DYNAMICAL ANATOMY OF EMERGENT ABILITIES IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In large language models (LLMs), the phenomenon of *emergent abilities*, characterized by abrupt performance improvements at specific scales or training stages, has received considerable attention. Notably, the learning process is observed to transition through three distinct phases: an initial non-linguistic output phase, a formally and syntactically consistent but semantically hollow phase, and finally, a meaningfully valid output phase. However, the underlying properties of internal information processing within these three phases remain unknown. In particular, it remains to be clarified what qualitative shifts in the model's internal processes underpin the acquisition of linguistic proficiency and higher-order intelligence. In this study, we focus on the high-dimensional internal dynamics of the LLM and analyze its dynamical properties across the three identified phases. We employ power spectral density (PSD) to investigate the long-range temporal correlations within the model's internal dynamics during text processing. Our analysis reveals that the transition from the first to the second phase is characterized by the emergence of a power-law spectrum with an exponent of $\beta \approx 0.5$, which reflects the characteristic strength of long-range correlations typically observed in natural language corpora. In the third phase, measuring β across individual hidden dimensions shows a heavy-tailed broadening of the exponent distribution while maintaining a constant layer-mean, marked by the emergence of distinct outlier units. Further inspection of these outlier units, specifically those exhibiting weakened temporal correlations, reveals that they selectively respond to functional vocabulary like punctuations. Together, these observations suggest that analyzing processing dynamics along the temporal axis offers a valuable new lens for understanding how language capabilities emerge in LLMs.

1 INTRODUCTION

In recent years, large language models (LLMs) have achieved remarkable breakthroughs in various domains, including natural language processing, reasoning and code generation (Vaswani et al., 2017; Brown et al., 2020). A widely observed empirical rule underpinning these advances is the *scaling law*, which dictates that model performance improves predictably in proportion to the logarithm of parameter count, training data volume, and computational resources (Kaplan et al., 2020; Hoffmann et al., 2022). However, in contrast to such continuous improvement, recent observations have identified a phenomenon known as *emergent abilities* (Wei et al., 2022; Ganguli et al., 2022; Srivastava et al., 2023). In this regime, sophisticated capabilities emerge abruptly once specific thresholds of compute, training tokens, or parameters are surpassed. A prominent characteristic of emergent abilities is their universality, manifesting across broad tasks and domains (Wei et al., 2022). In particular, the emergent abilities are observed regardless of tasks, model architecture, and algorithms employed. These similarities observed across diverse domains suggest that they stem from an underlying common mechanism and a qualitative reorganization of the model's internal information processing.

The universality of emergent abilities allows for analogies with phenomena in various disciplines outside of machine learning. For example, such discontinuous behaviors are sometimes discussed in terms of phase transitions in physics (Wei et al., 2022; Chen et al., 2023; Sun & Haghighat, 2025; Nakagi et al., 2025). Indeed, the learning process of neural networks can be formulated as nonlinear dynamics, i.e., both theoretical and experimental studies have shown that these transitions tend to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

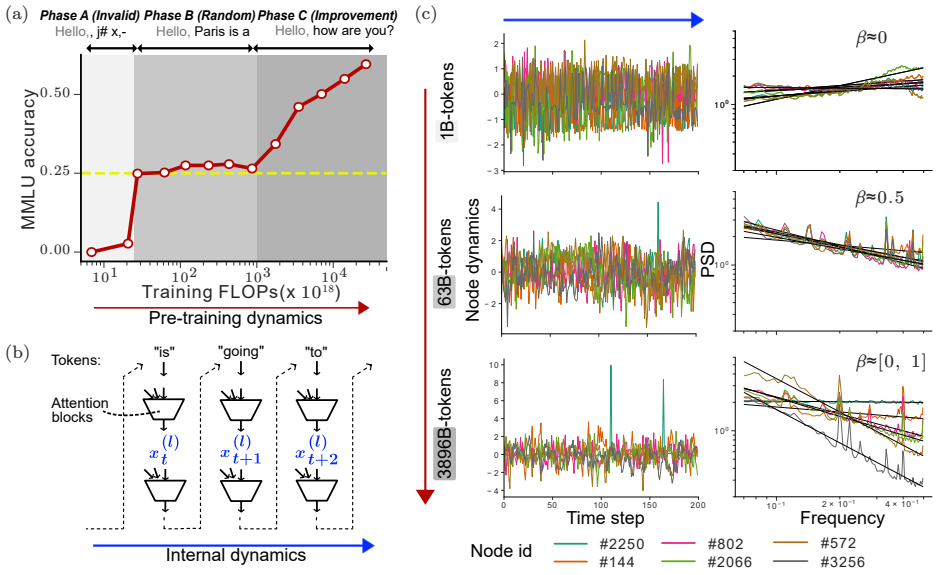

Figure 1: Conceptual overview. (a) Triple-phase transitions during the LLM pre-training process (Nakagi et al., 2025): The model transitions through three distinct stages: the invalid response phase (phase A), the formally valid but nonsensical phase (phase B), and the emergence phase characterized by the acquisition of advanced reasoning capabilities (phase C). The red line indicates the progression of benchmark scores for the OLMo-2-1124-7B model. Yellow line indicates chance rate. (b) Formulation of the internal LLM dynamics: By defining hidden states as variables and the state updates during token generation as temporal evolution, the inference process is formulated as time-series data. (c) Analytical framework with example dynamics: internal dynamics are recorded from models at distinct training stages. The left panel displays the dynamics captured during the 200 steps. The right panel shows the PSD along with its fitted slope. Through the analysis, we characterize the long range correlation within the system associated with the transitions between learning phases.

become more abrupt as the model size increases (Saxe et al., 2013). Similar nonlinear behaviors appear in the grokking phenomenon (Power et al., 2022; Liu et al., 2022), where generalization emerges after prolonged performance plateaus, as well as in human developmental milestones such as the vocabulary spurt (Nelson, 1973) and critical periods (Lenneberg, 1967). Thus, these abrupt changes in learning and developmental processes, common to both artificial and biological intelligence, can be understood as nonlinear phenomena occurring during qualitative transitions of the system.

A recent perspective proposes that emergent abilities should be viewed not merely as benchmark improvements above random chance, but as a *triple-phase transition* (Nakagi et al., 2025) (Fig.1a). Using the pre-training checkpoints released for several language models, the study demonstrated that internal states can be classified into three discontinuous phases through cross-correlation analysis. The outputs in each phase change from "invalid responses" (phase A) to "formally valid but incorrect responses" (phase B), and finally to "semantically correct responses" (phase C) suggesting a clear correspondence between internal states and functional characteristics. Furthermore, a comparison with human brain activity during the performance of the same task revealed a sharp increase in the brain-alignment score during the final phase C. This observation suggests that the emergence of advanced capabilities is accompanied by a functional shift toward representations that mirror biological information processing. However, such correlation analyses only demonstrate superficial similarities in internal representations and are limited in clarifying the mathematical properties of internal processing that underlie functional acquisition. To uncover the true nature of emergent capabilities, it is necessary to directly clarify the dynamical computational processes, which we term *dynamical anatomy*—specifically, how input token information is processed as temporal dependencies and reflected in the transitions of internal states. In other words, it is an important next step to

identify the specific mechanisms by which the integration of information along the token generation time series leads to advanced reasoning capabilities.

Internal dynamics analysis (Fig. 1b) is an effective method for elucidating these functional aspects. This approach formulates the inference process as multivariate time-series data by taking internal states as variables and token generation as temporal evolution. Previous studies employing this approach have described LLM behavior as a dynamical system and clarified its mathematical properties. For example, an analysis of BERT (Devlin et al., 2019; Lan et al., 2019) revealed chaotic properties in the models (Inoue et al., 2022; Liu et al., 2025). These properties emerge only in trained models. These findings suggest the sensitivity to initial conditions may play a crucial role in the acquisition of advanced natural language processing capabilities, amplifying minute differences in input sentences into changes in internal states. Recent studies using energy landscape analysis on the autoregressive generative models have highlighted similarities between the internal dynamics of models such as Llama 3 (Grattafiori et al., 2024) and GPT-2 (Radford et al., 2019) and neural activity patterns seen in Wernicke's aphasia (Watanabe et al., 2025), thereby raising new questions regarding the essential validity of semantic understanding in LLMs.

In this study, we aim to elucidate the information processing characteristics underlying the three phases observed during the LLM pre-training process through internal dynamics analysis. Specifically, using a series of models at various stages of pre-training, we sampled the trajectories of internal states during the performance of a free-writing task as multivariate time-series data. To quantify the dynamics in terms of autocorrelation properties, we performed power spectral density (PSD) analysis on each dimension (Fig. 1c). This allowed us to measure the spectral exponent β (a standard metric where natural text typically exhibits $\beta \approx 0.5$ (Ebeling & Pöschel, 1994; Drożdż et al., 2016)) to evaluate both context-dependency preservation and dynamical transitions throughout learning. The key findings are as follows:

- **First Phase Shift (Phase A $\rightarrow$ B):** As non-linguistic output transitions into syntactically valid text, the PSD of internal state dynamics transitions from noise to a power-law spectrum with an exponent converging toward $\beta \approx 0.5$. This reflects the emergence of scale-invariant, weak long-range temporal correlations characteristic of natural language corpora.
- **Second Phase Shift (Phase B $\rightarrow$ C):** Synchronous with the sudden improvement in benchmark performance, the β distribution broadens into a heavy-tailed structure, marked by a sharp spike in intra-layer kurtosis around a constant layer-mean (≈ 0.5), signaling the emergence of specialized outlier units.
- **Token Selectivity in Outliers:** Qualitative inspection reveals that a subset of flatter outlier units exhibit high activation selectivity toward specific functional tokens or structural boundaries (e.g., punctuation and chunk-constraining words), providing preliminary evidence for potential temporal functional differentiation within hidden representations.

This study clarifies that the acquisition of emergent capabilities in LLMs is underpinned by the establishment of power-law behavior in the autocorrelation structure of internal dynamics, followed by the hierarchical differentiation of temporal processing. These results redefine the acquisition of emergent abilities and the three-phase transitions not merely as incremental improvements in benchmark scores, but as a process of qualitative transition in temporal information processing. This evolution suggests a transition beyond superficial statistical mimicry, offering a key perspective on how models may develop sophisticated computational mechanisms with complex, multi-scale temporal dynamics.

2 Related works

Early research on emergent abilities primarily centered on external metrics, specifically performance evaluations across various benchmark tasks (Wei et al., 2022; Ganguli et al., 2022; Srivastava et al., 2023). However, following the observation by Schaeffer et al. (Schaeffer et al., 2023) that these performance leaps might be an artifacts of discrete or nonlinear evaluation metrics, the focus has shifted toward elucidating the nature of emergent abilities from an internal perspective. In particular, as an approach to understanding how these abilities are acquired within models, research focusing on the formation of circuits responsible for specific algorithmic functions has advanced significantly. For instance, the identification of *induction heads*, which are considered the foundation of in-context

learning, has shown that specific heads coordinate at particular stages of training to suddenly acquire the ability to reuse in-context patterns (Olsson et al., 2022). Furthermore, research has highlighted the formation of syntactic attention involved in grasping grammatical structures (Chen et al., 2023). The quantization hypothesis further posits that individual discrete skills are acquired in stages (Michaud et al., 2023), suggesting that external leaps in performance are underpinned by the internal formation of mechanisms for processing specific linguistic structures. However, while these analyses have succeeded in identifying local computational units responsible for specific tasks, they have yet to explain vastly diverse emergent abilities in LLMs within a unified framework.

Several approaches for the internal analysis of LLMs have been proposed. These include *probing* (Tenney et al., 2019; Hewitt & Manning, 2019; Clark et al., 2019; Jawahar et al., 2019), which measures the maturity of internal information representations through predictive tasks; *causal tracing* (Meng et al., 2022), which assesses contributions to output by intervening in activation values; and *sparse autoencoders (SAE)* (Bricken et al., 2023; Cunningham et al., 2023; Gao et al., 2024), which aim to decompose internal states into interpretable concepts. However, these methods remain primarily static, targeting snapshots of activation at specific points. Although effective for localizing specific knowledge, they struggle to capture the dynamic interplay of how the system as a whole processes information and manages complex context as text generation evolves over time.

Analyzing internal dynamics offers a promising perspective to address the limitations of circuit-based and static analyses. By treating the LLM as a dynamical system, this approach captures the temporal evolution of internal states as multivariate time series, focusing on latent temporal structures that are often overlooked in existing research. Previous studies analysis have demonstrated that the emergence of transient chaos in encoder-based models may be linked to linguistic capabilities (Inoue et al., 2022). Furthermore, energy landscape analysis has demonstrated that the dynamical trajectories of LLMs share common mathematical properties with human brain activity (Watanabe et al., 2025). These findings suggest a potential convergence between artificial and biological information processing. In addition, recent works analyze dynamics across model scales and pre-training trajectories, demonstrating increased dynamical coherence in larger models (Zhang et al., 2025) alongside non-monotonic geometric shifts that track distinct learning phases Li et al. (2026). Furthermore, recent works leverage internal representation dynamics and structural entropy as auxiliary targets to shape reinforcement learning advantages, thereby incentivizing deeper exploratory reasoning (Cheng et al., 2026; Huang et al., 2026).

Power-law behavior and long-range temporal correlations, measurable via PSD analysis, are well-established statistical hallmarks of both natural language corpora and LLM outputs. Corpus analyses reveal that natural language universally exhibits weak long-range correlations ($\beta = 0.5$) (Ebeling & Pöschel, 1994; Montemurro & Pury, 2002; Altmann et al., 2012; Drożdż et al., 2016; Tanaka-Ishii, 2021). In recent years, similar analyses have been utilized to evaluate LLM outputs. For instance, research analyzing the statistical manifolds of token ratio distributions has revealed that they exhibit finite-dimensional fractal structures, which are distinct from simple Markov processes (Du & Tanaka-Ishii, 2024). Furthermore, comparisons of fractal dimensions have shown that LLM-generated text tends to exhibit a wider range of values than human-written text (Alabdulmohsin & Steiner, 2025). Additionally, evaluations using correlation dimensions suggest that LLMs acquire these self-similarities in a stepwise manner throughout the learning process (Du & Tanaka-Ishii, 2025).

3 METHODS

To ensure the transparency and universality of our analysis, we selected several open-source decoder-only language models. For Transformer-based models, we included OLMo-2-1124-7B (OLMo et al., 2024), Tri-7B (Han et al., 2025), SmolLM3-3B (Bakouch et al., 2025), Pythia-6.9B (Biderman et al., 2023), Llama-3.1-8B (Grattafiori et al., 2024), and Qwen2.5-7B (Team, 2024). In addition, we incorporated the Mamba-based Falcon-Mamba-7B (Zuo et al., 2024) and Zamba-7B (Glorioso et al., 2024), a hybrid model combining Transformer and Mamba architectures. For a subset of these models, we analyzed intermediate checkpoints saved during the training process. This allows for a retrospective tracking of how dynamic properties transform alongside the emergence abilities. Detailed specifications for each model and the list of analyzed checkpoints are provided in the Appendix A.

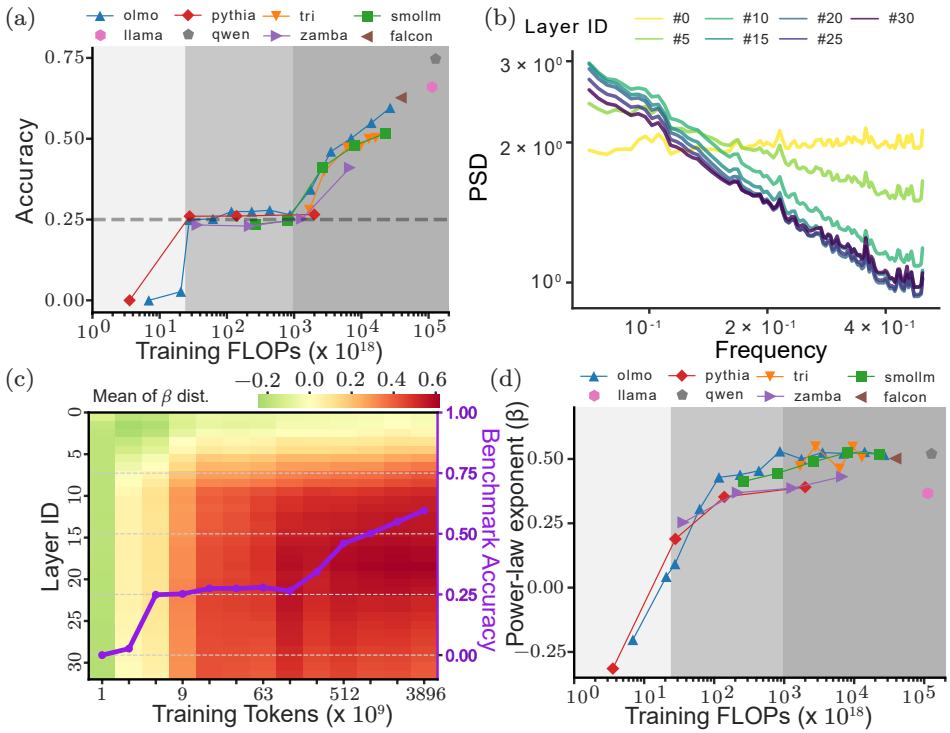

Figure 2: (a) MMLU scores accross models. The background shading represents the three distinct phases (phase A, B and C). (b) Ensemble-averaged PSD spectra across layers for the OLMo model in phase C. Power-law relationships were observed between frequency and spectral density, with the exponents increasing toward the deeper layers. (c) Evolution of the power-law index vs. benchmark performance during OLMo training. The color maps represent the values of the power-law exponent calculated across layers and training stages. The emergence of the power-law structure coincides precisely with the first phase shift. (d) Cross-model comparison. The same trend is observed across all tested models, with the layer-mean β averaged across layers 10 and deeper. In this analysis, the power-law indices are averaged over both the embedding dimensions and all layers.

To analyze internal dynamics, we extracted activation trajectories by providing a random 1-token prompt and allowing the model to generate 4,096 tokens autoregressively, resulting in a time series of 4,096 timesteps. This sequence length was chosen to respect the context limits of all evaluated models. From the resulting $L \times D$ (layers $\times$ dimensions) trajectories, we calculated the PSD and power-law exponent β for each neuron. The statistics reported in this study represent the ensemble average across all trials, capturing the model's inherent dynamical state while minimizing prompt-specific bias. For comparison, to evaluate the robustness of our findings across different input texts, we similarly examined the dynamics during prompt processing (without generating new tokens) using long context texts drawn from multiple domains, including the benchmark task problems evaluated in this study, yielding results consistent with those from free-form generation. These supplementary results are provided in the Appendix.

4 Experiments

4.1 Preliminary Experiment: Benchmark Evaluation

Prior to this analysis, in order to reinforce the validity of the three phases, we conducted benchmark measurements on models other than those employed in the study by Nagaki et al. (Nakagi et al., 2025). We evaluated the performance of each checkpoint using the massive multitask language understanding (MMLU) (Hendrycks et al., 2020). Following standard protocols, we measured 5-shot performance via the argmax of next-token likelihoods across the 57 subjects. To facilitate cross-model comparison, we utilize cumulative compute (Floating-point Operations; FLOPs (Ka-

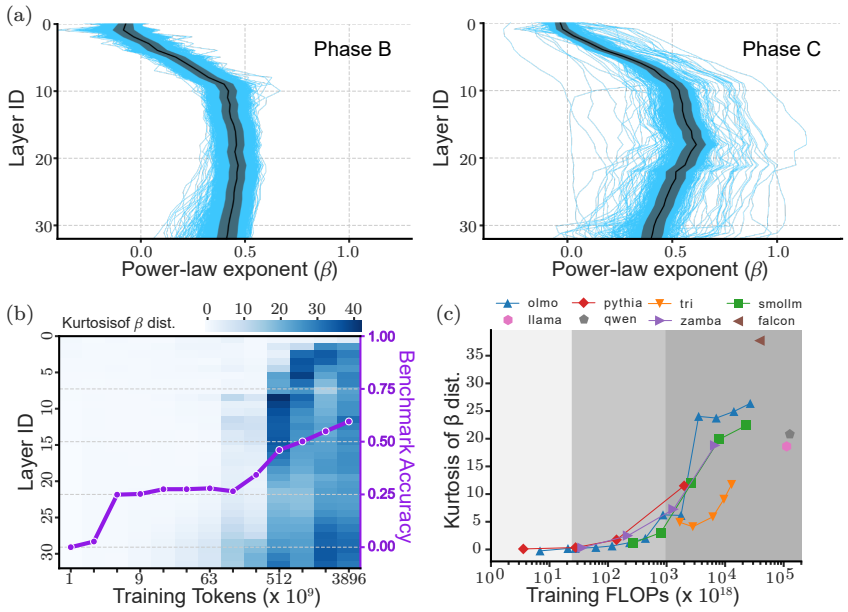

Figure 3: (a) Distribution of power-law exponents across layers and dimension in the OLMo phase B and phase C models represented by blue lines, with layer-wise means and standard deviations. While the mean β converges toward 0.5 in the middle layers for both phases, phase C exhibits a significantly higher diversity from the middle to deeper layers. (b) Evolution of intra-layer kurtosis vs. benchmark performance during OLMo training. A sharp spike in kurtosis coincides with the second phase shift, indicating the emergence of outlier nodes. (c) Cross-model comparison. A consistent trend is observed across all tested models, with the kurtosis averaged over all layers. The background shading represents the three distinct phases (phase A, B and C).

plan et al., 2020)) as the horizontal axis. Our measurements reveal discontinuous performance improvements as a function of model scale (Fig.2a). These results support the triple-phase transition proposed by Nakagi et al. (Nakagi et al., 2025), and we further confirmed the universality of this phenomenon across a broader range of model families. For the purpose of our analysis, we categorize the training process into the following three phases based on MMLU scores and FLOPSs:

- Phase A ($< 2.5 \times 10^{19}$ FLOPs): The model fails to follow the response format, producing invalid outputs.
- Phase B ($< 1.0 \times 10^{21}$ FLOPs): Accuracy remains at the chance level (approx. 25%) .
- Phase C ($> 1.0 \times 10^{21}$ FLOPs): The model demonstrates capabilities exceeding the random level.

We define the transition from phase A to B as the first phase shift and the transition from B to C as the second phase shift. In the following sections, we provide a detailed PSD-based analysis of the transformations in internal dynamics across these stages.

4.2 Self-organization of Power-law Statistics and the First Phase Shift

To clarify the fundamental properties of the internal dynamics, we first examined the PSD averaged across all hidden dimensions within each layer. The experimental results confirmed a prominent power-law behavior in the internal states of the model. Figure 2b illustrates the spectrum for the OLMo model (phase C). The log-log plot exhibits linear decay over a wide frequency range, indicating that the transitions of the internal states in LLMs follow a power law distribution. Notably, the emergence of a power-law characteristic coincides almost perfectly with the onset of the first phase shift (Fig. 2c). This pattern is consistently observed in both the OLMo and Pythia models where the first phase shift is identifiable (Fig. 2d). From a layer-wise perspective, in the phase B

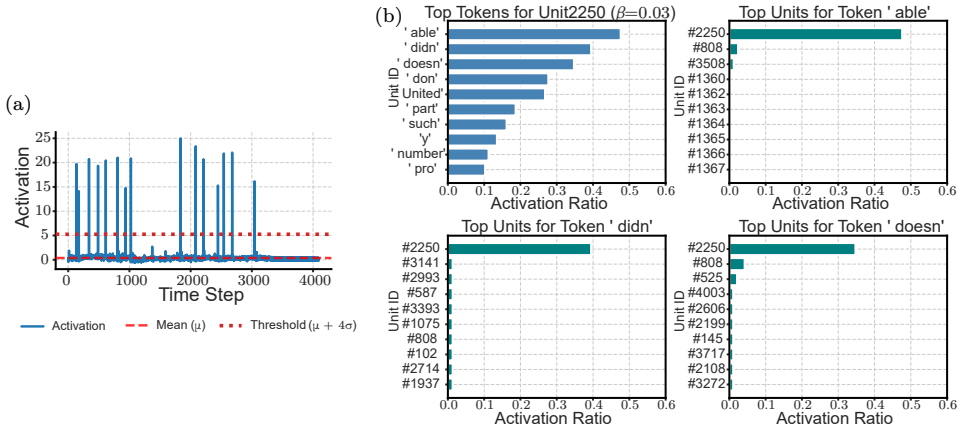

Figure 4: (a) Raw time-series activation of Unit 2250 in Layer 20 of OLMo, showing sharp, high-amplitude transient spikes. (b) Response ratio for each unit-token pair, defined as the fraction of token occurrences that trigger threshold-crossing activations. Unit 2250 shows high response ratios toward specific structural tokens such as "able", "didn", and "part" (blue bars). Conversely, when evaluating all units responsive to these tokens, Unit 2250 exhibits a uniquely dominant response ratio compared to any other unit (green bars).

and phase C models, a clear power-law decay is observed from the fifth layer onward, with the exponent converging toward $\beta \approx 0.5$. In contrast, shallower layers maintain a low exponent ($\beta \approx 0$), showing no significant temporal correlation even after training is complete. These results clarify that the first phase shift is characterized by the acquisition of temporal correlations with a specific intensity. Layer-wise averages of the exponent β for other models are presented in Appendix E.3. Across architectures, β consistently transitions from 0 toward ≈ 0.5 within approximately the first quarter of the model depth (e.g., layers 5–8 in a 32-layer model).

4.3 Increase in Power-Law Exponent Diversity and the Second Phase Shift

Moving beyond the layer-wise averages, we further investigate the diversity of internal dynamics across the hidden dimensions. Figure 3a illustrates the distribution of power-law exponents (β) for each layer at representative checkpoints during phases B and C. Our observations reveal that the transition from phase B to phase C is characterized by a broadening of the distribution. Specifically, while the most dimensions remain concentrated around the mean ($\beta \approx 0.5$), certain dimensions exhibit idiosyncratic exponents ranging from $\beta \approx 0$ to $\beta \approx 1$, resulting in a heavy-tailed distribution. To quantitatively evaluate this shift, we introduce kurtosis as a statistical metric to capture this increasing diversity. Figure 3b shows the layer-wise kurtosis for each checkpoint plotted against the MMLU score. Until the end of phase B, the kurtosis remains low, indicating a relatively uniform distribution of dynamics. However, as the model approaches the second phase shift, the kurtosis begins to rise sharply. Upon entering phase C, the kurtosis reaches extremely high values (exceeding 10) and continues to climb. Finally, to assess universality, we plotted the mean kurtosis averaged across all layers (Fig. 3c). Our analysis confirms that the abrupt surge in kurtosis following the second phase shift is a consistent behavior across all tested architectures, signaling the emergence of outlier nodes with distinct dynamical properties. These experiments clarify that the second phase shift is characterized by the emergence of nodes with highly specialized dynamics. The full β distributions across all examined models and training checkpoints are compiled in Appendix E.3.

4.4 Selective Response to Specific Tokens in Flatter Outliers

To observe potential linguistic functions associated with outlier nodes, we conducted a preliminary inspection of the raw activation dynamics of these outliers. As a qualitative case study focused specifically on Layer 20 of the final OLMo model (Figure 4), we monitored internal hidden state dynamics under the same experimental conditions described previously. Defining a response threshold as three standard deviations above the mean (mean $+ 4\sigma$) for each dimension, we quantified

the response ratio for each unit-token pair as the ratio of threshold-exceeding activations to the total occurrences of that token. Evaluating units with high response probabilities revealed distinct token selectivities, though notably confined to a small subset of flatter outliers. For instance, Unit 2250—one of the flatter outliers ($\beta = 0.03$)—selectively responds to tokens that constrain subsequent context or form structural chunks (e.g., " able", " didn", and " doesn"; Figure 4b, blue bars). Specifically, Unit 2250 achieves response ratios exceeding 0.30 for these target tokens, whereas no other unit in the model exceeds 0.10 (Figure 4b, green bars), highlighting its exceptional specificity. Similarly, Unit 3130 ($\beta = 0.29$) selectively activates for tokens containing sentence-ending punctuation, while Unit 2928 ($\beta = 0.24$) responds preferentially to pronouns (see Appendix F for full profiles).

5 Discussion

Regarding the first phase shift, we indicate that the power-law exponent β converges to approximately 0.5. Unlike common physical systems where $\beta = 0$ (white noise) or $\beta = 1$ ($1/f$ noise), $\beta \approx 0.5$ represents a state of moderate long-range dependency. This value aligns with the weak long-range correlations widely reported in quantitative linguistics (Ebeling & Pöschel, 1994; Montemurro & Pury, 2002; Altmann et al., 2012; Drożdż et al., 2016; Tanaka-Ishii, 2021). Statistical analyses across various linguistic scales—such as word recurrence times and sentence length fluctuations—frequently reveal exponents characteristic of weak long-range dependencies. The emergence of an exponent in this range implies that the model's internal dynamics come to reflect universal temporal properties inherent in natural language during the first phase shift. Consistent with this internal shift, output-level metrics confirm that long-range correlations within generated texts are also enhanced during the Phase A→B transition (see Appendix G).

Regarding the second phase shift, we observed that the sharp increase in intra-layer kurtosis occurs in synchronization with the onset of emergent capabilities. This signifies the appearance of outlier nodes whose temporal correlations deviate significantly from the baseline ($\beta \approx 0.5$). This observation suggests that the model's internal processing has evolved to mirror the multi-scale structures inherent in natural language, spanning from local lexical items to global context. We hypothesize that this dynamic divergence reflects a functional specialization across time scales: while high-β units are hypothesized to facilitate long-range context preservation over extended sequences, low-β units specialize in immediate, localized operations.

Qualitative inspection supports this functional specialization in low-β units. Specifically, these low-β units activate transiently in response to functional tokens or structural boundary markers (e.g., chunk-constraining words, punctuation, and pronouns). Because their computational role is restricted to signaling immediate syntactic events rather than tracking historical context, their activations lack temporal persistence, naturally yielding a flat power-law spectrum. Connecting spectral exponent properties to distinct token-processing behaviors thus bridges internal dynamic statistics with linguistic functionality, revealing how LLMs achieve hierarchical context processing through internal temporal differentiation.

These dynamic shifts offer novel conceptual perspectives on neural representations. First, while functional specialization in LLMs has traditionally been conceptualized along the vertical, layer-wise axis, our observation of temporal correlation variance during the second phase shift suggests the possibility of intra-layer functional differentiation across hidden dimensions. This node-wise variation in temporal properties, alongside its co-occurrence with emergent task performance, points toward a more nuanced view of LLM processing where functional roles diversify within individual layers. Second, while phase transitions are frequently invoked to conceptualize learning dynamics in machine learning, they are classically framed around global order and synchronization (e.g., Kuramoto models). Our findings suggest that simple synchronization alone cannot account for the second phase shift, which instead manifests as a dynamic divergence—captured by a sharp spike in kurtosis—highlighting how LLMs may develop advanced capabilities through increased internal diversity rather than uniform alignment.

Practically, these findings open new avenues for intrinsic model monitoring. Because these spectral signatures reflect structural learning milestones independently of specific task prompts or evaluation formats, calculating PSD statistics from free-form generation could serve as a streamlined, benchmark-free diagnostic tool for tracking LLM pre-training trajectories in real time.

A primary limitation of this study is that our findings remain observational, demonstrating the co-occurrence of temporal dynamic shifts and performance leaps without definitively establishing a causal relationship. To move beyond observational correlations, future research must incorporate interventional frameworks to establish a direct causal link between internal dynamics and emergent model capabilities. In particular, deeper verification is required to test whether steeper outlier nodes (exhibiting large β) actively drive long-range context preservation, thereby establishing their mechanistic role in enabling emergent task performance.

6 CONCLUSION

In this study, we demonstrate that the emergent abilities during the learning process of LLMs are characterized by the transformation of temporal statistics within their internal dynamics. By detailing the three-stage transition observed in MMLU performance through PSD analysis, we confirmed that the first phase shift marks a shift from a state lacking long-range dependency to power-law dynamics characterized by $\beta \approx 0.5$, where the model acquires the long-range dependencies inherent in natural language. Furthermore, we revealed that the second phase shift is characterized by a spike in the intra-layer kurtosis of the power-law exponents, signaling the emergence of outlier nodes specialized in specific temporal scales.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code, drafting parts of a research paper, brainstorming, sourcing/searching for information, editing a research paper to improve readability, identifying relevant literature, formatting references, and suggesting a structure for a research paper. We have reviewed all AI-assisted work. Specifically, all LLM-generated code was verified and tested for correctness by the authors, and all translated or edited text was thoroughly reviewed to maintain scientific accuracy. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, all large language models evaluated in this work are open-source and publicly accessible via the Hugging Face `transformers` library, with specific model architectures and checkpoint details summarized in Appendix A.1. All evaluation corpora are likewise freely available. The complete codebase for inference and temporal signal processing (e.g., PSD calculation and β fitting) is provided in the supplementary materials. Crucially, we also include the exact generated token sequences from our experiments in the repository, enabling deterministic, step-by-step reproduction of all reported figures and scaling trajectories. Detailed experimental protocols, signal analysis configurations (such as frequency ranges for power-law fitting and thresholding parameters), and compute hardware specifications are comprehensively described in Appendix E.

REFERENCES

Ibrahim Alabdulmohsin and Andreas Steiner. A tale of two structures: Do llms capture the fractal complexity of language? *arXiv preprint arXiv:2502.14924*, 2025.

Eduardo G Altmann, Giampaolo Cristadoro, and Mirko Degli Esposti. On the origin of long-range correlations in texts. *Proceedings of the National Academy of Sciences*, 109(29):11582–11587, 2012.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. `https://huggingface.co/blog/smollm3`, 2025.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2, 2023.

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

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in mlms. *arXiv preprint arXiv:2309.07311*, 2023.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30377–30385, 2026.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Stanisław Drożdż, Paweł Oświęcimka, Andrzej Kulig, Jarosław Kwapień, Katarzyna Bazarnik, Iwona Grabska-Gradzińska, Jan Rybicki, and Marek Stanuszek. Quantifying origin and character of long-range correlations in narrative texts. *Information Sciences*, 331:32–44, 2016.

Xin Du and Kumiko Tanaka-Ishii. Correlation dimension of natural language in a statistical manifold. *Physical Review Research*, 6(2):L022028, 2024.

Xin Du and Kumiko Tanaka-Ishii. Correlation dimension of auto-regressive large language models. *arXiv preprint arXiv:2510.21258*, 2025.

Werner Ebeling and Thorsten Pöschel. Entropy and long-range correlations in literary english. *Europhysics Letters*, 26(4):241, 1994.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 1747–1764, 2022.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Paolo Glorioso, Quentin Anthony, Yury Tokpanov, James Whittington, Jonathan Pilault, Adam Ibrahim, and Beren Millidge. Zamba: A compact 7b ssm hybrid model. *arXiv preprint arXiv:2405.16712*, 2024.

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

Sungjun Han, Juyoung Suk, Suyeong An, Hyungguk Kim, Kyuseok Kim, Wonsuk Yang, Seungtaek Choi, and Jamin Shin. Trillion 7b technical report. *arXiv preprint arXiv:2504.15431*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, DDL Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10, 2022.

Fanding Huang, Guanbo Huang, Xiao Fan, Yi He, Xiao Liang, Xiao Chen, Qinting Jiang, Faisal Nadeem Khan, Jingyan Jiang, and Zhi Wang. Semantic-space exploration and exploitation in rlvr for llm reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38402–38449, 2026.

Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. Codesearchnet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*, 2019.

Katsuma Inoue, Soh Ohara, Yasuo Kuniyoshi, and Kohei Nakajima. Transient chaos in bidirectional encoder representations from transformers. *Physical Review Research*, 4(1):013204, 2022.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Eric H Lenneberg. The biological foundations of language. *Hospital Practice*, 2(12):59–67, 1967.

Melody Li, Kumar Krishna Agrawal, Arna Ghosh, Komal Teru, Adam Santoro, Guillaume Lajoie, and Blake Richards. Tracing the representation geometry of language models from pretraining to post-training. *Advances in Neural Information Processing Systems*, 38:54691–54724, 2026.

Shuhong Liu, Nozomi Akashi, Qingyao Huang, Yasuo Kuniyoshi, and Kohei Nakajima. Exploiting chaotic dynamics as deep neural networks. *Physical Review Research*, 7(3):033031, 2025.

Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35:34651–34663, 2022.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36:28699–28722, 2023.

Marcelo A Montemurro and Pedro A Pury. Long-range fractal correlations in literary corpora. *Fractals*, 10(04):451–461, 2002.

Yuko Nakagi, Keigo Tada, Sota Yoshino, Shinji Nishimoto, and Yu Takagi. Triple phase transitions: Understanding the learning dynamics of large language models from a neuroscience perspective. *arXiv preprint arXiv:2502.20779*, 2025.

Katherine Nelson. Structure and strategy in learning to talk. *Monographs of the society for research in child development*, pp. 1–135, 1973.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

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

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in neural information processing systems*, 36:55565–55581, 2023.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Youran Sun and Babak Haghighat. Phase transitions in large language models and the $o(n)$ model. *arXiv preprint arXiv:2501.16241*, 2025.

Kumiko Tanaka-Ishii. *Statistical universals of language*. Springer, 2021.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Takamitsu Watanabe, Katsuma Inoue, Yasuo Kuniyoshi, Kohei Nakajima, and Kazuyuki Aihara. Comparison of large language model with aphasia. *Advanced Science*, 12(22):2414016, 2025.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Shimin Zhang, Ziyuan Ye, Yinsong Yan, Zeyang Song, Yujie Wu, and Jibin Wu. Koopstd: Reliable similarity analysis between dynamical systems via approximating koopman spectrum with timescale decoupling. In *Forty-second International Conference on Machine Learning*, 2025.

Micha
l Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3530–3534, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1561.

Jingwei Zuo, Maksim Velikanov, Dhia Eddine Rhaiem, Ilyas Chahed, Younes Belkada, Guillaume Kunsch, and Hakim Hacid. Falcon mamba: The first competitive attention-free 7b language model. *arXiv preprint arXiv:2410.05355*, 2024.

A MODEL DETAILS

We utilized the following models available through the Hugging Face Transformers library for our experiments. Model specifications were cross-referenced with the official model cards and technical reports, while the parameter counts were confirmed via the library's metadata.

The total compute cost during pre-training, in terms of FLOPs, was estimated using on the following formulation proposed Kaplan et al. (Kaplan et al., 2020):

$$C \approx 6ND, \tag{1}$$

where N and D denotes the number of non-embedding parameters and the total number of training tokens, respectively.

To ensure a balance between linguistic diversity and structural coherence during generation, we employed a standard stochastic decoding strategy with a temperature of $T = 1.0$. We further applied Top-k filtering ($k = 50$) and Top-p (nucleus) sampling ($p = 0.95$). These decoding conditions were maintained consistently across all models and checkpoints. Inference was performed on a hardware setup featuring NVIDIA GeForce RTX 3090 and an NVIDIA Tesla V100S-PCIE-32GB.

Table A.1: Model parametors and other details

Model	**Non-embedding Params**	**Training Tokens**	**Context Length**	**Num of Heads**
OLMo	6,887,575,555	3.9T (3,896B)	4,096	32
Tri	7,001,083,904	2T	32,768	32
Pythia	6,650,732,544	0.3T (299B)	2,048	32
SmolLM	2,812,430,336	8T (8,118B)	8,192	32
Llama	7,504,924,672	15T	128,000	32
Qwen	7,070,619,136	18T	128,000	28
Zamba	7,113,706,496	0.95T	4096	77
Falcon	7,006,326,784	5.5T	8,192	65

Table A.2: Selected checkpoints of OLMo

Checkpoints	Training tokens (B)
stage1-step150-tokens1B	1
stage1-step600-tokens3B	3
stage1-step900-tokens4B	4
stage1-step2000-tokens9B	9
stage1-step4000-tokens17B	17
stage1-step8000-tokens34B	34
stage1-step15000-tokens63B	63
stage1-step30000-tokens126B	126
stage1-step61000-tokens256B	256
stage1-step122000-tokens512B	512
stage1-step244000-tokens1024B	1024
stage1-step488000-tokens2047B	2047
stage1-step928646-tokens3896B	3896

Table A.3: Selected checkpoints of Tri

Checkpoints	Training tokens (B)
0000120000	240
0000200000	400
0000440000	880
0000680000	1360
0000920000	1840
0001160000	2320

Table A.4: Selected checkpoints of SmolLM

Checkpoints	Training tokens (B)
stage1-step-40000	94
stage1-step-120000	283
stage1-step-400000	944
stage1-step-1200000	2832
stage1-step-3440000	8118

Table A.5: Selected checkpoints of Pythia

Checkpoints	Training tokens (B)
stage1-step-120000	0.5
stage1-step-400000	4
stage1-step-1200000	20
stage1-step-3440000	299

Table A.6: Selected checkpoints of Zamba

Checkpoints	Training tokens (B)
iter2500	5
iter15000	31
iter90000	185
iter462070	950

B BENCHMARK DETAILS

MMLU (Hendrycks et al., 2020) is a comprehensive benchmark designed to evaluate the world knowledge and problem-solving capabilities of LLMs. Covering 57 subjects across domains such as the humanities, social sciences, STEM, it serves as a rigorous testbed. The auxiliary_train subset was intentionally excluded from the present analysis.

The following are multiple choice questions about anatomy.

What is the embryological origin of the hyoid bone?
(A) The first pharyngeal arch (B) The first and second pharyngeal arches (C) The second pharyngeal arch (D) The second and third pharyngeal arches
Answer: D

Which of these branches of the trigeminal nerve contain somatic motor processes?
(A) The supraorbital nerve (B) The infraorbital nerve (C) The mental nerve (D) None of the above
Answer: D

The pleura
(A) have no sensory innervation. (B) are separated by a 2 mm space. (C) extend into the neck. (D) are composed of respiratory epithelium.
Answer: C

In Angle's Class II Div 2 occlusion there is
(A) excess overbite of the upper lateral incisors. (B) negative overjet of the upper central incisors. (C) excess overjet of the upper lateral incisors. (D) excess overjet of the upper central incisors.
Answer: C

Which of the following is the body cavity that contains the pituitary gland?
(A) Abdominal (B) Cranial (C) Pleural (D) Spinal
Answer: B

A lesion causing compression of the facial nerve at the stylomastoid foramen will cause ipsilateral
(A) paralysis of the facial muscles. (B) paralysis of the facial muscles and loss of taste. (C) paralysis of the facial muscles, loss of taste and lacrimation. (D) paralysis of the facial muscles, loss of taste, lacrimation and decreased salivation.
Answer:

Figure B.5: MMLU example.

C EXAMPLES OF GENERATIVE OUTPUTS

We present sample outputs generated by the models in phases A, B, and C. Furthermore, we provide outputs from the phase C model following the ablation of the 64 units with the steepest, flattest, and median β (as detailed in Section 4.4).

Model output (OLMo-1B phase A)

```
Kurt Gildisch a their their to with and (.  but a on up, at the two
the time with the other of of an who (I. This is ( that the 1.  You
is ( than at the about in most the " got that in 0
0.  6, can • 7 4.  were 7 in 2.  The) to at the2.  This a 2 is.  In
the the to 2, not 1.  That 3:  only 4.  It is not no 2.  20.3 from
a does of a at 200 does 5.  5.  They is is of with an by 1.  This's
a 8 for the first at 3.  2.:  " be the other lot.  future, 1, but
that is more 4.  2.  The.0 is a
2.  As do new have been 4.5 (2- future.  It's some is the lot, 3..,
4 was only.  (11.  0, it will 3.  2.?
:  4.3.  1.  ( future.3) than this.  2 is a these is that the time
6.  line.  only.4.  does 2.  1.  The 2.1.  number 4 is all 11 for
1.  0.3.  (8:  lot.0 by 4.  ( does with:  0
2.  12 to 2 for 1.  until for 2
1.0 and the 4.3.3.3 1 is &
1.4.0.  :  12.3
number3.  In 2.  (2 of 3.  1.  The --.  in 1.  lot.6, 1.0 in
does.9...  4.0.  In the 3.3:  2.  mean.9 is the (7.  lot.  -- with.
- 1..  only, in 2.?
```

Model output (OLMo-63B phase B)

```
Kurt Gildisch:  The best is yet to come.  Kurt Gildisch
It is a wonderful time in business to be making computers.  Every
four to six months, I am amazed by the number of new designs that
show up.  In recent months we have seen some dramatic improvements
to both the size and price of integrated circuitry.
As a result, the PC is no longer a computer that you buy because
you want a computer.  It is a computer that you buy to get the
most out of your PC, as you plan to use it.  A lot of people use
their computer to play games on-line and have been trained to see
that the best graphics in the games do not really matter unless the
game is extremely difficult.  These are mostly games for the mass
market.  In these games it is not so important to have the best
graphics.  So in these games it is not really a matter of having
the best graphics, but of having better graphics than other people
in the game.
```

Model output (OLMo-3896B phase C)

```
Kurt Gildisch, 34, died of head injuries suffered in the fall
outside the 20th and Oakley Avenue home where he lived with his
wife and three children.  He was taken by ambulance to a nearby
hospital, where he died that morning.
The coroner, who was still gathering information Thursday, said
Gildisch's blood-alcohol level was 0.22 percent, more than twice
the legal limit.  The cause of death was determined to be a "fall
down a stairway with blunt trauma head injuries," said the coroner,
Michael Baden.
Baden said a preliminary investigation determined that Gildisch
fell down the stairs by himself.<|endoftext|>Sign up
Here's how it works:
1.  Anybody can ask a question
2.  Anybody can answer
I am looking at some solutions for making money (or more precisely,
generating a profit) via the Forex or the stock markets.
What is the most successful method to make money in these
situations?  Is it to go long or short on something and buy at a
lower price and sell at a higher one?
Or is it to use leverage and make bets on the movements on the
market?
Or is there some other way?  I have heard rumours about trading the
news, but have not really researched that so far.
What is the most successful strategy?  The question is based on the
assumption that the market will not go up long term.  I am aware
that this can happen, but I am concerned that it is more likely to
fluctuate in the long term.
share|improve this question
up vote 3 down vote accepted
It's true that market is more likely to fluctuate rather than go
up, but what is called up is just fluctuation with higher ratio of
the price going up.  You can profit on fluctuation to the extent
you can forecast that fluctuation.  If you are doing it on small
scale, then probably fundamental news analysis will work.  There
are many professional investors doing that.  For more, I recommend
you to read Taleb's Fooled by randomness, or you can read a summary
of it here.  There are others who prefer to base their investment
decision based on technical analysis, which also seem to work at
times.  In either case, it might be better to play with a small
amount of money to start with.  This way, if you loose your money,
you won't lose your family too :).
```

D OTHER DYNAMICAL INDICATORS

To complement the spectral characterization via PSD, we evaluated several foundational time-domain indicators on the internal dynamics of the OLMo model during free generation. For an activation time series $x_i(t)$ of unit i over T tokens, we quantified four primary metrics:

- **Time Average (μ):** Represents the baseline activation magnitude, defined as $\mu = \frac{1}{T}\sum_{t=1}^{T} x(t)$.
- **Time Variance (σ^2):** Measures the energy of temporal fluctuations around the mean, defined as $\sigma^2 = \frac{1}{T}\sum_{t=1}^{T}(x(t)-\mu)^2$.
- **Peak-to-Peak Amplitude:** Quantifies the dynamic range of activation, given by $A_{\text{p-p}} = \max_t x(t) - \min_t x(t)$.
- **Crest Factor:** Indicates the presence of bursty or extreme activation spikes relative to the root-mean-square (RMS) level, defined as $\frac{\max_t |x_i(t)|}{x_{\text{rms}}}$, where $x_{\text{rms}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} x_i(t)^2}$.

As illustrated in figure D.6 to D.9, while these scalar metrics provide a simpler time-domain view compared to the spectral exponent β, certain indicators exhibit distinct shifts across training checkpoints. These changes effectively mark the transitions between training phases, demonstrating that the macroscopic phase shifts in internal dynamics leave identifiable footprints even in basic time-domain statistics.

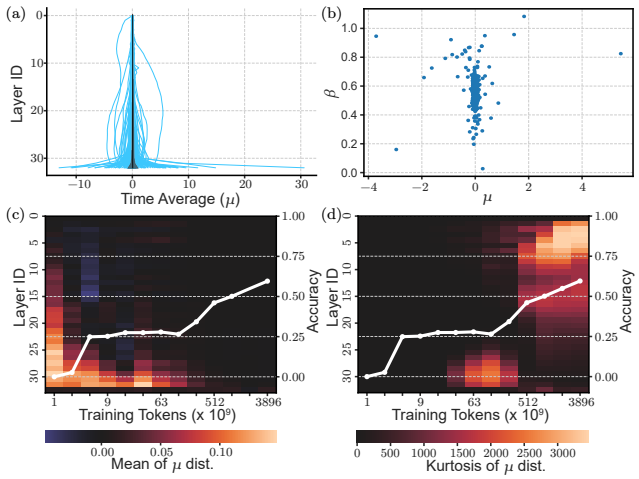

Figure D.6: (a) Layer-wise values of activation time averages across all layers in the final OLMo model. (b) Correlation between activation time average and power-law exponent (β) for individual units in Layer 20 of the final model. (c) Trajectory of the intra-layer mean of activation time averages plotted alongside benchmark performance across training checkpoints. (d) Trajectory of the intra-layer kurtosis of activation time averages versus benchmark performance across training checkpoints.

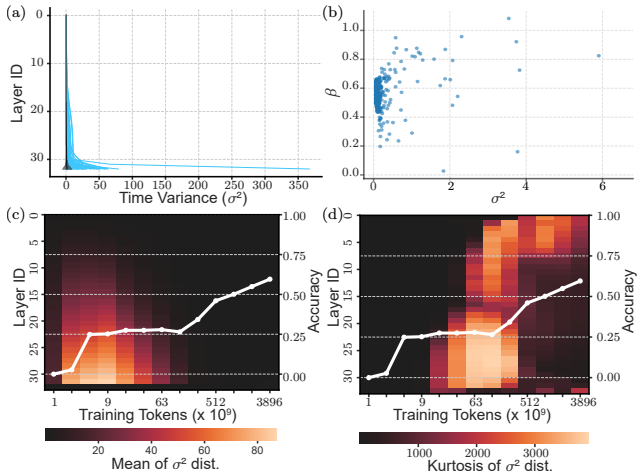

Figure D.7: (a) Layer-wise values of activation time variances across all layers in the final OLMo model. (b) Correlation between activation time variance and power-law exponent (β) for individual units in Layer 20 of the final model. (c) Trajectory of the intra-layer mean of activation time variances plotted alongside benchmark performance across training checkpoints. (d) Trajectory of the intra-layer kurtosis of activation time variances versus benchmark performance across training checkpoints.

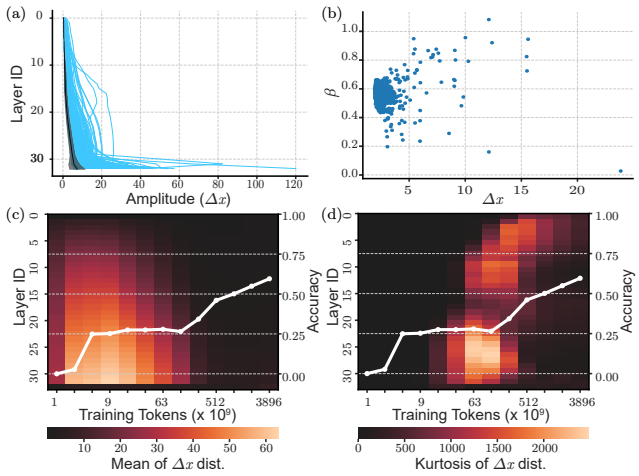

Figure D.8: (a) Layer-wise values of activation amplitudes across all layers in the final OLMo model. (b) Correlation between activation amplitude and power-law exponent (β) for individual units in Layer 20 of the final model. (c) Trajectory of the intra-layer mean of activation amplitudes plotted alongside benchmark performance across training checkpoints. (d) Trajectory of the intra-layer kurtosis of activation amplitudes versus benchmark performance across training checkpoints.

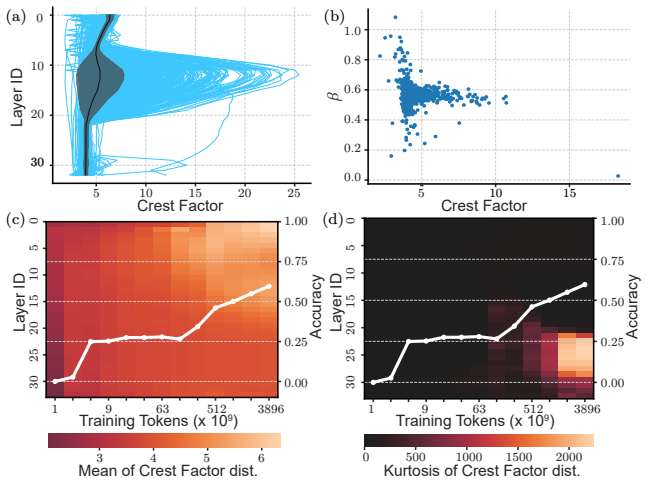

Figure D.9: (a) Layer-wise values of activation crest factors across all layers in the final OLMo model. (b) Correlation between activation crest factor and power-law exponent (β) for individual units in Layer 20 of the final model. (c) Trajectory of the intra-layer mean of activation crest factors plotted alongside benchmark performance across training checkpoints. (d) Trajectory of the intra-layer kurtosis of activation crest factors versus benchmark performance across training checkpoints.

E Details of PSD experiments

E.1 Spectral Estimation and Power-Law Exponents

For each activation time series $\{x_t^{(l,d)}\}$, we first standardized the data to ensure that the analysis focuses on the temporal structure rather than the absolute magnitude of activations. We then calculated the PSD, denoted as $\hat{P}_{l,d}(f)$, using Welch's method. To balance signal stationarity with low-variance estimation, periodograms were averaged using a Hann window with a segment length of 256 and an overlap of 50%. To quantify the long-range temporal correlations, we derived the power-law exponent β by performing linear regression on the log-log scale, based on the following relationship:

$$\log \hat{P}_{l,d}(f) \approx -\beta_{l,d} \log f + C. \tag{2}$$

The fitting is executed via SciPy's linregress. We filtered the frequencies to remove ultra-low frequency (> 0.025) and near Nyquist artifacts (~ 0.5), yielding 110 data points for each regression. Finally, the exponent β for each neuron is averaged over multiple independent trials.

To evaluate the non-Gaussianity of the activation distributions, the excess kurtosis was calculated for each activation time series, defined as follows:

$$g_2 = \frac{\mu_4}{\sigma^4} - 3, \tag{3}$$

where μ_4 and σ denote the fourth central moment and standard deviation, respectively. In this convection, the excess kurtosis is zero or Gaussian distributions, with positive values indicating heavy-tailed characteristics. As with the PSD analysis, values were averaged across all 50 trials.

E.2 Fitting Examples and Validation of Spectral Exponents

To assess the validity of the power-law characterization, we calculated the Root Mean Square Error (RMSE) of the linear regression fitted to the PSD in the log-log domain. Since all activation time series were standardized (zero mean and unit variance) prior to spectral analysis, the resulting RMSE provides a normalized, reliable metric for evaluating the goodness of fit across units and layers.

Fig. E.10(c) illustrates the layer-wise RMSE values for the Phase C OLMo model. Across all layers, nearly all units exhibit remarkably low RMSE values, remaining almost strictly below 0.1. This indicates that the spectral dynamics across the network are well-approximated by a linear fit in the log-log space, validating the power-law assumption.

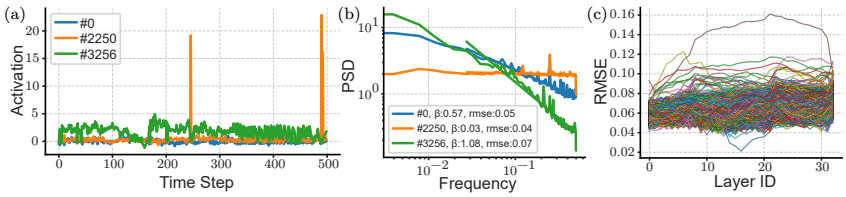

Figure E.10: Validation of PSD fitting and power-law dynamics for the final OLMo model. (a) Representative examples of raw activation time series from Layer 20. (b) Corresponding power spectral densities and their linear fits in log-log space, illustrating $1/f^{\beta}$-like scaling. (c) Layer-wise distribution of Root Mean Square Error (RMSE) of the linear fits across all units. Because activation time series were standardized (zero mean and unit variance), consistently low RMSE values (mostly below 0.1) across all layers demonstrate robust adherence to power-law dynamics.

E.3 Additional Figures

Figures E.11 through E.16 show the layer-wise distributions of the power-law exponent across various training checkpoints for each model evaluated in the main experiments. These plots provide a detailed overview of the training trajectory of spectral exponents throughout training across different model architectures and dataset scales.

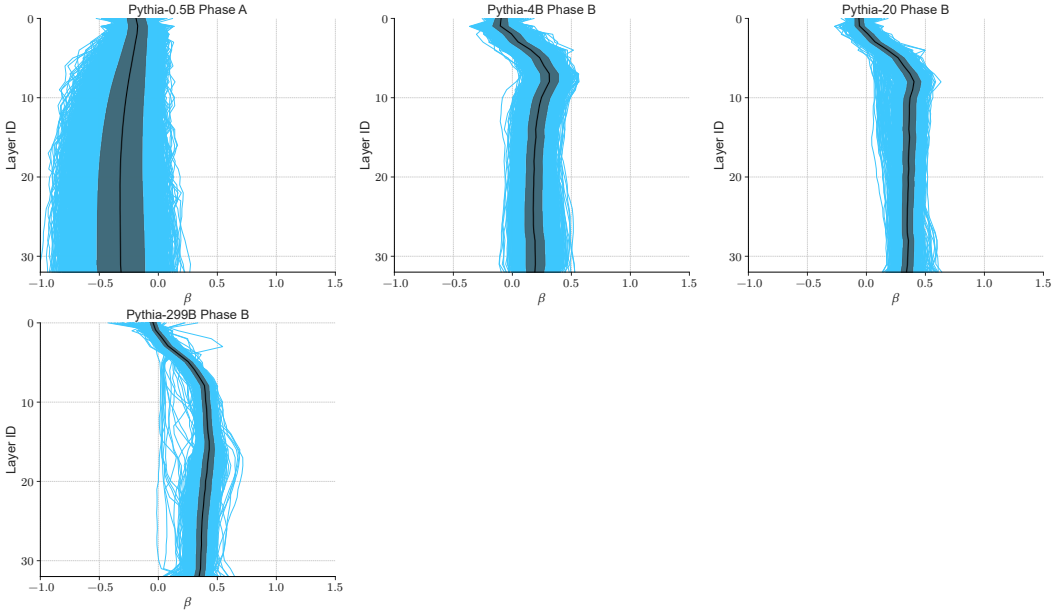

Figure E.11: β distribution of Pythia.

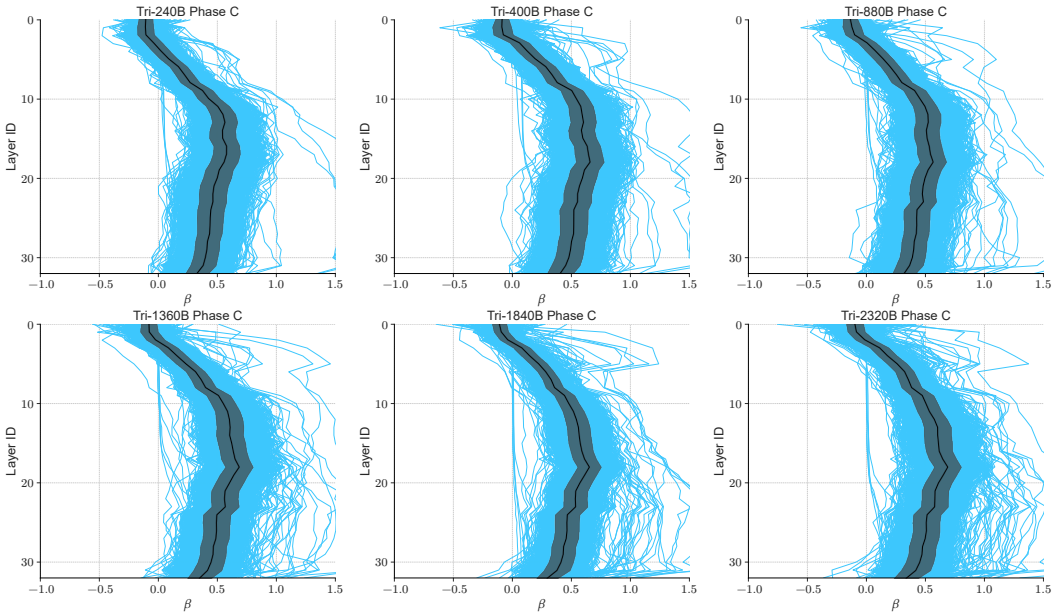

Figure E.12: β distribution of Tri.

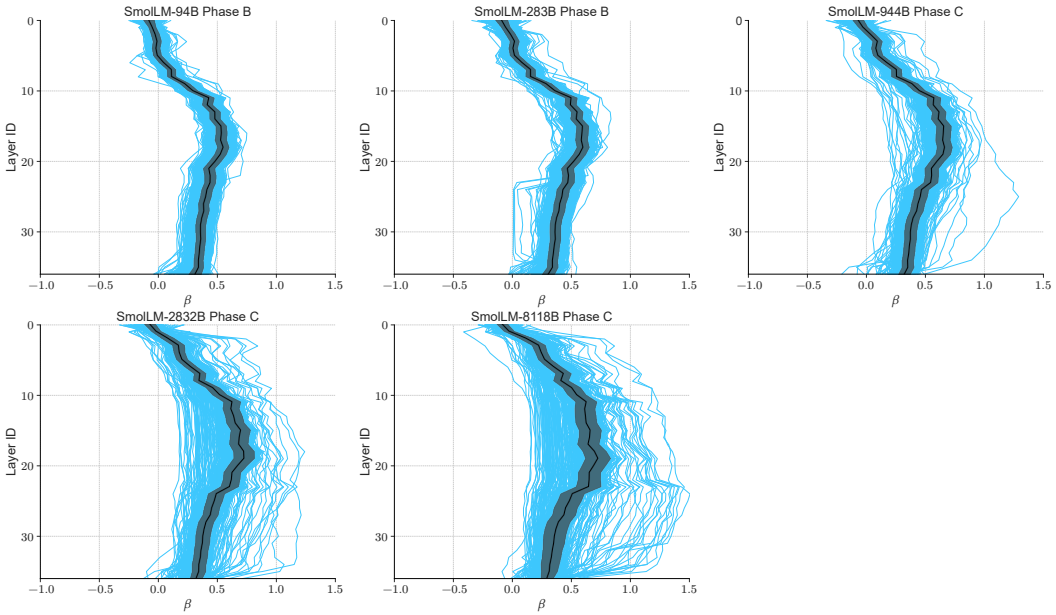

Figure E.13: β distribution of SmolLM.

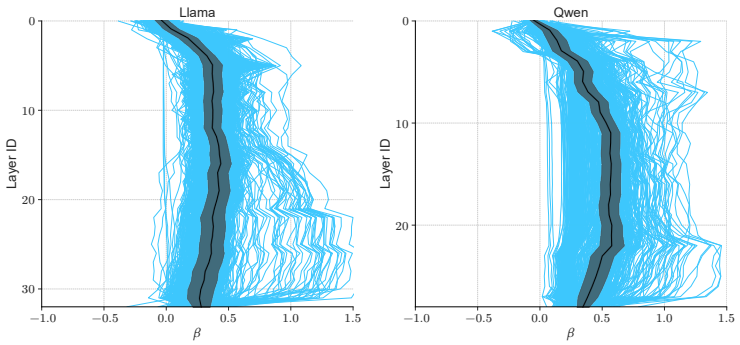

Figure E.14: β distribution of Llama and Qwen.

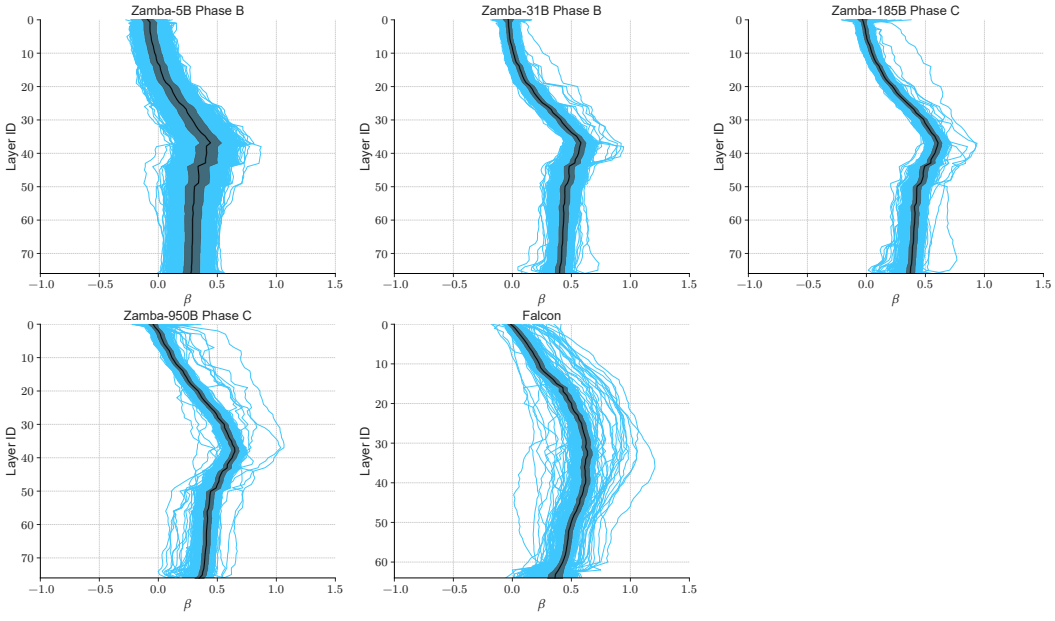

Figure E.15: β distribution of Zamba and Falcon.

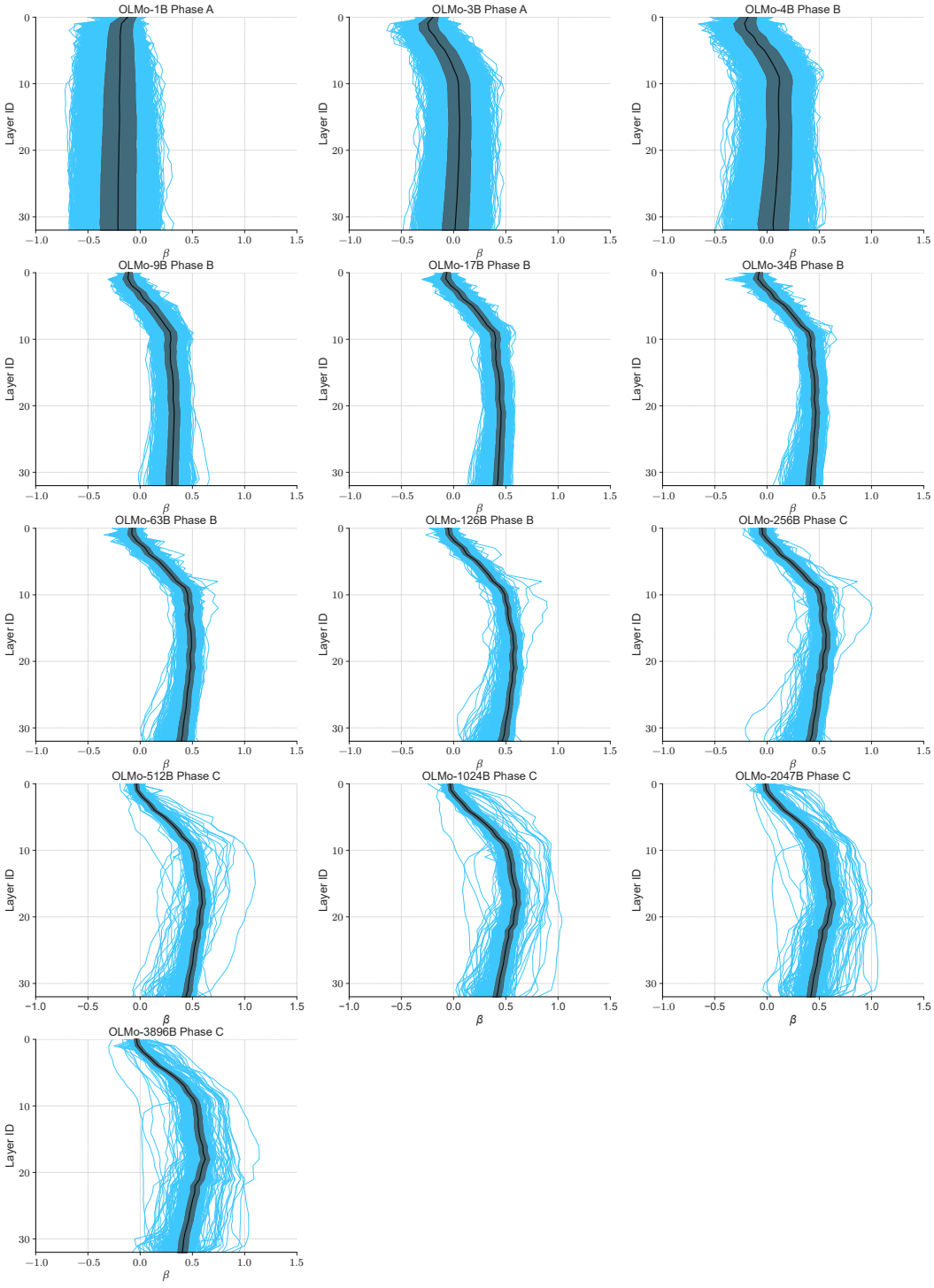

Figure E.16: β distribution of OLMo.

E.4 Encoding Dynamics and Domain Robustness

To verify the domain robustness of our power-spectral density findings and assess whether similar temporal characteristics emerge during benchmark evaluation, we analyzed the model's internal dynamics under an encoding regime (i.e., processing full input sequences) rather than autoregressive generation.

We evaluated input sequences across five distinct domains:

1. **Benchmark prompts:** Benchmark questions from MMLU (Hendrycks et al., 2020),
2. **Narrative prose:** Literary texts extracted from Project Gutenberg,
3. **Official documents:** Administrative text from the UN Parallel Corpus (Ziemski et al., 2016),
4. **Source code:** Programming code from CodeSearchNet (Husain et al., 2019), and
5. **Random sequences:** Synthetic sequences sampled uniformly at random from the model's vocabulary.

For each domain, text snippets were concatenated at the token level to reach a fixed sequence length of 2,048 tokens, and experiments were conducted across 10 independent trials.

The spectral dynamics observed under this encoding setup reinforce our main generation-based findings (see fig E.20):

- **Structured Text Domains:** The initial shift of the mean β toward ≈ 0.5 during the first phase shift and the subsequent rise in kurtosis during the second phase shift were consistently reproduced across all natural text domains. Notably, for domains composed of short, discrete snippets (e.g., MMLU prompts and programming code), while the absolute magnitude of kurtosis remained smaller than in continuous prose, its upward trend in the second phase was clearly preserved.
- **Random Sequences (Control):** In contrast, for random token sequences, β remained near zero ($\beta \approx 0$) even after full training, showing no emergence of power-law dynamic structures.

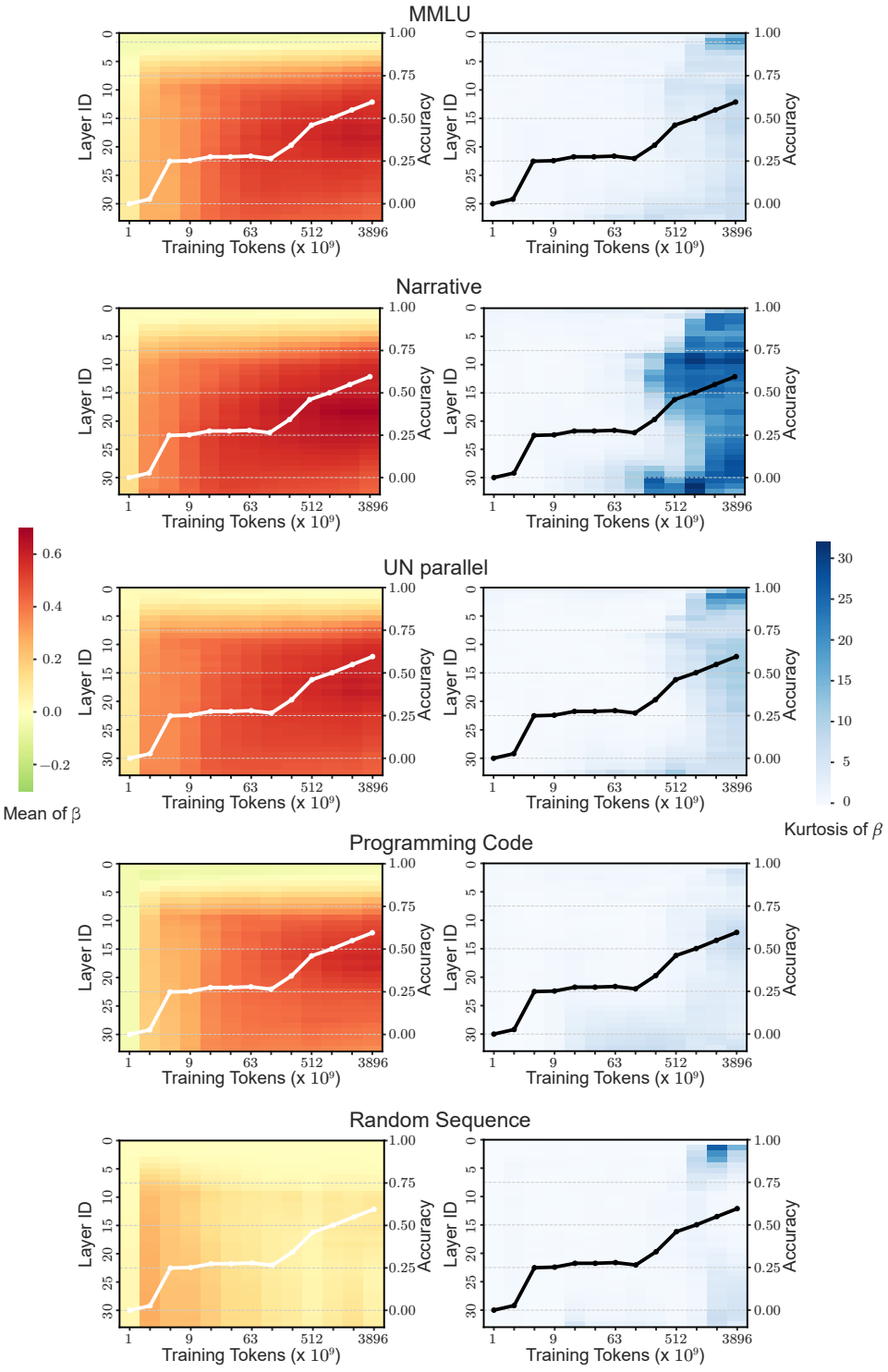

Figure E.17: Layer-wise and checkpoint-wise metrics of intra-head β distributions across corpora.

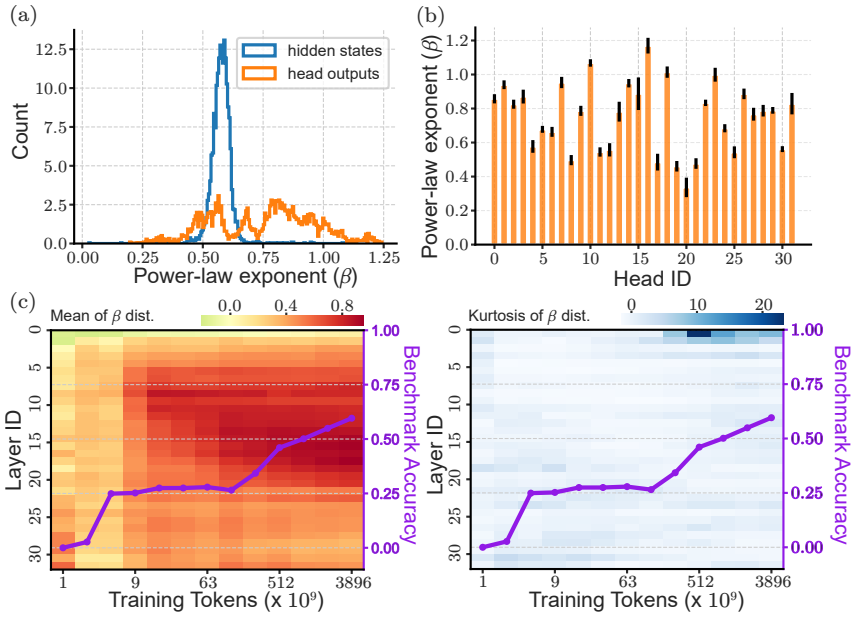

Figure E.18: (a) Distribution of power-law exponents β obtained via PSD from hidden states and attention head outputs at Layer 20 of the final model (Phase C). Attention head outputs (orange histogram) exhibit a flat, wide, and multimodal distribution. (b) Mean and standard deviation of β for individual attention heads derived from (a). The small intra-head standard deviations indicate highly consistent β values within each head. (c) Layer-wise and checkpoint-wise mean and kurtosis of the intra-head mean β distributions. Key trends include high mean β in deep layers during early training, a departure above $\beta = 0.5$ in Layers 10–20 from Phase B to C, and a marked increase in kurtosis at Layers 1–2 in Phase C.

E.5 Spectral Dynamics Across Architectural Loci

To identify the architectural source of the power-law dynamics observed in hidden states, we analyzed the activations within individual attention heads. Specifically, we extracted the attention head outputs, defined as the weighted sum of Value vectors (AV) prior to the output projection matrix W_O, under the same experimental conditions.

As shown in Figure E.18(a), the distribution of β across attention head outputs is flatter and more multimodal than that of global hidden states. When grouping hidden dimensions by their corresponding heads (128 dimensions per head), the intra-head standard deviation (σ_{intra}) of β was markedly smaller than the inter-head variation (σ_{inter}), indicating high dynamic consistency within each head (Figure E.18(b)). Tracking these head-level statistics across training checkpoints revealed three distinct temporal shifts (Figure E.18(c)): (1) the average β across attention heads was high ($\beta \in [0.5, 1.0]$) in deep layers during Phase A, but converged toward ≈ 0.5 upon transitioning to Phase B; (2) in middle layers (Layers 10–20), β gradually increased from 0.5 to ≈ 1.0 during the second phase shift (Phase B to C); and (3) a selective increase in kurtosis emerged specifically within shallow layers (Layers 1–2) during Phase C. PSD evaluations for other architectural components (e.g., MLP sub-layers) are detailed in Appendix E.5.

To further trace the structural origins and propagation of power-law dynamic characteristics within the transformer architecture, we extended our power-spectral density (PSD) analysis to several intermediate components beyond the residual hidden states and individual attention head outputs reported in the main text.

All evaluations were conducted under identical experimental settings and measurement protocols as described in Section 3. Specifically, we probed the following three intermediate loci within each layer:

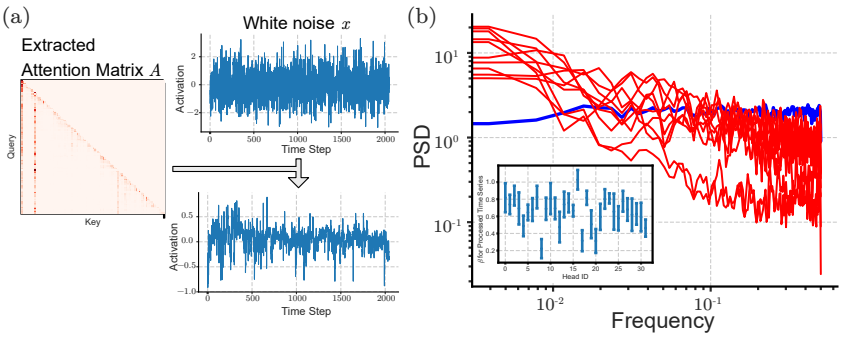

Figure E.19: Attention matrices A were extracted individually from all 32 attention heads in Layer 20 under the same processing conditions as the previous experiments, and subsequently applied to temporally uncorrelated white noise vectors x. By measuring the power-law exponent β of the resulting output time series Ax, we evaluated the intrinsic ability of the attention mechanism to generate temporal dynamics. The results demonstrate that even when provided with completely uncorrelated inputs, the attention matrices in Layer 20 systematically induce long-range temporal correlations with $\beta \approx 0.5$ in the output Ax.

1. **Value Representations (V):** The unprojected Value vectors within each attention head, evaluated prior to weighting by the attention matrix A and application of the output projection matrix W_O.
2. **Attention Block Outputs:** The aggregated output of the multi-head attention mechanism after applying the projection W_O and subsequent RMSNorm, corresponding to the primary residual contribution added to the hidden states.
3. **MLP Block Outputs:** The output of the Feed-Forward Network (MLP) block following RMSNorm, representing the secondary residual contribution.

Formally, for a hidden state x_l at layer l, the layer update is defined as:

$$x_{l+1} = x_l + \text{RMSNorm}(\text{Attn}(x_l)) + \text{RMSNorm}(\text{MLP}(x'_l)) \tag{4}$$

Our analysis isolates the spectral properties of each constituent term ($\text{RMSNorm}(\text{Attn}(\cdot))$ and $\text{RMSNorm}(\text{MLP}(\cdot))$) alongside the unprojected Value states V.

E.5.1 Key Empirical Observations

Comparative analysis across these architectural loci revealed several distinct properties regarding how power-law dynamics are generated and propagated through the network:

- **Locus-Specific Multimodality:** Multi-peaked (multimodal) distributions of β were uniquely observed in individual attention head outputs. In contrast, all other intermediate components (including unprojected V, Attention block outputs, and MLP block outputs) exhibited unimodal distributions.
- **Role of Attention Aggregation (V vs. Head Output):** Across all layers, head outputs consistently exhibited higher β values than their corresponding V representations ($\beta_V < \beta_{\text{head_out}}$). This indicates that the attention weighting and projection operations actively amplify long-range temporal correlations rather than merely reflecting the spectral properties of raw inputs or features.
- **Spectral Hierarchy Across Sub-Layers:** Comparing the residual streams, the average β value followed a strict hierarchical order:

  $$\beta_{\text{Attn_out}} > \beta_{\text{MLP_out}} > \beta_{\text{Hidden_State}} \tag{5}$$

  This ordering suggests that the attention mechanism serves as the primary engine for generating long-range temporal correlations, whereas the MLP blocks play a relatively diminished role in driving power-law dynamics, potentially acting to smooth or regulate the spectral characteristics before integration into the main residual stream.

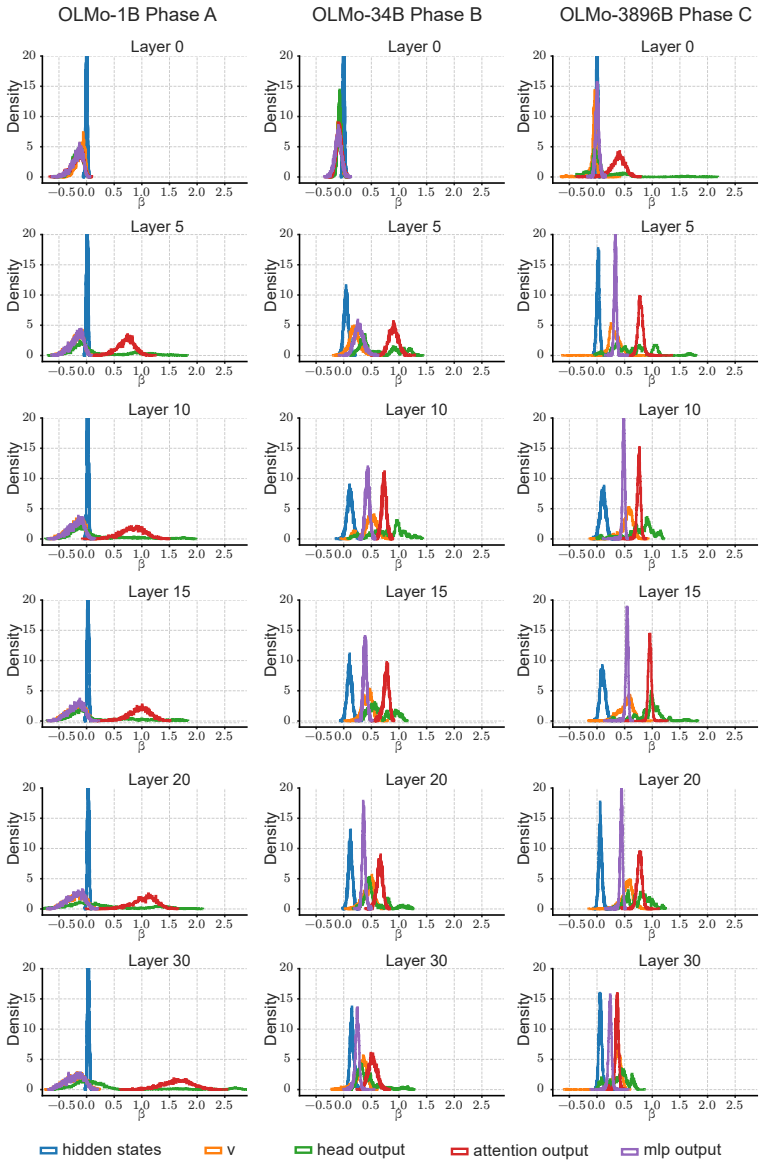

Figure E.20: β distribution metrics evaluated across training checkpoints for hidden states, unprojected value representations, head outputs, attention block outputs, and MLP block outputs.

F SELECTIVE RESPONSE ANALYSIS

To investigate whether specific internal units exhibit domain- or token-specific selectivity, we quantified the responsiveness of individual units during free-form text generation. This analysis utilizes the generated output tokens and their corresponding internal activations across time steps.

F.1 CORPUS AND TARGET TOKEN SELECTION

We collected a dataset consisting of 50 independent generation trials, each with a sequence length of 4,096 tokens (yielding a total corpus of $50 \times 4,096 = 204,800$ tokens). To ensure statistical reliability in measuring selectivity, we restricted our target set to tokens that appeared at least 50 times within the corpus.

F.2 THRESHOLDING AND RESPONSE RATIO CALCULATION

For each unit u and token t, we evaluated the temporal activation dynamics to determine token-level responsiveness:

1. Across all time steps in a given trial, a unit u is defined to have "responded" to token t if its activation $a_u(t)$ exceeds a standard deviation threshold of 4σ:

$$a_u(t) > \mu_u + 4\sigma_u \tag{6}$$

   where μ_u and σ_u represent the mean and standard deviation of unit u's activations across the entire trial, respectively.

2. We aggregated responses across all 50 trials. The degree of selective responsiveness for unit u toward token t is quantified by the *response ratio* $R(u,t)$, defined as:

$$R(u,t) = \frac{N_{\text{response}}(u,t)}{N_{\text{total}}(t)} \tag{7}$$

   where $N_{\text{total}}(t)$ is the total number of occurrences of token t in the corpus, and $N_{\text{response}}(u,t)$ is the number of times unit u met the 4σ activation condition upon receiving token t.

In the following section, we present representative examples of flatter outlier units that exhibited remarkably high response ratios toward specific semantics or functional token classes.

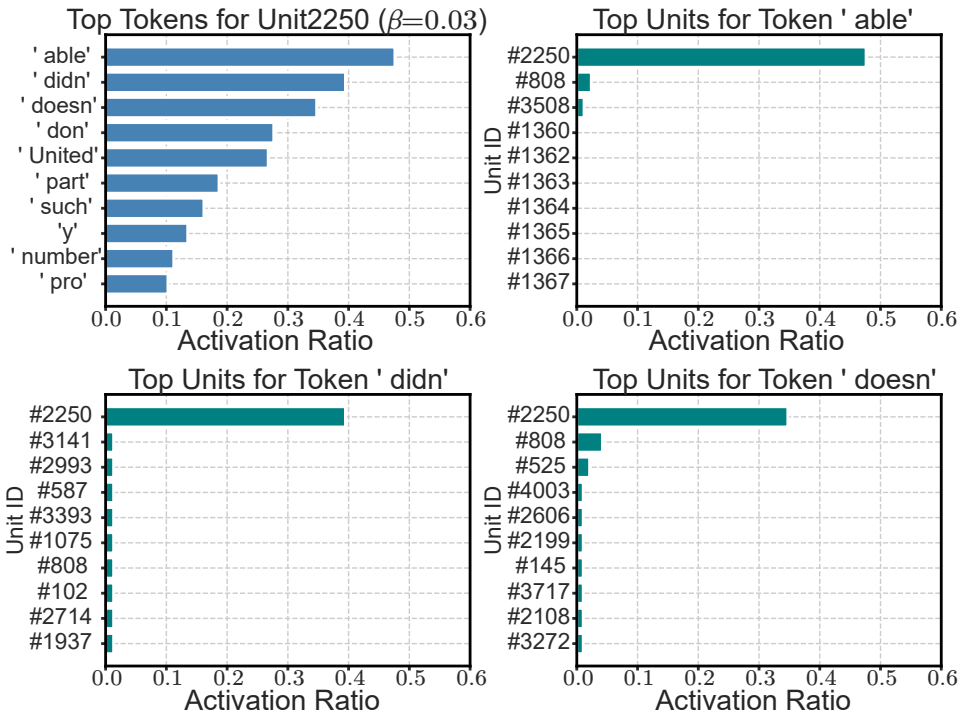

Figure F.21: Response ratio for each unit-token pair is defined as the fraction of token occurrences that triggered a threshold-crossing activation. Unit 2250 shows a high response ratio toward the tokens ”able”, ”didn”, and ”part” (blue bars). Conversely, among all units responding to these tokens, Unit 2250 exhibits by far the highest response ratio (green bars).

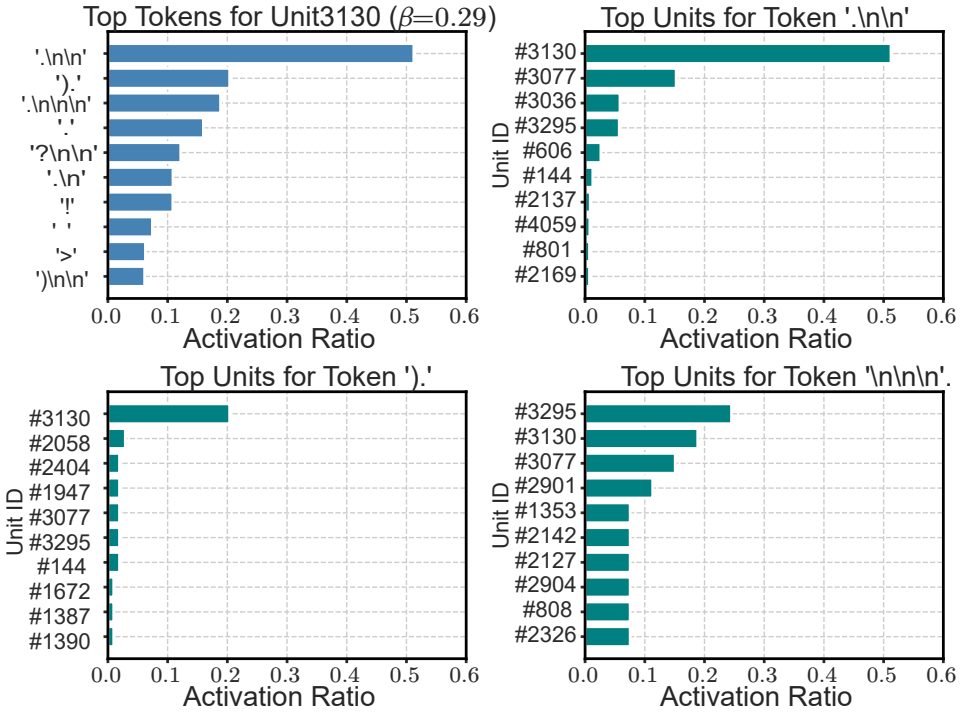

Figure F.22: Response ratio for each unit-token pair is defined as the fraction of token occurrences that triggered a threshold-crossing activation. Unit 3130 shows a high response ratio toward the punctuation tokens. Conversely, among all units responding to these tokens, Unit 3130 exhibits by far the highest response ratio (green bars).

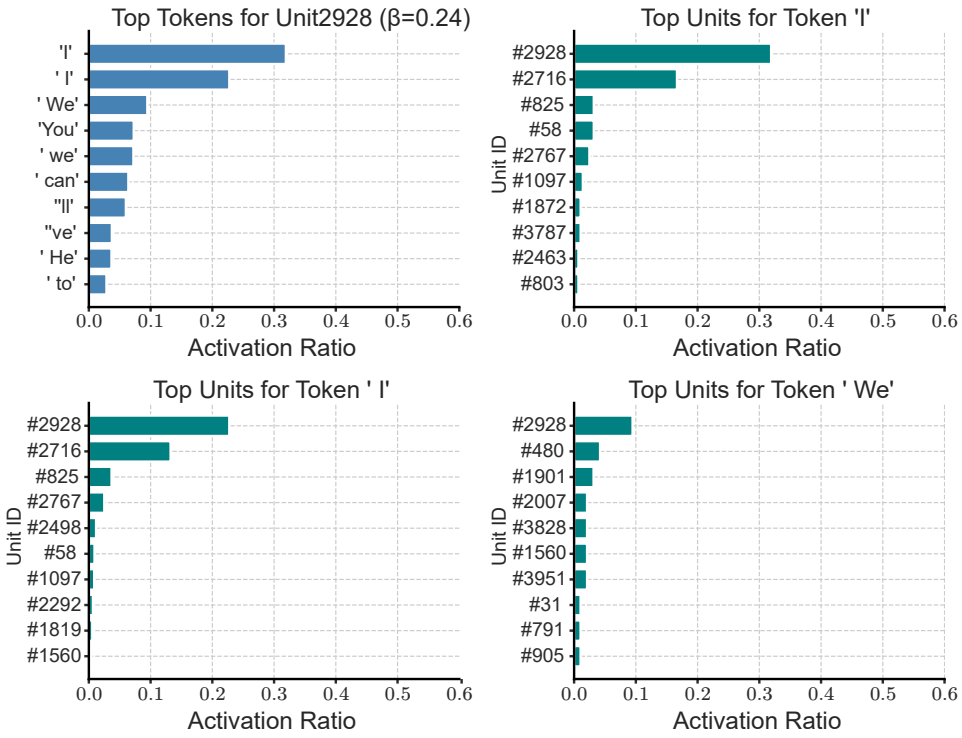

Figure F.23: Response ratio for each unit-token pair is defined as the fraction of token occurrences that triggered a threshold-crossing activation. Unit 2928 shows a high response ratio toward the personal pronoun tokens. Conversely, among all units responding to these tokens, Unit 2928 exhibits by far the highest response ratio (green bars).

G Long-range Correlation of Output Token Sequences

To speculate how the acquisition of internal long-range correlations manifests in the model's generated surface-level text, we analyzed the temporal fluctuation properties of the output token sequences.

G.1 Random Walk Fluctuation Analysis

Following the methodology of (Altmann et al., 2012), we quantify long-range correlations by mapping output sequences into binary indicator time series and measuring their windowed count variance.

Specifically, for a given target token α in a generated output sequence of length N, we define a binary sequence $x = (x_1, x_2, \ldots, x_N)$ as:

$$x_i = \begin{cases} 1, & \text{if token}_i = \alpha \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

To evaluate the scaling behavior over temporal window sizes t, we consider the cumulative occurrence $X_j(t)$ over a window of length t starting at index j:

$$X_j(t) = \sum_{k=j}^{j+t-1} x_k \tag{9}$$

We then calculate the variance $\sigma_X^2(t)$ across all valid starting positions $j \in \{1, \ldots, N - t + 1\}$:

$$\sigma_X^2(t) = \text{Var}_j[X_j(t)] = \langle X_j(t)^2 \rangle - \langle X_j(t) \rangle^2 \tag{10}$$

Following standard fluctuation analysis, t is sampled at 30 logarithmically spaced intervals ranging from $t = 1$ to a maximum window of $t_{\text{max}} = \lfloor N/5 \rfloor$ to ensure sufficient sample sizes for variance estimation.

G.2 Empirical Results and Discussion

Across the analyzed target tokens, we observed a systematic evolution in the scaling slope of $\sigma_X^2(t)$ as training progressed from Phase A to Phase B.

With the exception of `"."` and `"is"`, all evaluated tokens—including structural markers (`","`, `" "`) and core function words (`"the"`, `"in"`, `"to"`, `"of"`)—exhibited a distinct steepening of the scaling slope in Phase B relative to Phase A. This change in slope indicates an increase in the fluctuation exponent, directly signifying a transition toward stronger, longer-range temporal dependencies within the generated token streams.

Crucially, this output-level shift occurs in close synchronization with the internal phase shift identified in Section 4.2, where internal hidden dynamics transition to power-law regimes ($\beta \approx 0.5$). These results speculate a direct link between representation dynamics and surface-level output: as the internal state space establishes long-range temporal correlations, the model simultaneously acquires the capacity to structure long-range statistical dependencies in its generated language.

G.3 Target Token Selection and Corpus Setup

We applied this fluctuation measurement to the corpus generated during the free-generation experiments described in Section 3.

To guarantee statistical reliability when comparing across different training checkpoints, we restricted our target vocabulary to tokens that occurred at least $1,000$ times across all evaluated checkpoints. The target set spans high-frequency structural elements, including punctuation marks (e.g., `","`, `"."`) and whitespace (`" "`), as well as common function words (e.g., `"the"`, `"in"`, `"to"`, `"of"`, `"is"`).

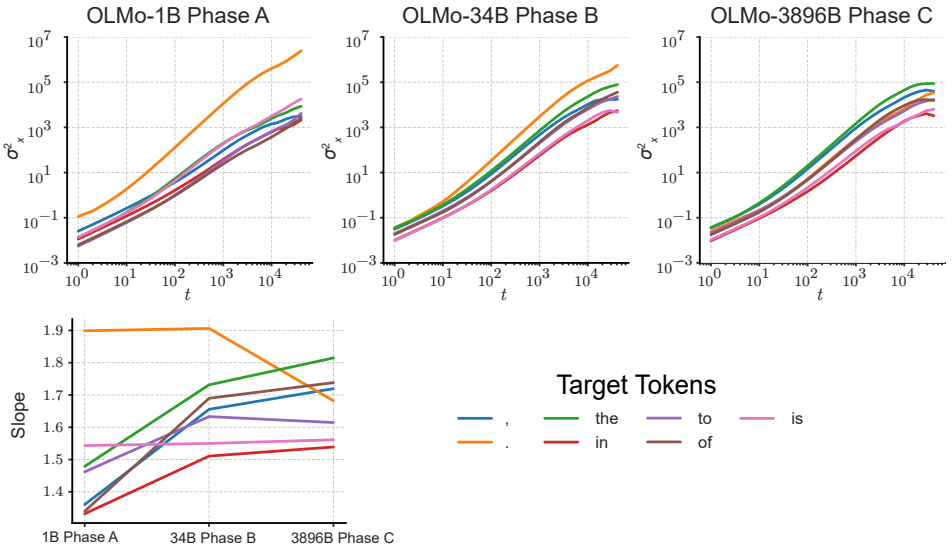

Figure G.24: Measured fluctuation functions $\sigma^2_X(t)$ and the evolution of their corresponding scaling slopes across Phase A, B, and C checkpoints of OLMo. We observe the increase in slope from Phase A to Phase B.

H Filter Analysis

To investigate the intrinsic spectral filtering properties of the attention mechanism, we evaluated how attention matrices transform input signals with varying temporal correlation properties into output time series.

H.1 Experimental Setup and Noise Filtering

We extracted attention matrices $A \in \mathbb{R}^{T\times T}$ from all 32 attention heads in Layer 20 across 10 independent text generation trials for each model checkpoint, using the same free-generation setup described in Section 3.

To test the frequency-dependent transmission characteristics of these matrices, we generated synthetic colored noise vectors $x \in \mathbb{R}^T$ with controlled power-law spectral exponents $\beta_{\text{in}} \in \{-0.5, -0.25, 0.0, 0.25, \ldots, 2.0\}$ (sampled in steps of 0.25). For each extracted attention matrix A and each target β_{in}, the attention transformation was computed via matrix multiplication:

$$y = Ax \tag{11}$$

where $y \in \mathbb{R}^T$ represents the filtered output sequence.

To ensure statistical stability, the filtering process $y = Ax$ was repeated over 5 independent random noise realizations for every combination of attention matrix and β_{in}. We subsequently computed the Power Spectral Density (PSD) of the resulting output sequence y and estimated its output spectral exponent β_{out} via linear regression in log-log space.

By systematically mapping $\beta_{\text{in}} \to \beta_{\text{out}}$, this setup allows us to isolate whether the attention mechanism acts as a passive linear filter, an amplifier of long-range correlations, or an intrinsic generator of power-law dynamics ($\beta \approx 0.5$) independent of input correlation.

H.2 Filtering Characteristics and Functional Differentiation of Attention Heads

To isolate the structural contribution of the attention mechanism from language-specific data statistics, we evaluated whether attention matrices intrinsically act as temporal filters. Specifically, we applied attention matrices A saved during inference to non-linguistic, uncorrelated white noise vectors

x. Remarkably, a single attention operation Ax transforms uncorrelated noise into long-range correlated dynamics, demonstrating that attention maps independently possess the inherent capacity to generate power-law behaviors. Crucially, in Appendix H.1, the resulting exponent β is head-specific, indicating that individual attention heads operate as distinct, functional dynamic filters. Furthermore, this transformation is not a trivial low-pass filtering effect, but a distinctive spectral shaping characteristic that systematically modulates dynamic properties across various colored noise inputs. Together, these findings establish self-attention as an intrinsic temporal filter that actively shapes the time-scale dynamics within LLMs.

Figure H.25 illustrates the input-output spectral transfer dynamics ($\beta_{\text{in}} \rightarrow \beta_{\text{out}}$) across multiple layers of the fully trained OLMo model. In each panel, individual curves represent the response profile of distinct attention heads within the corresponding layer.

Our spectral analysis reveals two functionally distinct categories of attention heads:

1. **Input-Invariant Generators:** A subset of attention heads maintains a nearly constant output exponent (β_{out}) regardless of the input noise exponent β_{in}. These heads act as intrinsic power-law generators, steadily imposing long-range temporal correlations onto incoming representations.
2. **Selective Correlation Amplifiers:** Another distinct subset exhibits a smooth, non-monotonic response curve with a clear peak at intermediate β_{in} values. Rather than uniformly filtering all frequencies, these heads act as band-pass temporal filters that selectively amplify input signals possessing specific moderate levels of long-range correlation.

Control Experiment with Random Attention. To confirm that these spectral modulation properties arise from learned attention parameters rather than architectural artifacts (such as Softmax normalization or Rotary Position Embeddings; RoPE), we conducted a control experiment. We generated synthetic attention matrices using randomly initialized query/key projection matrices (W_Q, W_K) and random hidden vectors h, applying identical RoPE positional encodings. Under this null model, the resulting output exponent remained centered near zero ($\beta_{\text{out}} \approx 0$) across all β_{in}. This explicitly demonstrates that the ability to generate and selectively amplify long-range temporal correlations is a non-trivial property acquired through pretraining.

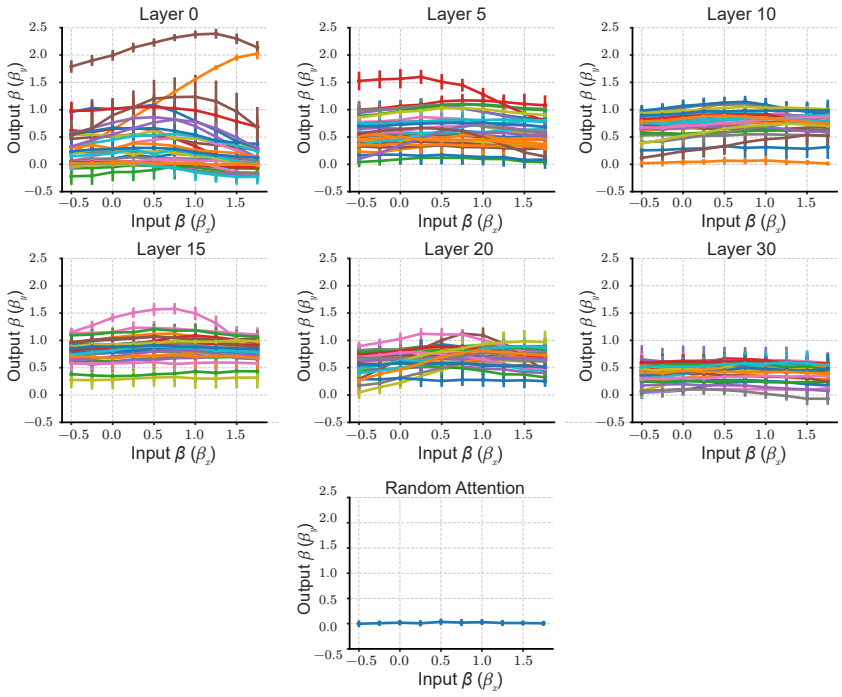

Figure H.25: Response curves of output power-law exponent β_{out} against input colored noise exponent $\beta_{\text{in}} \in [-0.5, 2.0]$ across attention heads in multiple layers. Each curve represents an individual attention head. (a) Attention heads exhibit functional specialization into *input-invariant generators* (flat profiles maintaining constant β_{out}) and *selective amplifiers* (peaked profiles that maximally accentuate intermediate input correlation scales). (b) Control experiment using synthetic attention matrices constructed from random weights and hidden states with RoPE; β_{out} consistently remains near zero ($\beta_{\text{out}} \approx 0$), confirming that the observed power-law transmission is a learned dynamical behavior.