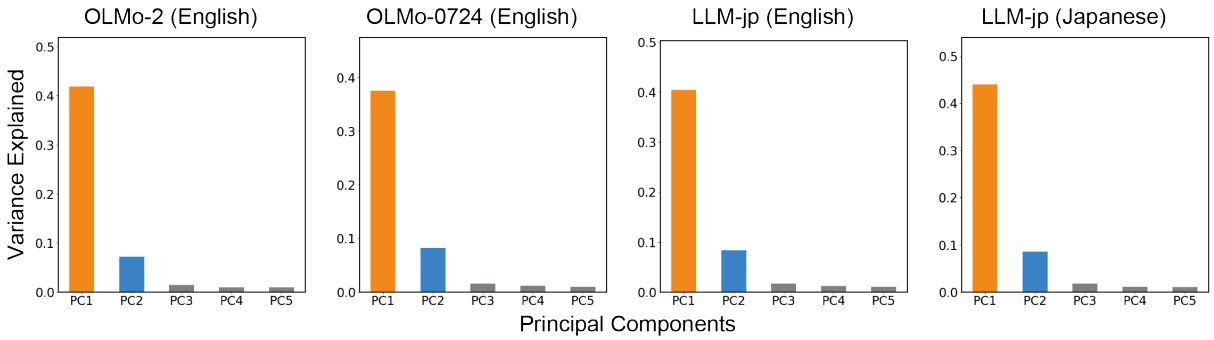

Figure B.13: Results for all models regarding the explained variance of the top five PCs of the estimated brain activations.

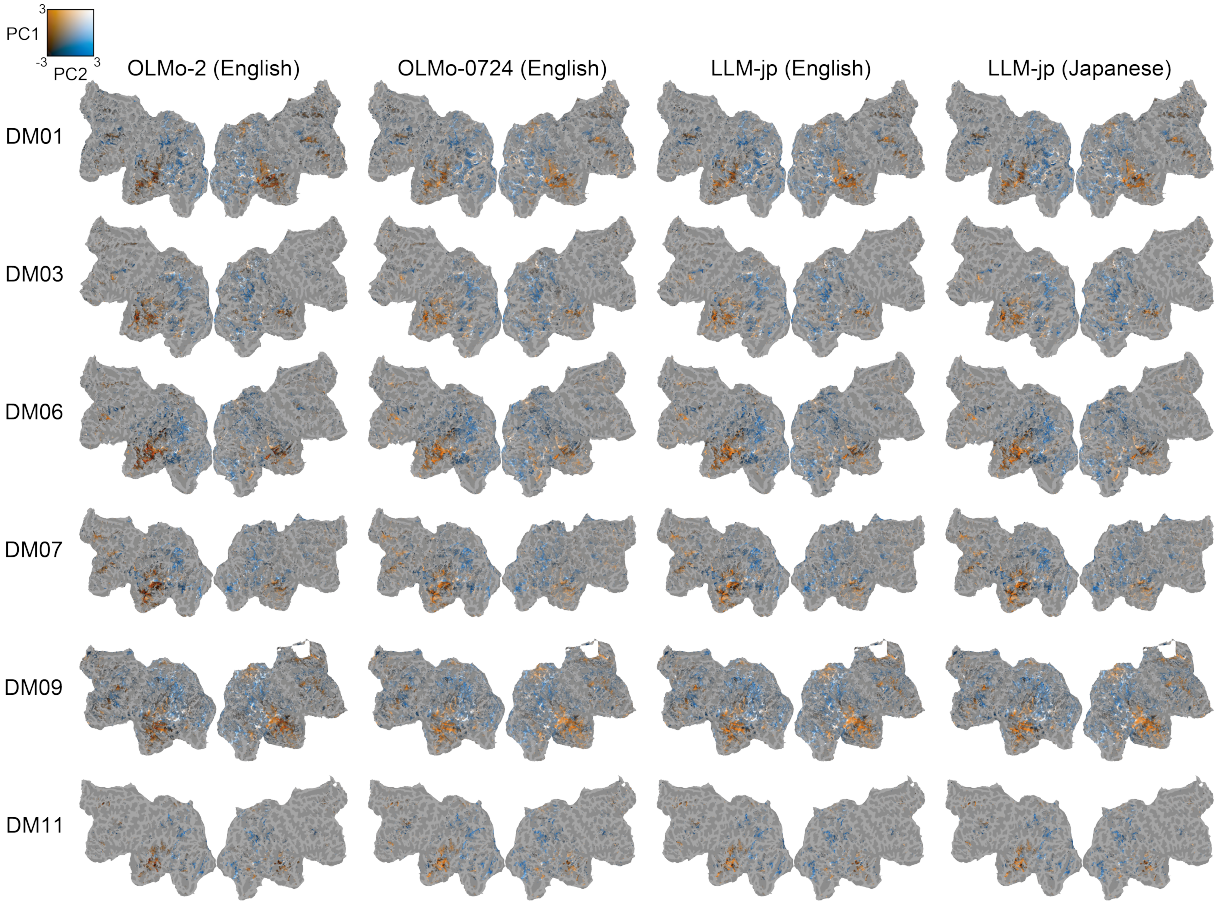

Figure B.14: Results for all participants and all models regarding PC1 and PC2 scores at the final checkpoint.

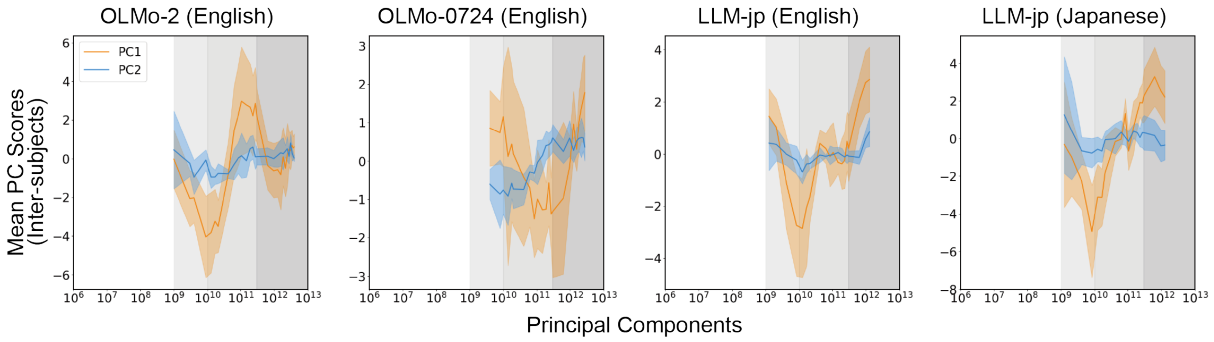

Figure B.15: Results for all models regarding changes of the mean scores of PC1 and PC2 scores for voxels that were significant at all checkpoints.

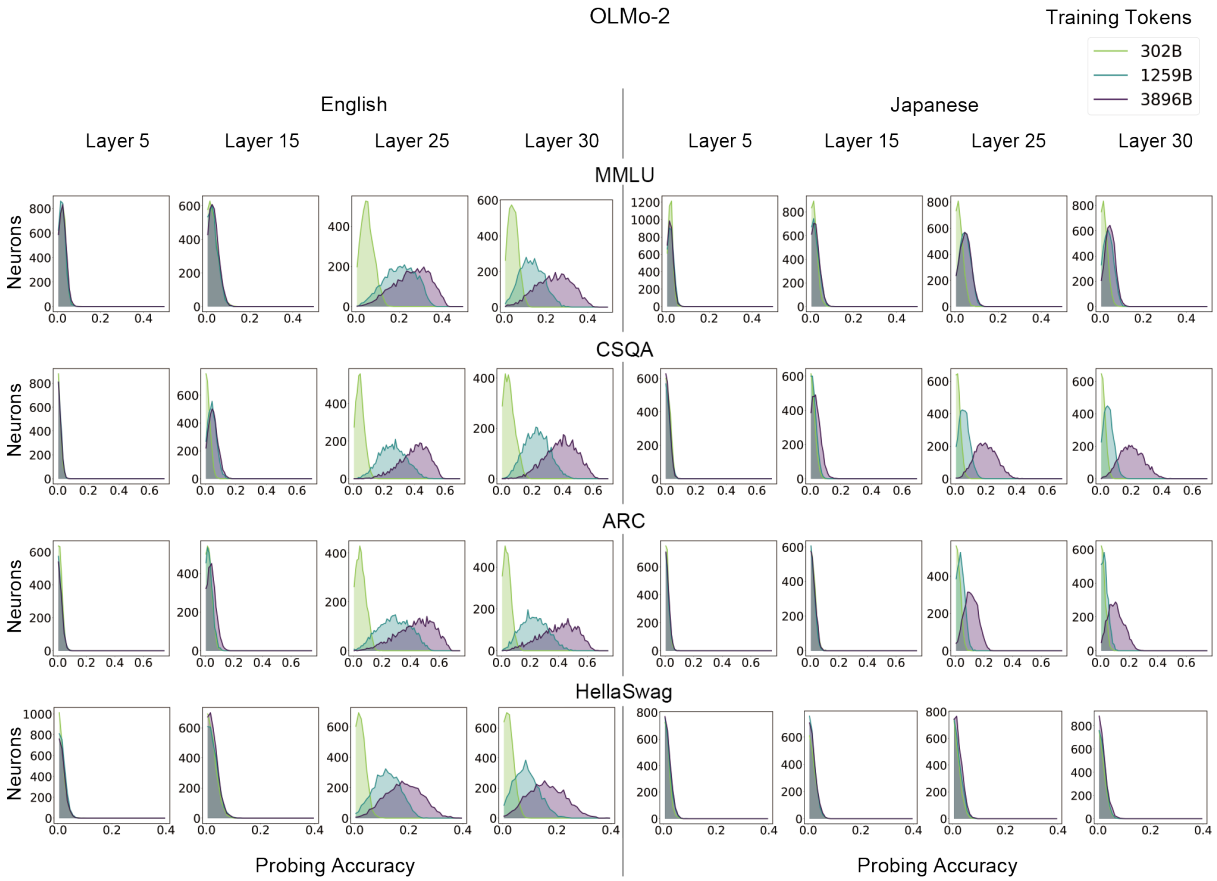

Figure B.16: Results for all downstream tasks regarding changes in the activations of OLMo-2.

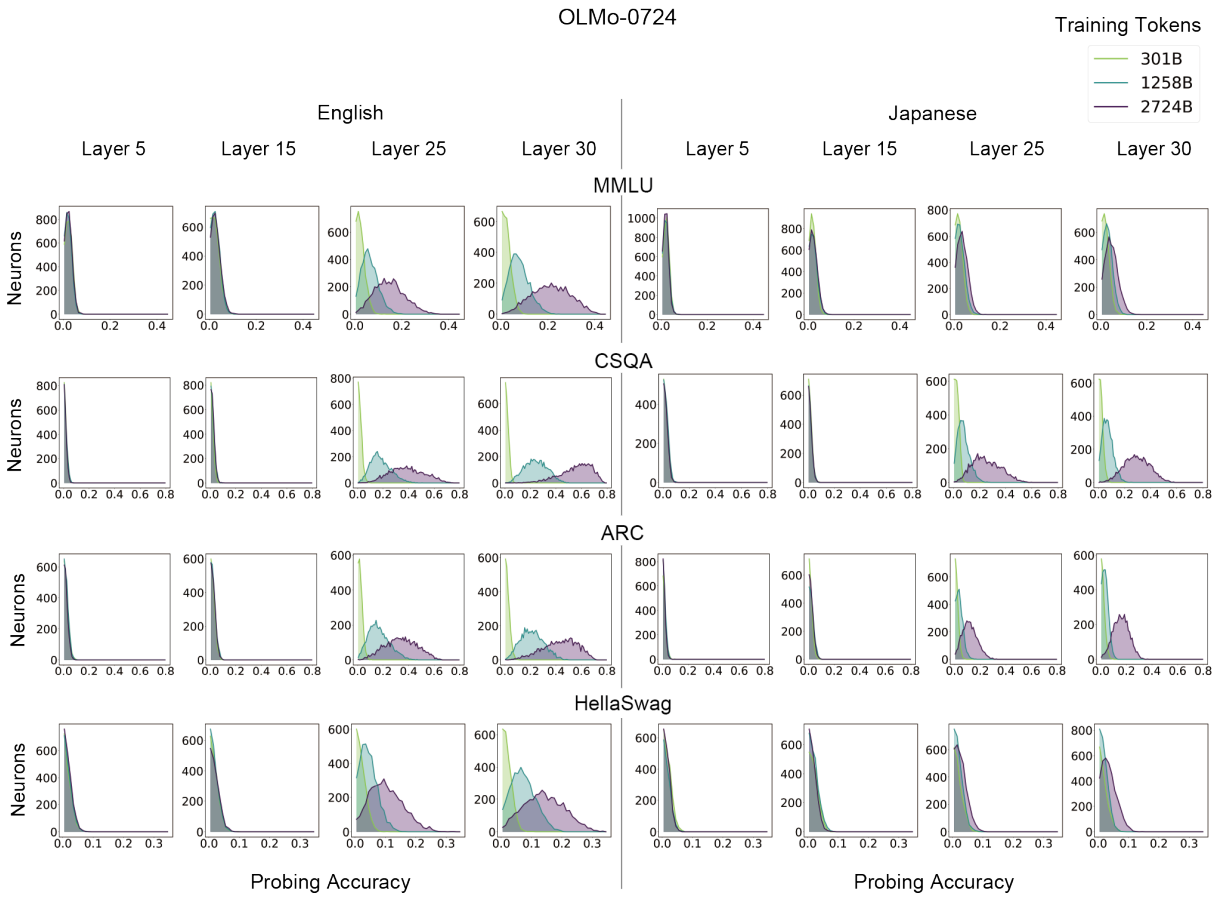

Figure B.17: Results for all downstream tasks regarding changes in the activations of OLMo-0724.

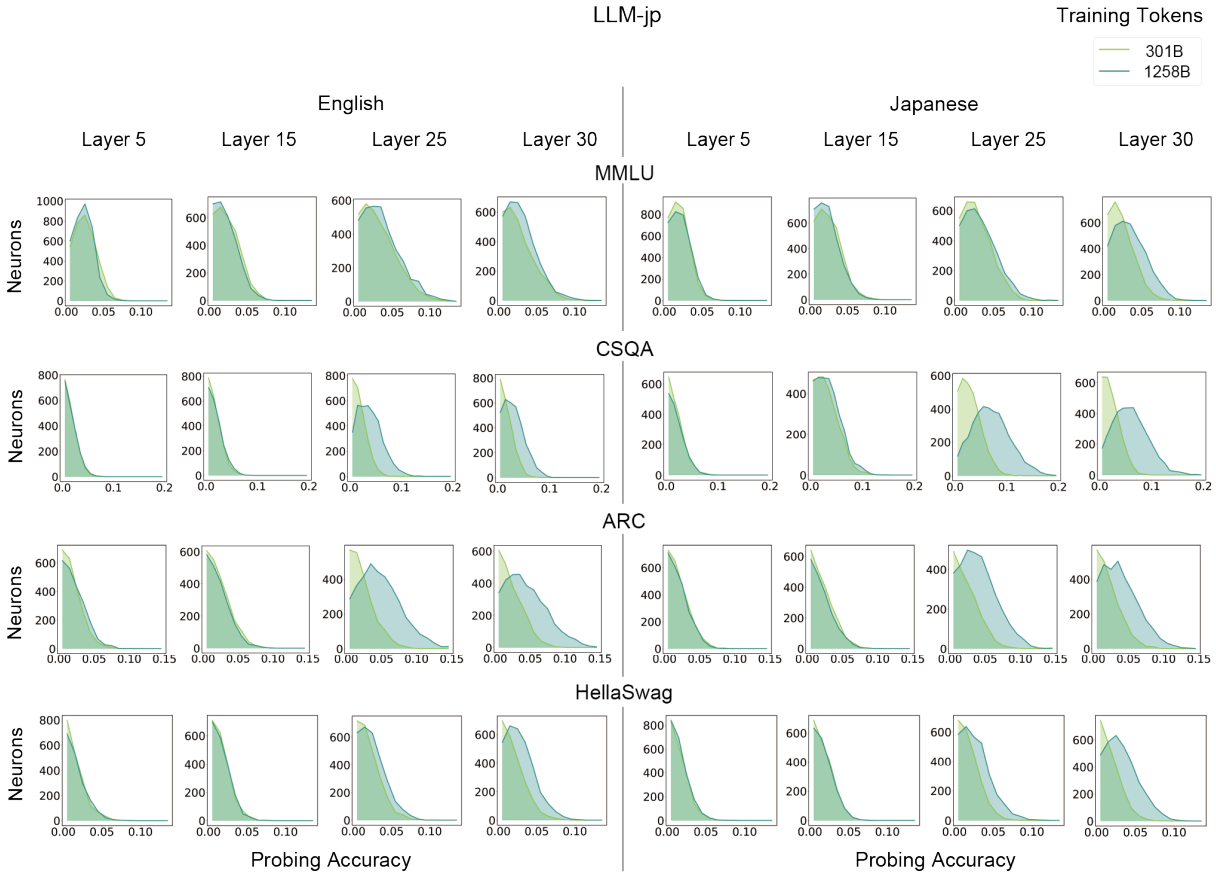

Figure B.18: Results for all downstream tasks regarding changes in the activations of LLM-jp.

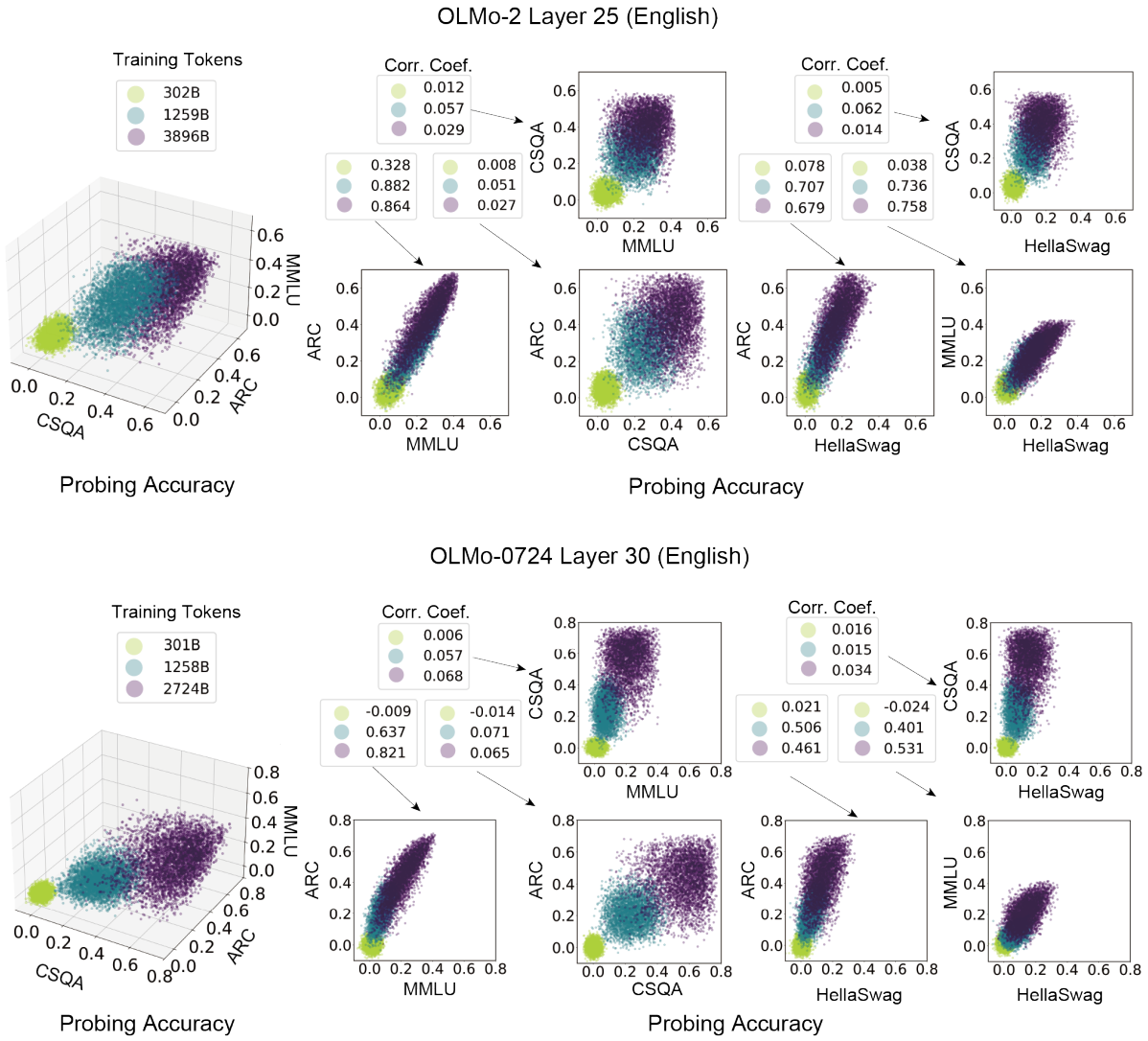

Figure B.19: Relationship between probing accuracies in OLMo-2 (layer 25) and OLMo-0724 (layer 30) across English MMLU, CSQA, ARC, and HellaSwag.

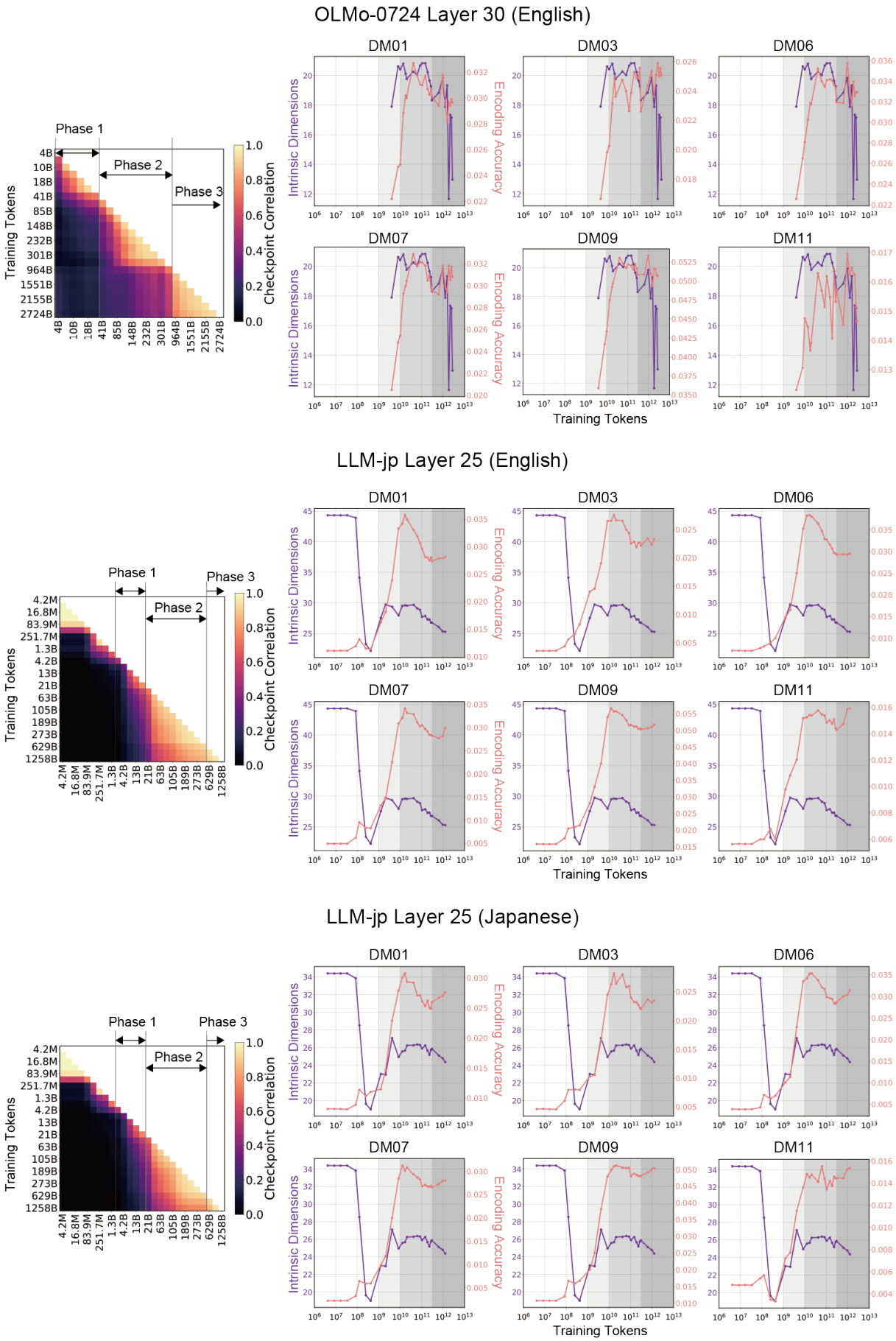

Figure B.20: Variations in correlation coefficients (left), encoding accuracy, and IDs (right) of the activations of OLMo-0724 (layer 30)/LLM-jp (layer 25) using learned languages across checkpoints.

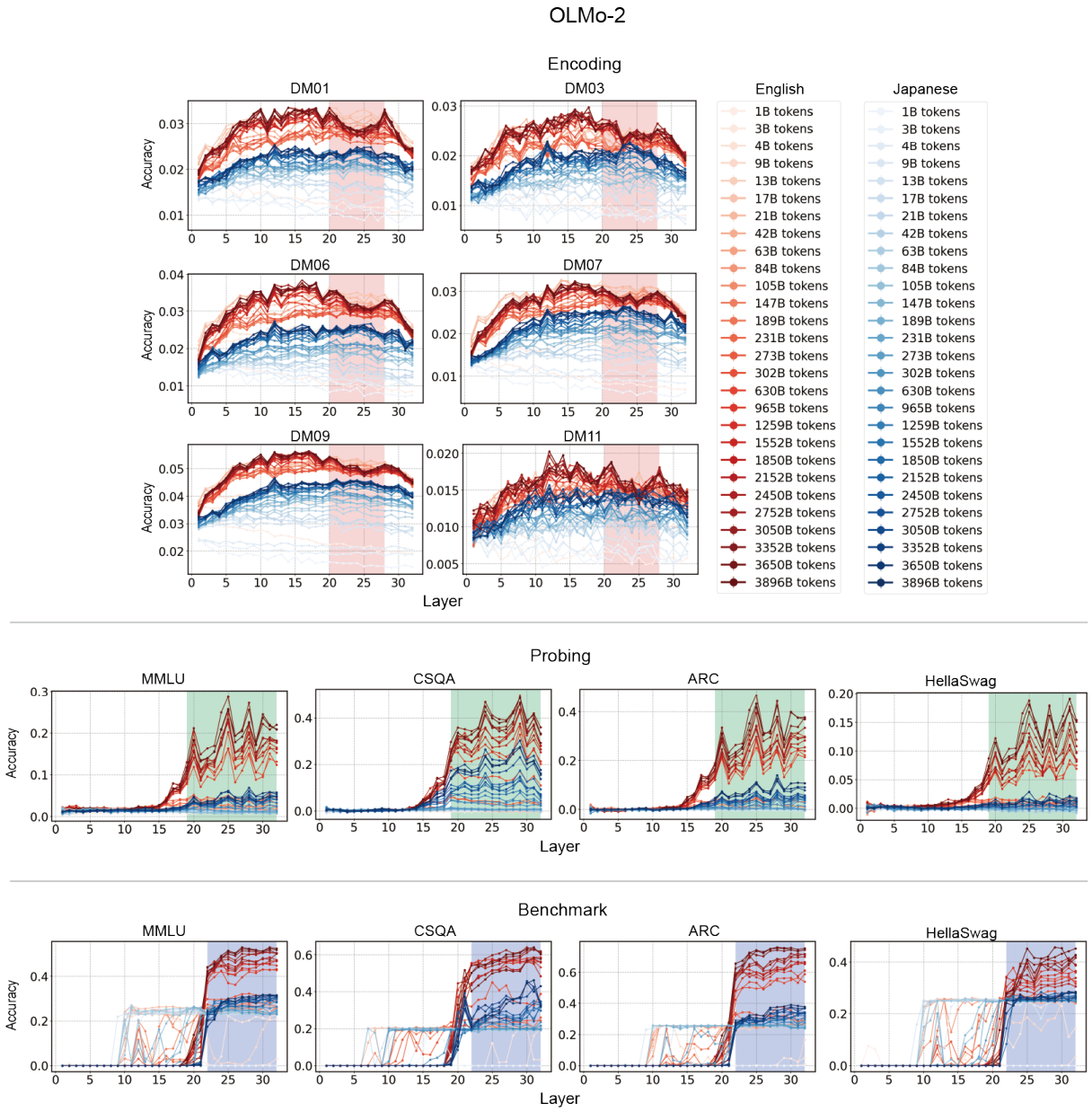

Figure B.21: Layers of interest for OLMo-2.

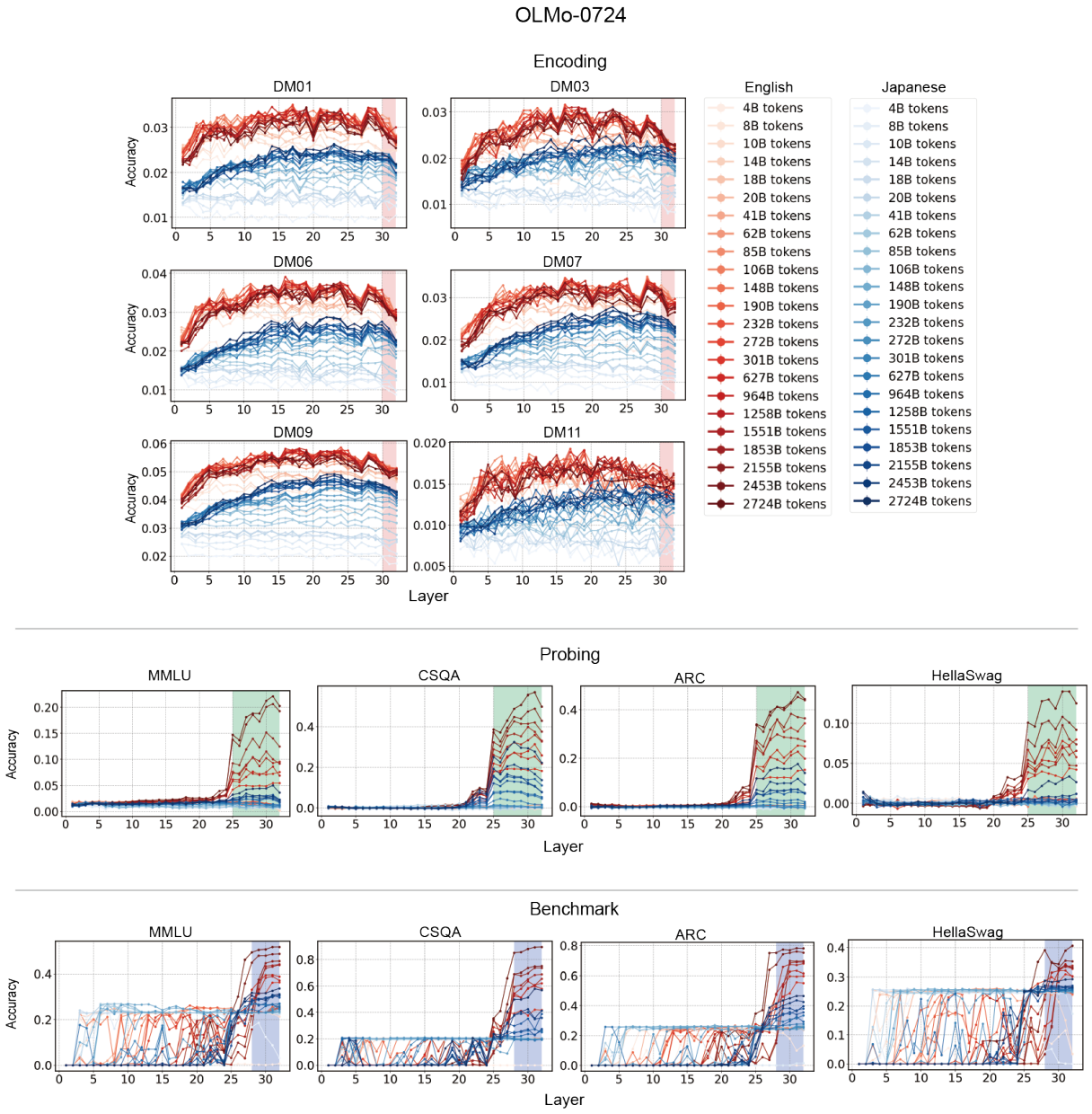

Figure B.22: Layers of interest for OLMo-0724.

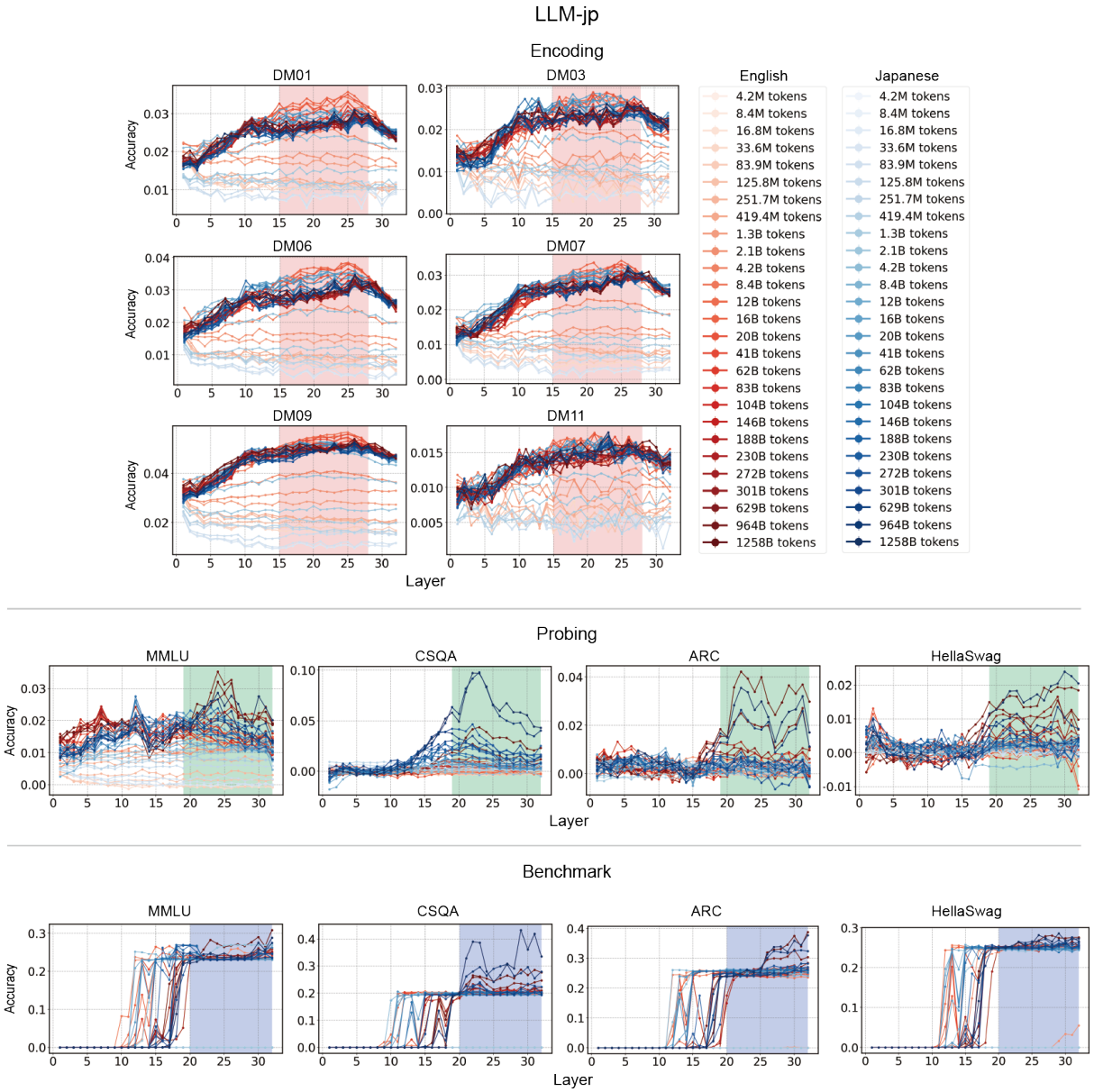

Figure B.23: Layers of interest for LLM-jp.