

A Graph Signal Processing Perspective on Numerical Sequence Representations in LLM In-Context Learning

Jiajun Bao^{1*}, Zihao Qi¹, Toni J.B. Liu¹, Gurbir Arora¹
 Raphaël Sarfati^{1,2}, Nicolas Boullé³, Christopher J. Earls¹
¹Cornell University ²Goodfire AI ³Imperial College London

Abstract

Pretrained large language models (LLMs) have demonstrated in-context learning (ICL) capabilities for numerical inference over sequences serialized as text. Prior work has identified and characterized this form of numerical inference primarily through output-level evaluations such as prediction error. However, how numerical information is organized within LLM representations remains much less understood. To study this internal organization, we adopt a graph signal processing perspective in which attention induces a weighted graph over tokens, while token hidden states define signals on its nodes. Quantitative graph-spectral diagnostics and qualitative token-graph visualizations reveal that representations become more clearly differentiated by input dynamical complexity as context length increases. Simpler inputs produce attention-induced token graphs with stronger global connectivity and smoother, spectrally concentrated hidden-state signals, whereas more complex inputs produce more localized graphs and hidden-state signals with broader spectral support and greater high-frequency energy. Together, these findings point to systematic, context-dependent internal signatures associated with numerical ICL that are conserved across model families.

1 Introduction

Large language models (LLMs) are increasingly studied for numerical inference over sequences serialized as text. Without task-specific parameter updates, they can infer and extrapolate numerical processes, a capability we refer to as numerical in-context learning (ICL). Prior work has demonstrated this capability in time-series forecasting (Gruver et al., 2023; Jin et al., 2024), numerical regression (Vacareanu et al., 2024; Requeima et al., 2024), dynamical-system inference

(Liu et al., 2024; Zhang and Gilpin, 2025), density estimation (Liu et al., 2025), Markov-model prediction (Dai et al., 2025; Zekri et al., 2024), and partial differential equation extrapolation (Bao et al., 2026). Depending on the task, longer contexts often yield more accurate point predictions, predictive distributions closer to the ground-truth distribution, or more faithful recovery of latent dynamics. Yet most existing studies characterize these improvements through model outputs, treating the model as a black box and leaving it unclear how its internal representations change as context length increases.

Recent studies have begun to examine internal representations associated with numerical ICL, providing evidence that LLM hidden states systematically encode numerical information. In particular, Sarfati et al. (2026) identify curved manifolds encoding beliefs about Gaussian parameters, while Piskorz et al. (2026) use trained probes to decode point predictions and uncertainty from LLM hidden states. Motivated by these findings, we jointly analyze token-level hidden representations and the structure of attention-induced token graphs. Specifically, we ask how their organization varies with context length and input dynamical complexity.

To study this interplay, we adopt the graph signal processing perspective summarized in Fig. 1: at each transformer layer, attention induces a weighted graph over tokens, while the corresponding token hidden states define signals on its nodes. Our framework supports two complementary analyses: 1) graph-spectral diagnostics, which quantify global attention-graph connectivity, high-frequency energy in hidden-state signals, and the breadth of their spectral support; 2) attention-induced token-graph visualizations, which provide a qualitative view of token-level organization consistent with the quantitative trends. Across both analyses, we find that, as context length increases, the model’s inter-

*Correspondence author: jb2777@cornell.edu

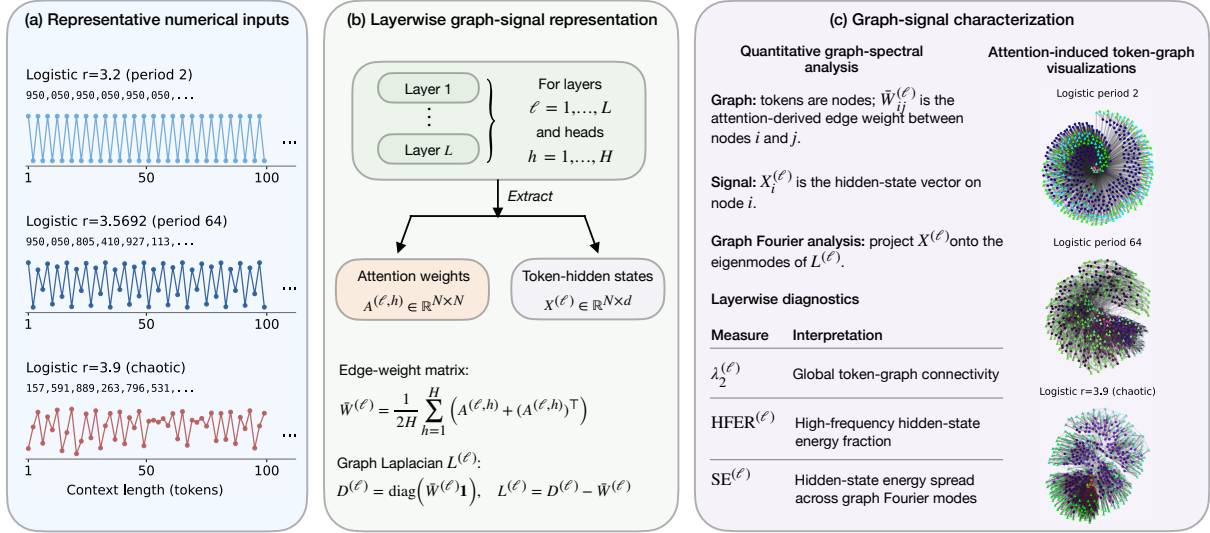


Figure 1: Overview of the graph signal processing framework for studying LLM internal representations during numerical in-context learning. **(a)** Numerical trajectories are quantized and serialized as comma-delimited sequences of three-digit numbers. **(b)** At each layer ℓ , head-aggregated attention defines a weighted, undirected token graph $\bar{W}^{(\ell)}$ over the N token nodes, hidden states $X^{(\ell)}$ define d -dimensional node signals, and Laplacian eigenvectors form the graph Fourier basis, enabling spectral analysis of these signals. **(c)** Layerwise graph-spectral diagnostics and attention-induced token-graph visualizations characterize these representations. Node positions reflect attention topology, node colors encode hidden-state signals, and circles and triangles denote numeric and separator tokens, respectively. The examples are from the final layer of Llama-3.2-3B at $N = 500$.

nal representations become more clearly ordered by input dynamical complexity. The diagnostics further distinguish input families whose output-level extrapolation errors remain tightly clustered, suggesting that representation-level analysis can reveal structure not apparent from extrapolation error alone (cf. Fig. 5 in Appendix A).

2 Background and Related Work

Graph spectra and graph signals. Spectral graph theory studies graph structure through operators such as the Laplacian, whose spectrum captures connectivity, cuts, and global organization (Fiedler, 1973; Chung, 1997). Graph signal processing extends this spectral viewpoint to data defined on graph nodes: for undirected graphs, Laplacian eigenvectors form a graph Fourier basis whose modes are ordered by how rapidly nodewise signals vary over the graph (Hammond et al., 2011; Shuman et al., 2013; Ortega et al., 2018). This classical framework relates the variation of graph-supported signals to the geometry of the underlying graph.

Graph-spectral analysis of LLMs. Recent graph-spectral work on LLMs has focused on hallucination detection and reasoning verification. Binkowski et al. (2025) use attention-derived Laplacian spectra for hallucination detection. Closer

to our setting, Noël (2025, 2026) treats attention-induced token graphs together with token-level hidden-state signals, using diagnostics such as spectral entropy, high-frequency energy ratio, and the Fiedler value for hallucination analysis and proof verification. At the output level, Balaji (2026) constructs sentence-level graph signals from generated chain-of-thought traces, enabling verification without access to internal attention weights or hidden states. We adopt closely related diagnostics for quantitative, representation-level analysis of numerical ICL rather than for hallucination detection or reasoning verification.

Applications of graph constructions in LLMs.

Several recent works represent LLM reasoning traces or token-level attention patterns as graphs, including sentence-level chain-of-thought dependency graphs (Bogdan et al., 2025), attention-induced token graphs for reinforcement-learning credit assignment (Dong et al., 2026), and attributed attention graphs used with graph neural networks for hallucination detection (Frasca et al., 2026). In contrast, we use training-free, layerwise visualizations of attention-induced token graphs to display attention topology alongside low-dimensional hidden-state structure.

Collectively, these studies leave open how the organization of LLM internal representations during numerical ICL varies with context length and input dynamical complexity. We investigate this question through complementary quantitative and qualitative analyses within a graph signal processing framework.

3 Methods

3.1 Numerical Inputs and Tokenization

Controlled numerical input suite. Our main experiments use ten one-dimensional sequence families spanning different levels of dynamical complexity under a fixed prompt format: a sampled constant baseline, logistic map trajectories (May, 1976), and trajectories from the x -component of the Lorenz system (Lorenz, 1963). The logistic map, $x_{t+1} = rx_t(1 - x_t)$, where $x_t \in [0, 1]$ and $r \in [0, 4]$ is a control parameter, serves as our primary controlled family. Our selected values of r span periodic regimes with periods 2, 4, 8, 16, 32, and 64, followed by two chaotic regimes along the classical period-doubling route (Strogatz, 2024). The logistic map therefore provides a controlled setting for isolating how input dynamics affect LLM representations: varying r produces trajectories with progressively longer periods and then chaotic dynamics while preserving the recurrence form.

Throughout, we use *dynamical complexity* operationally, with increasing complexity corresponding to the progression from constant inputs through periodic inputs of increasing period to chaotic inputs. This ordering is also supported quantitatively by a standard entropy-based characterization of the input dynamics, as detailed in Appendix B. The constant sequence provides a low-complexity baseline, whereas the Lorenz trajectory provides a complementary observation from a continuous-time chaotic system. We evaluate an ensemble of 20 trajectory realizations per family. Appendix A provides full input-generation details. To broaden the input coverage beyond the controlled dynamical-system suite, we additionally consider structured numerical sequences with distributional shifts, smooth periodic variation, and stochastic evolution, as described in Appendix I.

Quantization and serialization. Following prior work (e.g., Gruver et al., 2023; Liu et al., 2024), we encode each numerical input sequence as a quantized, comma-delimited text prompt. For a trajectory $(x_1, \dots, x_T)$, scalar values are linearly

rescaled to the interval $[50, 950]$ and rounded to obtain quantized integers $q_i \in \{50, \dots, 950\}$. Each q_i is formatted as a zero-padded three-digit decimal string. The encoded values are separated by commas, yielding strings such as “087,173,642,...”. To ensure comparability across context lengths, we quantize the full trajectory before truncation so that the numerical mapping remains fixed. We choose this format because widely used tokenizers, such as those used by GPT-4 (OpenAI, 2023) and Llama-3-family models (Grattafiori et al., 2024), encode three-digit integer strings (000–999) and commas as single tokens. Under this serialization, T numerical states yield T numeric and T separator tokens, so the analyzed context length is $N = 2T$. This avoids splitting numerical states across tokens and preserves one-to-one state-to-token alignment for graph analysis. Further quantization and serialization details are provided in Appendix A.

3.2 Graph-Spectral Diagnostics

We adapt three established spectral diagnostics employed by Noël (2026) to numerical ICL: the normalized Fiedler value, high-frequency energy ratio, and spectral entropy (all defined below). In our setting, these diagnostics characterize how global attention-graph connectivity, hidden-state roughness, and spectral diversity vary with input dynamical complexity and context length.

Layerwise representations. For an N -token input, we record two objects at each layer. First, for each layer $\ell \in \{1, \dots, L\}$ and attention head $h \in \{1, \dots, H\}$, let $A^{(\ell,h)} \in \mathbb{R}^{N \times N}$ denote the post-softmax attention matrix, where $A_{ij}^{(\ell,h)}$ is the attention weight from query token i to key token j .¹ Second, let $X^{(\ell)} \in \mathbb{R}^{N \times d}$ denote the hidden-state matrix entering layer ℓ , whose i -th row $X_i^{(\ell)}$ is the d -dimensional representation of analyzed token i . We pair each layer’s attention matrices with its input hidden states because those states generate the queries and keys underlying the attention weights.

Graph construction and Fourier modes. To support standard Laplacian-based graph-spectral analysis, we construct one undirected weighted graph per layer by representing each analyzed token as a node, symmetrizing each head’s attention matrix, and averaging the symmetrized matrices

¹When present, we exclude non-content special tokens introduced during model input preprocessing, such as <BOS>, and renormalize each attention row over the remaining tokens.

across heads (see Appendix Fig. 10 for representative final-layer attention matrices):

$$\bar{W}^{(\ell)} = \frac{1}{2H} \sum_{h=1}^H \left(A^{(\ell,h)} + (A^{(\ell,h)})^\top \right).$$

By construction, $\bar{W}_{ij}^{(\ell)}$ is the attention-derived edge weight between tokens i and j . We define the corresponding degree matrix as $D^{(\ell)} = \text{diag}(\bar{W}^{(\ell)} \mathbf{1})$ and the combinatorial graph Laplacian as $L^{(\ell)} = D^{(\ell)} - \bar{W}^{(\ell)}$. The resulting $L^{(\ell)}$ is symmetric positive semidefinite and therefore admits an orthonormal eigendecomposition $L^{(\ell)} = U^{(\ell)} \Lambda^{(\ell)} (U^{(\ell)})^\top$. The eigenvectors in $U^{(\ell)}$ define graph Fourier modes ordered from low to high graph frequency (Shuman et al., 2013). Treating the hidden-state matrix $X^{(\ell)}$ as a vector-valued graph signal whose value at node i is $X_i^{(\ell)}$, we project it onto the graph Fourier basis as $\hat{X}^{(\ell)} = (U^{(\ell)})^\top X^{(\ell)}$. Low-eigenvalue modes correspond to smoother hidden-state variation over the attention-induced graph, whereas high-eigenvalue modes correspond to sharper variation across token pairs joined by larger attention weights, as follows from the Laplacian quadratic-form identity in Appendix C.

Diagnostic 1: Normalized Fiedler value. The Fiedler value—the second-smallest eigenvalue of the combinatorial graph Laplacian $L^{(\ell)}$ —is a classical measure of global connectivity (Fiedler, 1973). To facilitate comparisons across graph sizes and degree scales, we use the standard normalized formulation (Chung, 1997):

$$\mathcal{L}^{(\ell)} = (D^{(\ell)})^{-1/2} L^{(\ell)} (D^{(\ell)})^{-1/2}.$$

The spectrum of $\mathcal{L}^{(\ell)}$ lies in $[0, 2]$. Its second-smallest eigenvalue, denoted by $\lambda_2^{(\ell)}$, is the normalized Fiedler value. Larger $\lambda_2^{(\ell)}$ indicates a more globally integrated attention graph, whereas smaller $\lambda_2^{(\ell)}$ indicates more localized graph structure. Appendix D.1 examines additional low-end eigenvalues to characterize multiway localized graph structure beyond $\lambda_2^{(\ell)}$ (Lee et al., 2014).

Diagnostic 2: High-frequency energy ratio (HFER). Let $\hat{\mathbf{x}}_m^{(\ell)} \in \mathbb{R}^d$ denote the graph Fourier coefficient vector for mode m , given by the m th row of $\hat{X}^{(\ell)}$, and $p_m^{(\ell)} = \|\hat{\mathbf{x}}_m^{(\ell)}\|_2^2 / \sum_{r=1}^N \|\hat{\mathbf{x}}_r^{(\ell)}\|_2^2$ the normalized modal energy. The HFER is the fraction of hidden-state energy assigned to high

graph-frequency modes,

$$\text{HFER}^{(\ell)} = \sum_{m=K+1}^N p_m^{(\ell)}.$$

We use $K = \lfloor N/2 \rfloor$ as the default cutoff, so HFER measures the fraction of energy in the high-frequency half of the spectrum. Appendix D.1 shows that the qualitative behavior remains consistent across alternative cutoffs. Larger HFER indicates that a greater fraction of hidden-state energy lies in modes associated with sharper variation over the attention-induced graph.

Diagnostic 3: Spectral entropy (SE). Whereas HFER measures the fraction of hidden-state energy in high-frequency modes, SE measures how broadly that energy is distributed across graph Fourier modes. Using the normalized modal energies $p_m^{(\ell)}$, we define

$$\text{SE}^{(\ell)} = - \sum_{m=1}^N p_m^{(\ell)} \log p_m^{(\ell)}.$$

We report $\exp(\text{SE}^{(\ell)})$ as the effective spectral support, following the standard effective-number interpretation of exponentiated Shannon entropy (Jost, 2006). If energy is distributed equally across exactly k modes, then $\exp(\text{SE}^{(\ell)}) = k$. More generally, it represents the number of equally weighted graph Fourier modes that would yield the observed spectral entropy. Larger values indicate broader spectral support, whereas smaller values indicate concentration in fewer modes.

Layer averaging. In the main text, we report diagnostics averaged across model layers. The corresponding layer-resolved results in Appendix D.2 broadly reproduce the same qualitative trends, supporting layer averaging as a concise and representative summary across model depth. We use λ_2 , HFER, and $\exp(\text{SE})$ to denote the resulting layer-averaged diagnostics.

3.3 Token-Graph Visualization Pipeline

We complement the graph-spectral diagnostics with layerwise two-dimensional graph visualizations in which node color summarizes token hidden-state representations and node position reflects attention topology. These visualizations provide a qualitative view of the organization characterized quantitatively by our spectral diagnostics. Figure 2 illustrates this joint encoding for three representative input families across context lengths.

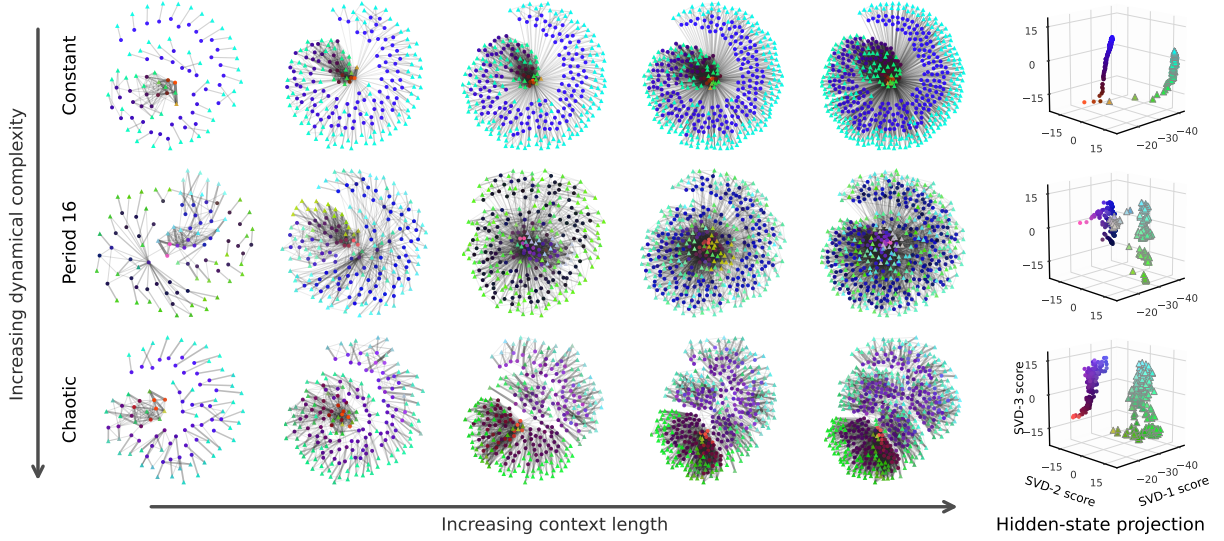


Figure 2: Illustration of attention-induced token graphs across context lengths. Rows show final-layer results from Llama-3.2-3B for representative constant, period-16 logistic map, and chaotic logistic map ($r = 4.0$) inputs. The first five columns show token graphs at $N \in \{100, 200, 300, 400, 500\}$. Node positions reflect input-specific attention weights through a Fruchterman–Reingold layout, with the same initialization used across inputs at each N . Edge thickness encodes attention-derived weight. Only edges above $2.5/N$ are displayed to reduce visual clutter. Node colors encode the leading three uncentered SVD coordinates of token hidden states, with the $N = 500$ projections shown in the rightmost column. Circles mark numeric tokens and triangles mark separator tokens.

Node-color construction for token graphs. In classical graph signal processing, a scalar signal is often visualized by mapping its value at each node to color or vertical height (Shuman et al., 2013). To obtain a color encoding of the vector-valued token hidden states, we project the hidden-state matrix at each layer onto its leading three uncentered SVD coordinates and assign them to the red, green, and blue (RGB) channels; Appendix E provides the full construction. At $N = 500$, the leading three components capture 75% of the hidden-state energy on average across the main numerical input families and analyzed layers (Appendix Fig. 12), supporting their use as a compact representation of the dominant hidden-state structure. Corresponding projections for all ten main numerical input families are provided in Appendix Fig. 11. For other inputs whose hidden-state energy is distributed across more SVD components, the visualization can be extended with additional color-coded views.

Attention-based node positions. Visualizing the attention-induced token graph as a two-dimensional node-link diagram requires choosing a layout. Layouts determined solely by the tokens’ sequential positions, such as the spiral initialization shown in Appendix Fig. 13, preserve input order. However, because attention weights do not influ-

ence node placement in these layouts, the resulting geometry may obscure nonlocal interactions and community structure reflected in the LLM’s internal attention organization. We therefore apply the Fruchterman–Reingold (FR) algorithm (Fruchterman and Reingold, 1991) to the full weighted graph defined by the symmetrized attention matrix $\bar{W}^{(\ell)}$. The algorithm iteratively updates the node positions by treating nodes as mutually repelling particles, which discourages node overlap, and edges as attractive forces weighted by the corresponding off-diagonal entries of $\bar{W}^{(\ell)}$, so larger attention-derived edge weights exert stronger attraction. At a fixed N , all inputs use the same spiral initialization and layout settings; differences in the relaxed node positions therefore reflect their input-specific attention weights rather than different starting coordinates. Starting from the token-ordered spiral, we run 100 iterations, allowing the layout to relax toward a stable configuration, and use the resulting positions for all subsequent token-graph visualizations. Appendix Fig. 13 shows that most visible layout reorganization occurs by 50 iterations, with only minor changes thereafter. Appendix E provides implementation details and demonstrates the robustness of the relaxed layouts to initialization perturbations.

Edge display threshold. The FR layout is computed using the full weighted attention graph, which contains $N(N - 1)/2$ off-diagonal edges. Rendering every connection would turn the visualization into a dense “hairball” in which meaningful patterns are no longer discernible. To reduce visual clutter, we display only edges satisfying $\bar{W}_{ij}^{(\ell)} > \kappa/N$, using $\kappa = 2.5$ for all main visualizations. Appendix Fig. 15 shows that the qualitative organization is robust to changes in the display threshold. This threshold affects only edge rendering; the FR node positions and all graph-spectral diagnostics are computed from the full weighted graph.

4 Results

We examine how the model’s internal representations vary with the dynamical complexity of the input across increasing context lengths. Section 4.1 quantifies this representational variation using graph-spectral diagnostics, and Section 4.2 visualizes the associated changes in the organization of attention-induced token graphs.

Choice of model. Our main analysis focuses on the pretrained base Llama-3.2-3B model. Appendix G extends this analysis to Llama-3.2-1B and Llama-3.1-8B, compares base and instruction-tuned variants, and broadens the evaluation to Microsoft Phi-4 (Abdin et al., 2024) and Hugging Face SmoLM3 (Bakouch et al., 2025). The broad context-length- and complexity-dependent trends persist across these models. Within the evaluated Llama 3 family, the 1B base model distinguishes the dynamical regimes less sharply than the 3B and 8B base models, and the instruction-tuned variants show weaker regime separation than their base counterparts. Within a shared architecture family, these comparisons suggest that model scale and instruction tuning both affect how distinctly dynamical regimes are expressed in the model’s internal representations.

4.1 Graph-Spectral Signatures of Dynamical Complexity Across Context Lengths

For our quantitative graph-spectral analysis, we compute the diagnostics defined in Section 3.2 for 20 realizations of each numerical input family introduced in Section 3.1 at 20 context lengths ranging from 100 to 2000 tokens in increments of 100.

Global attention connectivity reveals context-dependent regime separation. Fig. 3(a) shows that, with sufficient context, the normalized Fiedler value separates the inputs into two broad groups: the chaotic logistic map and Lorenz inputs exhibit lower λ_2 , whereas the constant and periodic inputs exhibit higher λ_2 . Because lower λ_2 indicates weaker global connectivity, the chaotic inputs induce more localized attention geometry, while the non-chaotic inputs retain more globally integrated graphs. Consistent with this interpretation, Appendix Fig. 7 shows that the chaotic inputs remain systematically below the non-chaotic inputs across additional low-end eigenvalues, indicating multi-way localized graph structure beyond λ_2 . Among the periodic inputs, those with higher periods exhibit a finer dependence on context length. Periods 32 and 64 initially follow the decline of the chaotic inputs before turning upward at $N = 300$ and $N = 400$, respectively, and approaching the lower-period curves. These ordered turning points suggest that the model needs progressively more context to resolve increasingly complex input dynamics in its internal attention organization.

Graph-signal smoothing varies with dynamical complexity. Fig. 3(b) shows an increasingly clear complexity-ordered HFER ladder as context length increases. HFER decreases most strongly for the constant and lower-period inputs, the decline weakens as period increases, and the chaotic inputs remain comparatively high after an initial decline. Together, these trends indicate that the effect of additional context on how smoothly hidden states vary over the attention-induced graph depends on input dynamical complexity: for simpler inputs, hidden-state energy becomes increasingly concentrated in low-frequency modes, whereas for more complex inputs, a larger fraction remains in high-frequency modes, preserving sharper variation across the graph.

Effective spectral support expands at complexity-dependent rates. Fig. 3(c) shows that the effective spectral support, $\exp(\text{SE})$, grows with context length for all inputs but remains well below its theoretical maximum N . Thus, increasing context does not simply cause the representation to spread uniformly across the larger set of available spectral modes; instead, it continues to favor a restricted spectral subset. Over the evaluated context range, effective support expands at a complexity-dependent rate: growth is

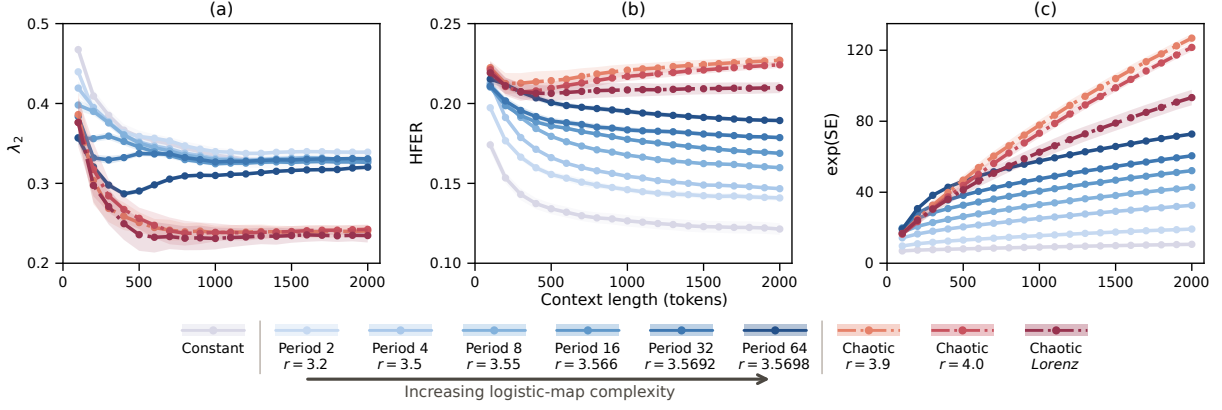


Figure 3: Layer-averaged graph-spectral diagnostics for the main numerical input families using Llama-3.2-3B. Panels report (a) normalized Fiedler value λ_2 , (b) HFER, and (c) effective spectral support $\exp(\text{SE})$. At longer contexts, chaotic inputs show weaker global attention connectivity (lower λ_2), greater high-frequency hidden-state energy (higher HFER), and broader effective spectral support (higher $\exp(\text{SE})$) than constant and periodic inputs; periodic families vary progressively with period. Curves and shading show the mean and ± 1 standard deviation across 20 realizations after averaging each diagnostic across transformer layers.

modest for the constant and lower-period inputs, whereas the higher-period and chaotic inputs show stronger expansion. These differences indicate that additional context affects the effective spectral dimensionality of internal representations differently across input families: hidden-state energy for simpler dynamics remains concentrated within a compact, slowly expanding spectral support, whereas for more complex dynamics it is distributed across an increasingly broad set of graph Fourier modes.

Taken together, the diagnostics reveal an ordered complexity gradient: greater input complexity corresponds to weaker global attention connectivity, higher HFER, and broader effective spectral support, with periodic inputs exhibiting graded changes as their period increases. Appendix H shows that the broad trends also appear in attention-only and hidden-state-only controls, while the graph-spectral diagnostics more clearly separate input families. Overall, these results show that the model’s internal representations reflect numerical input dynamics in a rich and systematically ordered manner as context length increases.

4.2 Attention-Induced Token-Graph Organization Across Dynamical Complexity and Context Length

Following the construction in Section 3.3, Fig. 4 presents representative final-layer attention-induced token graphs at $N = 500$. At this context length, the layer-averaged λ_2 trends in Fig. 3(a) already show broad separation between chaotic

and non-chaotic inputs in global attention-graph connectivity. Figure 2 further shows that the corresponding complexity-dependent visual organization is visible by $N = 300$ and remains qualitatively stable thereafter. We focus on the final layer because its attention-defined topology is closest to the representations used to compute the output logits. Appendix F provides complementary visualizations of additional transformer layers.

Token graphs become more localized with dynamical complexity. The layouts progress from comparatively regular, globally integrated organization for constant and lower-period inputs to less homogeneous organization as the period increases. The chaotic logistic map and Lorenz inputs contain multiple localized, clique-like subnetworks, consistent with their lower normalized Fiedler values. Beyond topology, the node colors reveal additional structure in the hidden-state signals: numeric and separator tokens generally occupy distinct regions of the representation. For the period-2 input, separator tokens split into two color groups aligned with the alternating numerical states, suggesting that their hidden states encode the trajectory’s alternating phase despite their shared token identity.

Long-period periodicity becomes more distinct from chaos as context increases. Period 64 provides an informative intermediate case: despite its periodic dynamics, its graph contains a prominent clique-like subnetwork resembling those of the chaotic inputs. The right panel maps this subnetwork and those in the chaotic logistic map and

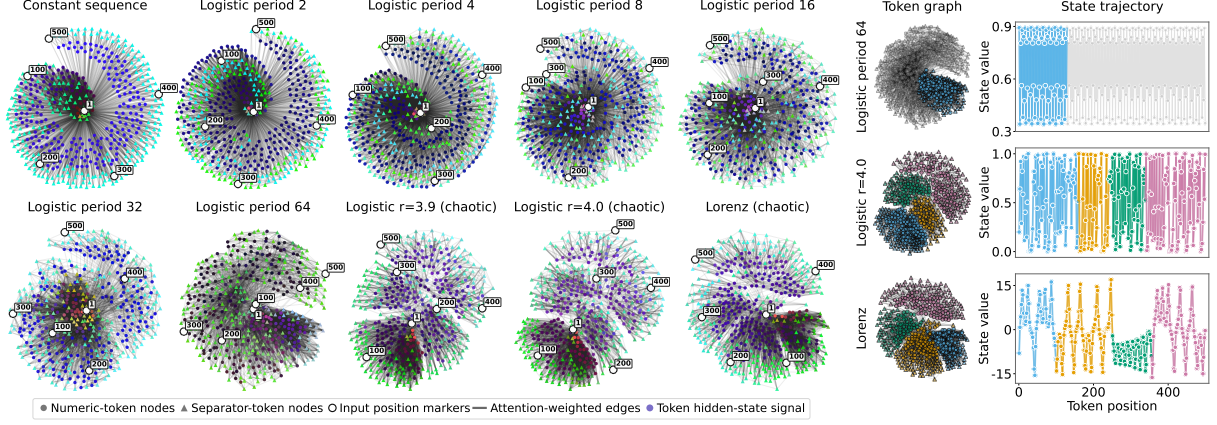


Figure 4: Final-layer attention-induced token graphs and trajectory mappings for Llama-3.2-3B at $N = 500$. **Left:** Node positions reflect the topology of the attention-induced token graph, node colors encode hidden-state graph signals, and circles and triangles denote numeric and separator tokens. Simpler inputs show globally integrated layouts, whereas chaotic inputs show localized, clique-like subnetworks. **Right:** Matched colors link these clusters to trajectory intervals for period 64, the chaotic logistic map, and Lorenz inputs. For period 64, a clique-like cluster appears early, before sufficient context reveals the recurrence, but no comparable clusters appear later; chaotic inputs exhibit multiple such clusters throughout.

Lorenz inputs to their corresponding trajectory intervals. For period 64, this subnetwork is concentrated within roughly the first 128 tokens, before the model has observed enough recurrence to distinguish its periodic structure from chaotic dynamics; later trajectory segments do not form similarly pronounced subnetworks. In contrast, the chaotic logistic map and Lorenz graphs contain multiple localized subnetworks throughout the context, each associated primarily with a contiguous trajectory interval. This difference suggests that additional context resolves the early ambiguity of the long-period input, whereas temporal partitioning persists for chaotic dynamics.

The visualizations complement the spectral diagnostics by showing a broad shift from globally integrated to more localized token-graph organization as dynamical complexity increases. Appendix I extends this qualitative analysis beyond the controlled dynamical-system suite to additional structured numerical inputs, including distributional shifts accompanied by distinct token-graph clustering.

5 Conclusion

We introduced a graph signal processing framework for numerical ICL, in which attention defines a token graph and hidden states form signals on its nodes. As context grows, our graph-spectral diagnostics and analyses of attention-induced token graphs reveal increasingly clear differences across trajectories of varying dynamical complex-

ity. These findings show that the models’ emergent capacity for numerical ICL manifests not only in the autoregressive continuation of numerical sequences at the output level but also in the fine-grained, context-dependent organization of their internal representations. More broadly, the framework provides a mesoscale characterization of LLM attention topology and hidden-state organization, complementing microscale neuron- and circuit-level analyses and macroscale output-based evaluations.

In numerical ICL, an important open question is whether the observed spectral signatures and attention-graph locality causally contribute to sequence continuation. Reconstructions from individual graph Fourier modes could localize hidden-state variation, while frequency-band ablations could test the functional importance of particular bands. Contrasting locality-preserving with locality-disrupting attention-edge interventions could test the role of graph locality.

Beyond numerical sequences, preliminary analyses in Appendix J show that, at matched context lengths, natural-language and code inputs consistently exhibit more localized attention topology, a greater high-frequency share of hidden-state energy, and broader effective spectral support than numerical inputs. Across natural languages, programming languages, and English-language domains, it remains to determine how input properties such as syntax, semantic content, formatting, and pre-

dictability, and model-dependent factors such as tokenization granularity and training-data exposure contribute to both the observed separation from numerical inputs and graph-spectral variation within these broader input classes.

Limitations

Mechanistic scope. Our graph-spectral diagnostics of attention-induced token graphs and hidden-state signals identify structural and representational signatures associated with numerical ICL, but do not establish the causal mechanisms that produce this behavior.

Model and tokenization choices. Our experiments cover several model families with Llama 3–style tokenization, for which each three-digit state corresponds to a single numeric node in the attention-induced token graph and each comma delimiter to a separator node. Other model families use different numerical tokenization schemes; for example, Gemma 4 tokenizes each digit separately (Gemma Team, 2026). Applying the same state-aligned construction to such models would require aggregating digit-level tokens into state-level nodes, which we leave to future work.

Access to internal representations. The proposed diagnostics and visualizations require access to layerwise attention weights and hidden-state activations. Hosted interfaces to frontier models, such as chat applications and APIs, typically do not expose these quantities. The framework is currently most directly applicable to open-weight models whose internal representations are accessible.

Data and Code Availability

The data and Python code needed to reproduce all experiments reported in the main text and appendices are publicly available at <https://github.com/Jiajun-Bao/LLM-Graph-Signal-Processing>.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. *Phi-4 technical report*. *arXiv preprint arXiv:2412.08905*.

Nasir Ahmed, T. Raj Natarajan, and Kamisetty R. Rao. 1974. *Discrete cosine transform*. *IEEE Trans. Comput.*, C-23(1):90–93.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Noumane 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, and 4 others. 2025. *SmolLM3: smol, multilingual, long-context reasoner*.

Arjun Balaji. 2026. *Sparse spectral signatures of reasoning: Model-agnostic verification via sentence-level graph signals*. In *ICLR 2026 Workshop on Logical Reasoning of Large Language Models*.

Jiajun Bao, Nicolas Boullé, Toni J.B. Liu, Raphaël Sarfati, and Christopher J. Earls. 2026. *Text-trained LLMs can zero-shot extrapolate PDE dynamics, revealing a three-stage in-context learning mechanism*. In *AI&PDE: ICLR 2026 Workshop on AI and Partial Differential Equations*.

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. 2025. *Hallucination detection in LLMs using spectral features of attention maps*. In *Proc. EMNLP*, pages 24354–24385.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. 2025. *Thought anchors: Which LLM reasoning steps matter?* In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. 2016. *Discovering governing equations from data by sparse identification of nonlinear dynamical systems*. *Proc. Natl. Acad. Sci. USA*, 113(15):3932–3937.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. *LexGLUE: A benchmark dataset for legal language understanding in English*. In *Proc. ACL*, pages 4310–4330.

Despina Christou and Grigorios Tsoumakas. 2025. *Artificial relationships in fiction: A dataset for advancing NLP in literary domains*. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 130–147.

Fan R. K. Chung. 1997. *Spectral Graph Theory*. AMS.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. *A discourse-aware attention model for abstractive summarization of long documents*. In *Proc. NAACL-HLT*, pages 615–621.

- Yijia Dai, Zhaolin Gao, Yahya Sattar, Sarah Dean, and Jennifer J. Sun. 2025. [Pre-trained large language models learn to predict hidden Markov models in context](#). In *NeurIPS*.
- Zhichen Dong, Yang Li, Yuhan Sun, Weixun Wang, Yijia Luo, Zinian Peng, Wenbo Su, Cheng Yu, Bo Zheng, and Junchi Yan. 2026. [How does reasoning flow? tracing attention-induced information flow for targeted RL in LLMs](#). In *ICML*.
- Andreas Eisele and Yu Chen. 2010. [MultiUN: A multilingual corpus from United Nation documents](#). In *Proc. LREC*.
- J. Doynne Farmer. 1982. [Information dimension and the probabilistic structure of chaos](#). *Z. Naturforsch. A*, 37(11):1304–1325.
- Miroslav Fiedler. 1973. [Algebraic connectivity of graphs](#). *Czechoslov. Math. J.*, 23(2):298–305.
- Fabrizio Frasca, Guy Bar-Shalom, Yftah Ziser, and Hagai Maron. 2026. [Neural message-passing on attention graphs for hallucination detection](#). In *ICLR*.
- Thomas M. J. Fruchterman and Edward M. Reingold. 1991. [Graph drawing by force-directed placement](#). *Softw. Pract. Exp.*, 21(11):1129–1164.
- Gemma Team. 2026. [Gemma 4 technical report](#). *arXiv preprint arXiv:2607.02770*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 540 others. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Nate Gruver, Marc Anton Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. [Large language models are zero-shot time series forecasters](#). In *NeurIPS*.
- Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. [Exploring network structure, dynamics, and function using NetworkX](#). In *Proc. SciPy*.
- David K. Hammond, Pierre Vandergheynst, and Rémi Gribonval. 2011. [Wavelets on graphs via spectral graph theory](#). *Appl. Comput. Harmon. Anal.*, 30(2):129–150.
- Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. [Teaching machines to read and comprehend](#). In *NIPS*.
- Niall Hurley and Scott Rickard. 2009. [Comparing measures of sparsity](#). *IEEE Trans. Inf. Theory*, 55(10):4723–4741.
- Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. [Time-LLM: Time series forecasting by reprogramming large language models](#). In *ICLR*.
- Lou Jost. 2006. [Entropy and diversity](#). *Oikos*, 113(2):363–375.
- Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Yacine Jernite, Margaret Mitchell, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. 2023. [The Stack: 3 TB of permissively licensed source code](#). *Trans. Mach. Learn. Res.*
- James R. Lee, Shayan Oveis Gharan, and Luca Trevisan. 2014. [Multiway spectral partitioning and higher-order Cheeger inequalities](#). *J. ACM*, 61(6):37:1–37:30.
- Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Olek Shliazhko, and 48 others. 2023. [StarCoder: may the source be with you!](#) *Trans. Mach. Learn. Res.*
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [DailyDialog: A manually labelled multi-turn dialogue dataset](#). In *Proc. IJCNLP*, pages 986–995.
- Toni J.B. Liu, Nicolas Boullé, Raphaël Sarfati, and Christopher J. Earls. 2024. [LLMs learn governing principles of dynamical systems, revealing an in-context neural scaling law](#). In *Proc. EMNLP*, pages 15097–15117.
- Toni J.B. Liu, Nicolas Boullé, Raphaël Sarfati, and Christopher J. Earls. 2025. [Density estimation with LLMs: a geometric investigation of in-context learning trajectories](#). In *ICLR*, pages 22163–22197.
- Edward N. Lorenz. 1963. [Deterministic nonperiodic flow](#). *J. Atmos. Sci.*, 20(2):130–141.
- Robert M. May. 1976. [Simple mathematical models with very complicated dynamics](#). *Nature*, 261(5560):459–467.
- Valentin Noël. 2025. [A graph signal processing framework for hallucination detection in large language models](#). *arXiv preprint arXiv:2510.19117*.
- Valentin Noël. 2026. [Geometry of reason: Spectral signatures of valid mathematical reasoning](#). In *ICML*.
- OpenAI. 2023. [GPT-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- Antonio Ortega, Pascal Frossard, Jelena Kovačević, José M. F. Moura, and Pierre Vandergheynst. 2018. [Graph signal processing: Overview, challenges, and applications](#). *Proc. IEEE*, 106(5):808–828.

- Julianna Piskorz, Kasia Kobalczuk, and Mihaela van der Schaar. 2026. [Eliciting numerical predictive distributions of LLMs without auto-regression](#). In *ICLR*.
- James Requeima, John Bronskill, Dami Choi, Richard E. Turner, and David Duvenaud. 2024. [LLM processes: Numerical predictive distributions conditioned on natural language](#). In *NeurIPS*, pages 109609–109671.
- Raphaël Sarfati, Eric Bigelow, Daniel Wurgaft, Siddharth Boppana, Jack Merullo, Atticus Geiger, Owen Lewis, Tom McGrath, and Ekdeep Singh Lubana. 2026. [The shape of beliefs: Geometry, dynamics, and interventions along representation manifolds of language models’ posteriors](#). *arXiv preprint arXiv:2602.02315*.
- Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. [Get to the point: Summarization with pointer-generator networks](#). In *Proc. ACL*, pages 1073–1083.
- Claude E. Shannon. 1948. [A mathematical theory of communication](#). *Bell Syst. Tech. J.*, 27(3):379–423.
- David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. 2013. [The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains](#). *IEEE Signal Process. Mag.*, 30(3):83–98.
- Steven H. Strogatz. 2024. [Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering](#), 3rd edition. Chapman and Hall/CRC.
- Jörg Tiedemann. 2012. [Parallel data, tools and interfaces in OPUS](#). In *Proc. LREC*, pages 2214–2218.
- Robert Vacareanu, Vlad Andrei Negru, Vasile Suciu, and Mihai Surdeanu. 2024. [From words to numbers: Your large language model is secretly a capable regressor when given in-context examples](#). In *COLM*.
- Peter Walters. 1982. [An Introduction to Ergodic Theory](#). Springer.
- Wikimedia Foundation. 2023. [Wikimedia downloads](#).
- Oussama Zekri, Ambroise Odonnat, Abdelhakim Benechehab, Linus Bleistein, Nicolas Boullé, and Ievgen Redko. 2024. [Large language models as Markov chains](#). *arXiv preprint arXiv:2410.02724*.
- Yuanzhao Zhang and William Gilpin. 2025. [Zero-shot forecasting of chaotic systems](#). In *ICLR*.

Appendix Contents

- A. Main Numerical Input Suite and Next-State Extrapolation
- B. Entropy-Based Quantification of Input Dynamical Complexity

- C. Laplacian Quadratic Forms and Graph Fourier Modes
- D. Robustness and Layerwise Consistency of Graph-Spectral Diagnostics
- E. Attention-Induced Token Graphs: Construction and Visualization Robustness
- F. Attention-Induced Token-Graph Visualizations at Additional Transformer Layers
- G. Graph-Spectral Diagnostics Across Model Families, Scales, and Instruction Tuning
- H. Attention-Only and Hidden-State-Only Non-Graph Baselines
- I. Attention-Induced Token Graphs for Additional Structured Numerical Inputs
- J. Preliminary Graph-Spectral Comparison of Numerical, Natural-Language, and Code Inputs
- A Main Numerical Input Suite and Next-State Extrapolation

Main numerical input suite. The main numerical input suite comprises ten one-dimensional trajectory families: a constant-sequence baseline whose value is sampled separately for each realization, six periodic logistic map families with periods 2, 4, 8, 16, 32, and 64, two chaotic logistic map families, and a family of trajectories obtained from the Lorenz x -component. For each family, we construct an ensemble of 20 trajectory realizations. The same realizations are used in the graph-spectral analysis presented in the main text and in the extrapolation analysis described below. Realization j uses a shared initial-state draw $x_0^{(j)} \sim \text{Unif}(0.3, 0.7)$ across these families. For the constant baseline, realization j is assigned the quantized state

$$q_{\text{const}}^{(j)} = \text{round}\left(50 + 900x_0^{(j)}\right).$$

This quantized value is repeated across all 1000 states of realization j and varies across realizations. Logistic map inputs are generated from $x_{t+1} = rx_t(1 - x_t)$. The periodic cases use $r = 3.2, 3.5, 3.55, 3.566, 3.5692, \text{ and } 3.5698$, corresponding to periods 2, 4, 8, 16, 32, and 64, and the chaotic cases use $r = 3.9$ and $r = 4.0$. For each realization of a periodic case with expected period

p , we generate a 20,000-state trajectory, identify the earliest point after which $|x_{t+p} - x_t| < 10^{-10}$ holds for the remainder of the trajectory, and extract a 1000-state window after applying a random phase offset in $\{0, \dots, p-1\}$. We use 20 realizations primarily to characterize variation in the graph-spectral diagnostics for families with greater across-realization trajectory diversity, particularly the higher-period and chaotic families. We retain the same ensemble size for the lower-period families to ensure consistent aggregation across inputs, although a converged period- p orbit has only p possible phases, so duplicate phase selections are expected. For the chaotic logistic families, we instead use direct 1000-state trajectories from the realization-specific $x_0^{(j)}$. The Lorenz trajectory family is generated from

$$\begin{aligned}\dot{x} &= \sigma(y - x), \\ \dot{y} &= x(\rho - z) - y, \\ \dot{z} &= xy - \beta z,\end{aligned}$$

with $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$, the classical chaotic parameter setting (Lorenz, 1963). For realization j , the initial condition is $(x, y, z) = (-8.0 + (x_0^{(j)} - 0.5), 7.0, 27.0)$, varying the initial x -coordinate around the standard $(-8, 7, 27)$ initialization (Brunton et al., 2016). We integrate over $\tau \in [0, 100]$ using SciPy’s `solve_ivp` with its default solver settings, record 1000 evenly spaced time points, and retain only the x -component, yielding, for each realization, a partially observed one-dimensional trajectory from a continuous-time chaotic system.

Trajectory quantization and prompt serialization. For a logistic or Lorenz realization $\mathbf{u} = (u_1, \dots, u_{1000})$, let

$$u_{\min} = \min_i u_i \quad \text{and} \quad u_{\max} = \max_i u_i.$$

If $u_{\max} = u_{\min}$, we set $q_i = 500$. Otherwise, the quantized state at position i is

$$q_i = \text{round}\left(50 + 900 \frac{u_i - u_{\min}}{u_{\max} - u_{\min}}\right).$$

Here, $\text{round}(\cdot)$ denotes rounding to the nearest integer, so $q_i \in \{50, \dots, 950\}$. The sampled constant baseline bypasses this mapping and uses $q_{\text{const}}^{(j)}$ as defined above. The scaling bounds are computed once from the full 1000-state realization before context truncation. Each quantized state is formatted

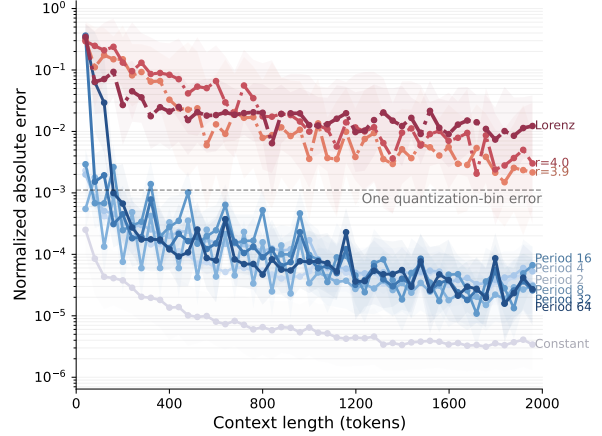


Figure 5: Context-length scaling of next-state extrapolation error with Llama-3.2-3B. Curves show normalized expected-value error for predicting the next quantized state from the preceding serialized numerical context. Longer context generally reduces extrapolation error across the main numerical input families. Several low-error inputs have tightly clustered output-level curves, motivating the representation-level analysis: graph-spectral diagnostics reveal more ordered differences in the model’s internal token organization that are less apparent from extrapolation error alone. The shaded bands show one log standard deviation across realizations.

as a zero-padded three-digit string and followed by a comma delimiter, including the final state. Thus, a realization is serialized schematically as “ $q_1, q_2, \dots, q_{1000}$ ”.

Next-state extrapolation evaluation. Fig. 5 reports output-level next-state extrapolation error across context lengths for the main numerical input families. We evaluate 49 context fractions from 2% through 98% in increments of two percentage points. For each fraction c , the prompt contains $t_c = \text{round}(c(1000 - 1))$ states, reserving the following state as the prediction target. For input family f , realization j , and state index t , the prompt contains the first t comma-delimited quantized states and the target is the $(t + 1)$ -st state, denoted $y_{j,t+1}^f \in \{50, \dots, 950\}$. At the final comma position of the prompt, the model produces next-token logits over its full vocabulary. For each integer $v \in \mathcal{V} = \{0, \dots, 999\}$, let $z_{j,t}^f(v)$ denote the logit assigned to the token encoding the zero-padded three-digit representation of v . We retain the full numeric output range 000–999, so values outside the input interval 050–950 provide buffers that reduce boundary truncation of probability mass near the quantization boundaries. We then evaluate a softmax distribution restricted to these numeric

tokens

$$p_{j,t}^f(v) = \frac{\exp z_{j,t}^f(v)}{\sum_{u \in \mathcal{V}} \exp z_{j,t}^f(u)}.$$

The corresponding expected next state is

$$\hat{y}_{j,t+1}^f = \sum_{v \in \mathcal{V}} v p_{j,t}^f(v).$$

We measure extrapolation error using the normalized absolute difference between the expected next state and the target:

$$e_{j,t}^f = \frac{|\hat{y}_{j,t+1}^f - y_{j,t+1}^f|}{950 - 50}.$$

The plotted curves are geometric means of $e_{j,t}^f$ over the 20 realizations at each t , with shaded bands corresponding to plus or minus one standard deviation on the log-error scale. This behavioral check shows that extrapolation accuracy generally improves with longer context across all main numerical input families, while chaotic logistic and Lorenz inputs exhibit larger extrapolation errors than simpler periodic and constant inputs. At the same time, several low-error inputs have tightly clustered output-level curves, motivating the representation-level analysis in the main text.

The dashed reference line marks one quantization-bin error, $1/900$: because the error is normalized by $950 - 50 = 900$, this corresponds to an absolute error of one quantized state unit. Periodic inputs often lie below this line, indicating that the restricted-softmax expected output is within one bin of the ground-truth quantized state; their curves can still decrease with longer context, indicating further reductions in expected-value error even after reaching this output-level scale.

B Entropy-Based Quantification of Input Dynamical Complexity

An invariant probability measure μ provides a time-independent summary of state-space occupancy in a dynamical system (Walters, 1982). Coarse-grained entropies of invariant state distributions are a classical tool for characterizing the statistical structure of dynamical systems (Farmer, 1982). We use Shannon entropy (Shannon, 1948) to summarize how broadly this distribution is spread across the visited states. This quantity provides quantitative support for the operational ordering of dynamical complexity used in Section 3.1.

Input type	Mean $\pm$ SD (bits)
Constant sampled	0.000 $\pm$ 0.000
Logistic $r = 3.2$ (period 2)	1.000 $\pm$ 0.000
Logistic $r = 3.5$ (period 4)	2.000 $\pm$ 0.000
Logistic $r = 3.55$ (period 8)	3.000 $\pm$ 0.000
Logistic $r = 3.566$ (period 16)	3.625 $\pm$ 0.000
Logistic $r = 3.5692$ (period 32)	4.039 $\pm$ 0.001
Logistic $r = 3.5698$ (period 64)	4.293 $\pm$ 0.002
Logistic $r = 3.9$ (chaotic)	6.222 $\pm$ 0.057
Logistic $r = 4.0$ (chaotic)	6.295 $\pm$ 0.053
Lorenz	6.360 $\pm$ 0.022

Table 1: Coarse-grained state-distribution entropy of the quantized numerical input families introduced in Section 3.1. Values are reported as the mean $\pm$ sample standard deviation in bits across 20 realizations at trajectory length $T = 1000$, estimated using 100 equal-width bins with shared edges across all input families and realizations.

For a period- p orbit with distinct states $x_1, \dots, x_p$, the invariant measure assigns probability $1/p$ to each state:

$$\mu_p = \frac{1}{p} \sum_{j=1}^p \delta_{x_j},$$

where δ_{x_j} denotes a unit point mass at x_j . Its state-distribution entropy is therefore

$$H_{\text{state}}(\mu_p) = - \sum_{j=1}^p \frac{1}{p} \log_2 \frac{1}{p} = \log_2 p.$$

Thus, the constant input has entropy 0, while periods 2, 4, 8, 16, 32, and 64 have ideal entropies 1, 2, 3, 4, 5, and 6 bits, respectively. To compare all input families on a common basis, we estimate state-distribution entropy directly from each finite quantized trajectory $q_1, \dots, q_T$ used as an LLM input, where $q_t \in \{50, \dots, 950\}$, using 100 equal-width bins with shared edges spanning $[49.5, 950.5]$ across all input families and realizations. Letting B_b denote the b -th bin, we compute

$$\begin{aligned} \hat{p}_{T,b} &= \frac{1}{T} \sum_{t=1}^T \mathbf{1}\{q_t \in B_b\}, \\ \hat{H}_T &= - \sum_{b: \hat{p}_{T,b} > 0} \hat{p}_{T,b} \log_2 \hat{p}_{T,b}. \end{aligned}$$

Applying this estimator to realization j of input family f gives $\hat{H}_{j,T}^f$. Table 1 reports the mean and sample standard deviation of $\hat{H}_{j,1000}^f$ over $j = 1, \dots, 20$ for each family f . The empirical entropies are consistent with the intended complexity ordering: entropy increases with logistic map

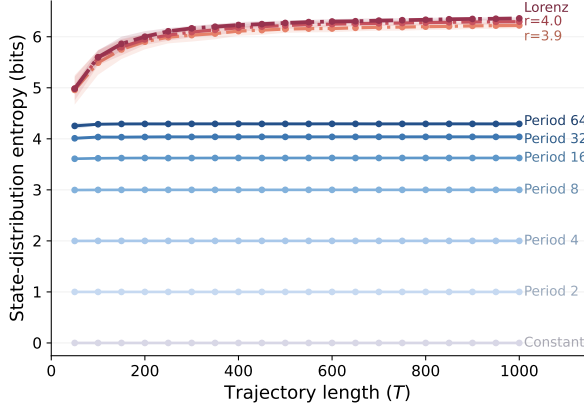


Figure 6: Empirical state-distribution entropy as a function of trajectory length T . Periodic estimates stabilize by $T = 100$, while chaotic logistic and Lorenz estimates also stabilize relatively early, with little change beyond $T = 400$. Entropy is estimated using 100 equal-width bins with shared edges across all input families and realizations. Curves and shading show the mean and ± 1 sample standard deviation across the same 20 realizations used in Section 3.1.

period, while the chaotic logistic and Lorenz inputs exhibit the highest state-distribution entropies. The higher-period values fall below the ideal values $\log_2 p$ because the orbit states are nonuniformly spaced across state space, causing multiple distinct states to fall within the same histogram bin. The reported quantity is therefore a coarse-grained estimate of the entropy of the invariant state distribution.

Figure 6 shows that empirical estimates of state-distribution entropy stabilize with increasing trajectory length, consistent with entropy characterizing the underlying invariant state distribution. The estimates are computed at the context lengths used in the main analysis: $T = 50, 100, \dots, 1000$ numerical states, corresponding to $N = 100, 200, \dots, 2000$ tokens under the $N = 2T$ serialization described in Section 3.1. The periodic estimates are already close to their $T = 1000$ values by $T = 100$. The chaotic logistic and Lorenz estimates increase most sharply between $T = 50$ and $T = 200$, then gradually level off, with only modest changes beyond $T = 400$. Thus, the context-dependent growth in the spectral entropy of the LLM’s internal representations cannot be explained simply by variation in the estimated state-distribution entropy of the inputs.

C Laplacian Quadratic Forms and Graph Fourier Modes

We briefly review why the graph Laplacian measures signal variation over a weighted graph. Let $W \in \mathbb{R}^{N \times N}$ be a symmetric nonnegative weighted adjacency matrix, let $D = \text{diag}(W\mathbf{1})$, and let $L = D - W$ be the combinatorial graph Laplacian.

First consider a scalar graph signal $\mathbf{x} = (x_1, \dots, x_N)^\top$, where x_i is the value assigned to node i . The Laplacian quadratic form expands as

$$\begin{aligned} \mathbf{x}^\top L \mathbf{x} &= \sum_i D_{ii} x_i^2 - \sum_{i,j} W_{ij} x_i x_j \\ &= \frac{1}{2} \sum_{i,j} W_{ij} (x_i - x_j)^2. \end{aligned}$$

Thus, $\mathbf{x}^\top L \mathbf{x}$ measures the roughness of the node-wise signal over the graph: differences across node pairs joined by larger weights contribute more strongly to the quadratic form.

This identity also motivates the graph Fourier interpretation of Laplacian eigenvectors. Writing

$$\begin{aligned} L &= U \Lambda U^\top, \\ \Lambda &= \text{diag}(\mu_1, \dots, \mu_N), \quad \hat{\mathbf{x}} = U^\top \mathbf{x}, \end{aligned}$$

where $0 \leq \mu_1 \leq \dots \leq \mu_N$, let $\mathbf{u}_m$ denote the m -th column of U . These eigenvectors form the graph Fourier basis used in Section 3.2, and $\hat{x}_m = \mathbf{u}_m^\top \mathbf{x}$ is the graph Fourier coefficient of $\mathbf{x}$ at mode m . The quadratic form can then be written as

$$\mathbf{x}^\top L \mathbf{x} = \mathbf{x}^\top U \Lambda U^\top \mathbf{x} = \hat{\mathbf{x}}^\top \Lambda \hat{\mathbf{x}} = \sum_{m=1}^N \mu_m \hat{x}_m^2.$$

In particular, $\mathbf{u}_m^\top L \mathbf{u}_m = \mu_m$, so μ_m gives the graph variation of the corresponding Fourier mode. Because $L\mathbf{1} = \mathbf{0}$ and L is positive semidefinite, its smallest eigenvalue is $\mu_1 = 0$. More generally, the multiplicity of the zero eigenvalue equals the number of connected components of the graph. Thus, for a connected graph, the zero eigenvalue is simple and the second-smallest eigenvalue satisfies $\mu_2 > 0$. The full weighted attention graphs used in our spectral diagnostics are connected. This connectedness condition applies throughout our analysis. Ordering the eigenvectors by increasing eigenvalue therefore orders the graph Fourier modes from low frequency, corresponding to smooth variation across strongly weighted edges, to high frequency, corresponding to sharper variation.

The same interpretation extends to vector-valued hidden-state signals. Suppose each node i is assigned a hidden-state vector $X_i \in \mathbb{R}^d$, and collect these vectors as the rows of a matrix $X \in \mathbb{R}^{N \times d}$. Applying the scalar identity to each hidden dimension gives

$$\text{tr}(X^\top LX) = \frac{1}{2} \sum_{i,j} W_{ij} \|X_i - X_j\|_2^2.$$

Equivalently, if $\hat{X} = U^\top X$, then

$$\text{tr}(X^\top LX) = \sum_{m=1}^N \mu_m \|\hat{\mathbf{x}}_m\|_2^2,$$

where $\hat{\mathbf{x}}_m$ denotes the m -th row of $\hat{X}$. In our setting, $W = \bar{W}^{(\ell)}$ is the attention-derived weighted adjacency matrix and $X = X^{(\ell)}$ is the hidden-state matrix entering layer ℓ . Therefore, the Laplacian quadratic form quantifies how rapidly hidden-state representations vary across attention-weighted token interactions, and the Laplacian eigenvectors provide the graph Fourier modes used in the spectral diagnostics.

Because U is orthonormal, Parseval's identity gives

$$\|X\|_F^2 = \|\hat{X}\|_F^2 = \sum_{m=1}^N \|\hat{\mathbf{x}}_m\|_2^2.$$

This identity justifies treating $\|\hat{\mathbf{x}}_m\|_2^2$ as the energy assigned to mode m in the HFER and spectral entropy diagnostics.

D Robustness and Layerwise Consistency of Graph-Spectral Diagnostics

D.1 Robustness of Graph-Spectral Diagnostics

Section 3.2 defines the normalized Fiedler value as the second-smallest normalized Laplacian eigenvalue and HFER as the fraction of hidden-state energy in the highest-frequency 50% of graph Fourier modes. Here, we test whether the observed Fiedler-value separation extends to multiway graph structure and whether the HFER trends are robust to the choice of high-frequency cutoff by examining additional low-end eigenvalues and alternative HFER cutoffs.

Low-end normalized Laplacian spectrum. Let the ordered eigenvalues of the layer- ℓ normalized Laplacian $\mathcal{L}^{(\ell)}$ be

$$0 = \lambda_1^{(\ell)} \leq \lambda_2^{(\ell)} \leq \dots \leq \lambda_N^{(\ell)} \leq 2.$$

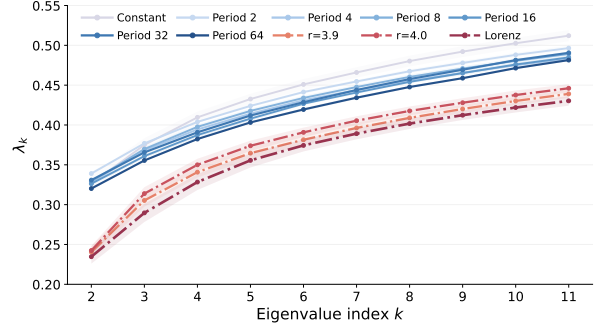


Figure 7: Low end of the normalized Laplacian spectrum at $N = 2000$, averaged across layers. For each $k = 2, \dots, 11$, curves show the mean λ_k across 20 realizations, and shading denotes ± 1 standard deviation. The chaotic logistic map and Lorenz inputs remain below the constant and periodic inputs throughout this range, showing that the separation observed in the normalized Fiedler value in Fig. 3 extends across the low end of the spectrum.

Writing λ_k without a layer superscript for the layer average, we define

$$\lambda_k := \frac{1}{L} \sum_{\ell=1}^L \lambda_k^{(\ell)}.$$

Thus, λ_2 is the layer-averaged normalized Fiedler value reported in the main analysis. To determine whether the observed separation extends beyond the two-way connectivity captured by λ_2 , we examine the first ten nontrivial eigenvalues, $k = 2, \dots, 11$. The additional eigenvalues at the low end of the normalized Laplacian spectrum are related to multiway sparse-partition structure through higher-order Cheeger inequalities (Lee et al., 2014). We focus on $N = 2000$, where Fig. 3 in Section 4.1 shows a clear separation in λ_2 between the chaotic and non-chaotic inputs. Fig. 7 shows that this separation persists throughout the examined low end of the spectrum: the chaotic logistic map and Lorenz inputs remain systematically below the constant and periodic inputs for $k = 2, \dots, 11$. This systematic downward shift is consistent with a greater tendency toward multiple weakly connected graph regions under chaotic inputs, indicating that the reduced global integration observed through λ_2 is not confined to a single eigenvalue.

HFER cutoff sensitivity. To test whether the observed HFER trends depend on the default 50% high-frequency tail, let $M_q = \text{round}(qN)$ and $K_q = N - M_q$, where M_q is the number of highest-frequency modes retained for tail fraction q . We

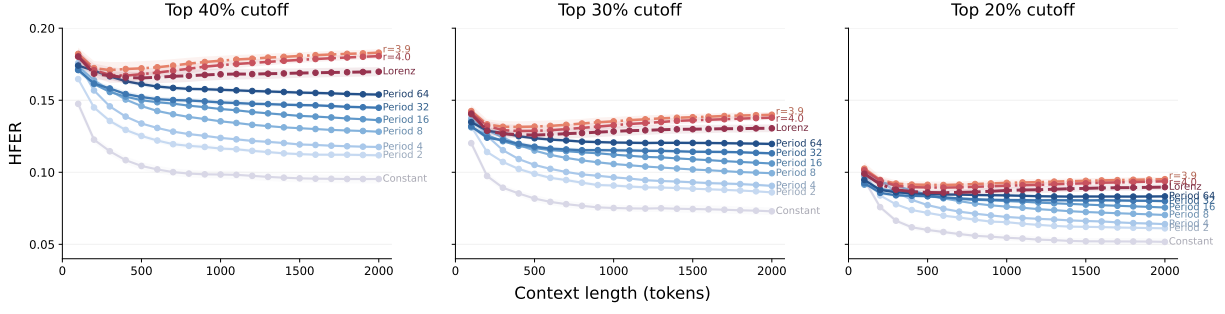


Figure 8: Layer-averaged HFER cutoff sensitivity. The panels report HFER_q using the highest-frequency 40%, 30%, and 20% of graph Fourier modes. Under the same evaluation setup as Section 4.1, curves and shading show the mean and ± 1 standard deviation across 20 realizations. Across all three cutoffs, chaotic inputs exhibit higher HFER at longer contexts, while the non-chaotic inputs retain the periodicity-dependent ladder observed with the default 50% high-frequency tail in Fig. 3. The qualitative separation is therefore robust to the number of high-frequency modes included.

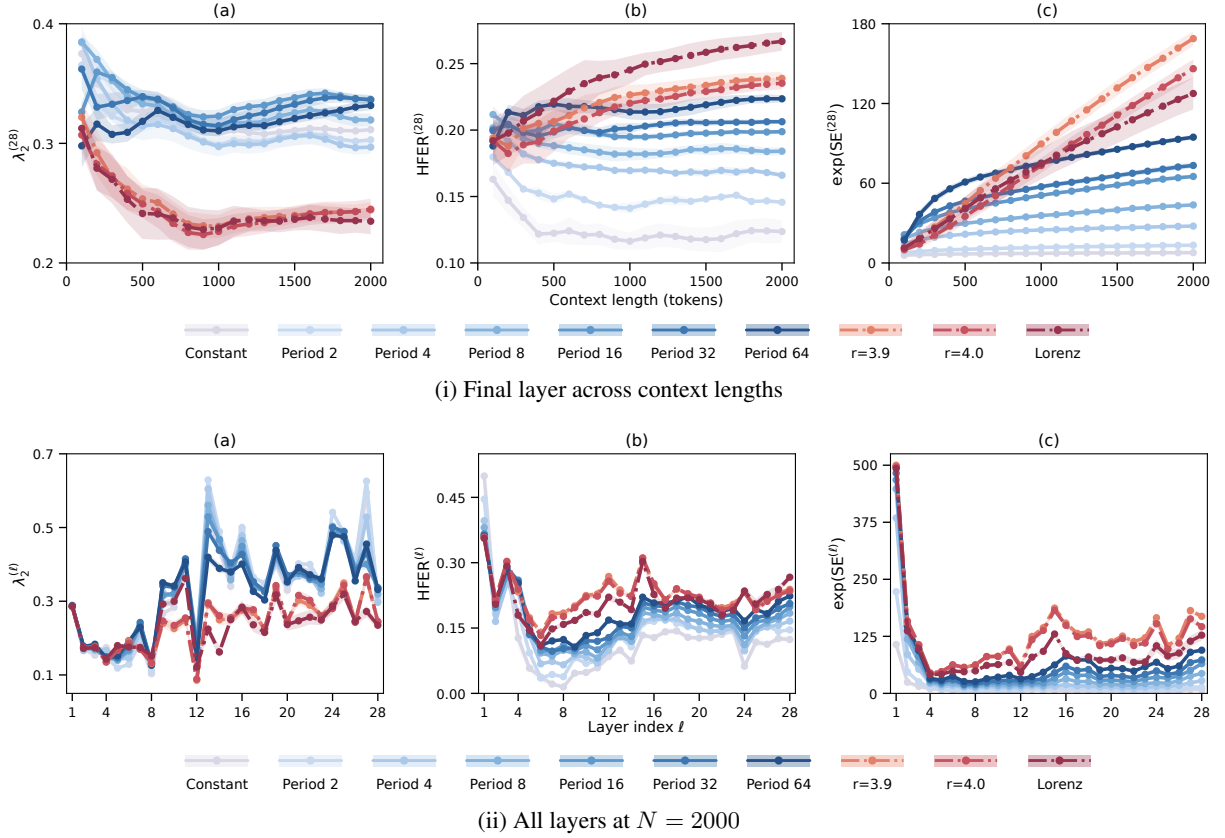


Figure 9: Layer-specific graph-spectral diagnostics for the main numerical input families using Llama-3.2-3B. Row (i) shows final-layer diagnostics across context lengths, while row (ii) shows the corresponding diagnostics across all 28 model layers at $N = 2000$. Within each row, panels (a)–(c) report normalized Fiedler value λ_2 , HFER, and effective spectral support $\exp(\text{SE})$, respectively. Curves and shading denote the mean and ± 1 standard deviation across the same 20 realizations analyzed in Section 4.1. The final-layer results reproduce the broad context- and complexity-dependent behavior of the layer-averaged analysis, although the single-layer curves are less smooth. In the fixed-context depth profiles, most layers exhibit the same broad complexity-dependent ordering as the layer-averaged and final-layer results, with only a few of the earliest layers showing weaker separation.

define

$$\text{HFER}_q^{(\ell)} = \sum_{m=K_q+1}^N p_m^{(\ell)},$$

$$\text{HFER}_q = \frac{1}{L} \sum_{\ell=1}^L \text{HFER}_q^{(\ell)}.$$

Thus, $q = 0.4, 0.3$, and 0.2 retain the highest-frequency 40%, 30%, and 20% of graph Fourier modes, corresponding to cutoffs at $K_q = 0.6N$, $0.7N$, and $0.8N$, respectively, for the evaluated context lengths. The default diagnostic in Section 3.2 is $\text{HFER}_{0.5}$. As shown in Fig. 3 of Section 4.1, at longer contexts this default separates the higher-complexity chaotic inputs from the simpler constant and periodic inputs and reveals a clear ladder among the periodic inputs. Fig. 8 shows the same qualitative behavior for $q = 0.4, 0.3$, and 0.2 : chaotic inputs retain higher HFER, while the periodic inputs preserve their period-dependent ordering. The persistence of these trends under increasingly selective high-frequency tails indicates that the observed behavior is not specific to the default 50% high-frequency tail.

D.2 Layerwise Graph-Spectral Diagnostics

The main analysis in Section 4.1 reports graph-spectral diagnostics averaged over transformer layers, providing a compact summary of how the diagnostic patterns vary across input families and context lengths. Here, we present two complementary views of the corresponding layerwise behavior. We first show the final-layer diagnostics across context lengths as an illustrative layer-specific example, and then examine all layers at $N = 2000$ to characterize how the same broad input-family structure is distributed across model depth. Together, these views show that the layer-averaged results provide a holistic summary of the recurring layerwise behavior while smoothing layer-specific variation.

Final-layer context scaling. The top row of Fig. 9 presents the three diagnostics at the final model layer as an illustrative layer-specific example. At longer contexts, the final-layer normalized Fiedler value primarily separates chaotic from non-chaotic inputs, whereas HFER and effective spectral support also reveal a finer period-dependent ordering among the periodic families. These patterns are broadly consistent with the layer-averaged analysis, although the single-layer curves are less smooth.

Consistency across model depth. Because presenting a separate context-length sweep for each layer of an L -layer model would require L figures, we complement the final-layer sweep with a fixed-context depth profile that compares all layers in a single view. The bottom row of Fig. 9 reports the normalized Fiedler value, HFER, and effective spectral support across all model layers at $N = 2000$. Overall, the layerwise profiles are consistent with the layer-averaged and final-layer results, with most layers exhibiting the same broad complexity-dependent ordering.

E Attention-Induced Token Graphs: Construction and Visualization Robustness

Attention matrices. Fig. 10 shows representative final-layer, head-averaged post-softmax attention matrices for the main numerical input families. These matrices are symmetrized to form the weighted adjacency matrices used in the spectral diagnostics and graph visualizations.

Hidden-state projections. Let $X^{(\ell)} \in \mathbb{R}^{N \times d}$ contain the token hidden states at layer ℓ , with one token per row. We compute the uncentered singular value decomposition $X^{(\ell)} = U_X^{(\ell)} \Sigma_X^{(\ell)} V_X^{(\ell)\top}$, where the subscript X distinguishes these hidden-state SVD quantities from the graph Laplacian eigenvectors $U^{(\ell)}$ introduced in Section 3.2. The full matrix of token-level SVD coordinates is

$$Z^{(\ell)} = U_X^{(\ell)} \Sigma_X^{(\ell)} = X^{(\ell)} V_X^{(\ell)}.$$

To construct the node colors, we assign the first three columns of $Z^{(\ell)}$ to the red, green, and blue channels, respectively. For each $k \in \{1, 2, 3\}$, let $m_k^{(\ell)} = \min_j Z_{jk}^{(\ell)}$ and $M_k^{(\ell)} = \max_j Z_{jk}^{(\ell)}$. Using a numerical tolerance of $\varepsilon = 10^{-12}$, we independently min-max scale each coordinate across nodes:

$$C_{ik}^{(\ell)} = \begin{cases} \frac{Z_{ik}^{(\ell)} - m_k^{(\ell)}}{M_k^{(\ell)} - m_k^{(\ell)}}, & M_k^{(\ell)} - m_k^{(\ell)} > \varepsilon, \\ 1/2, & \text{otherwise.} \end{cases}$$

Token i is then assigned the RGB color $\mathbf{c}_i^{(\ell)} = (C_{ik}^{(\ell)})_{k=1}^3 \in [0, 1]^3$. The fraction of hidden-state

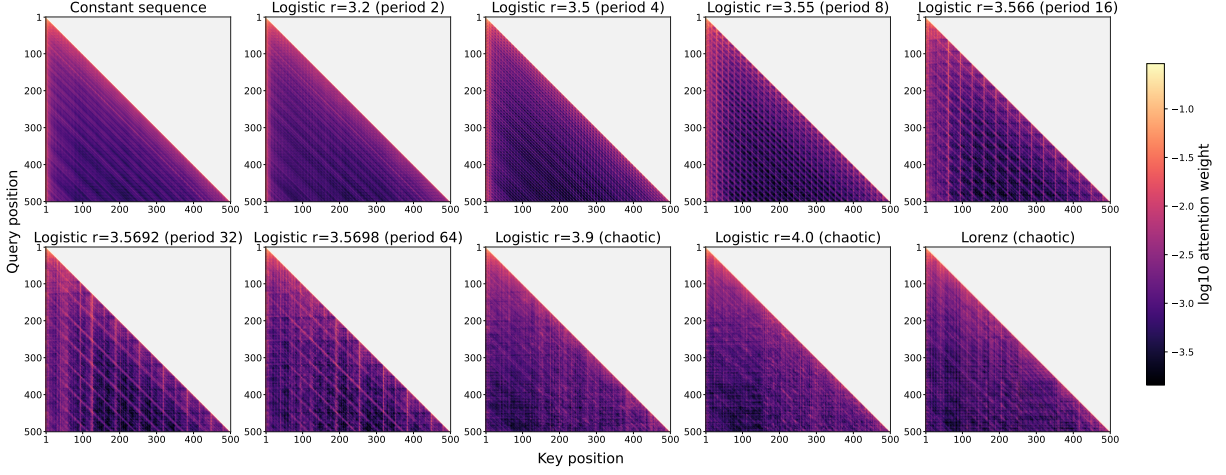


Figure 10: Last-layer attention matrices for Llama-3.2-3B across the main numerical input families at $N = 500$. Each panel shows the head-averaged matrix for a representative input.

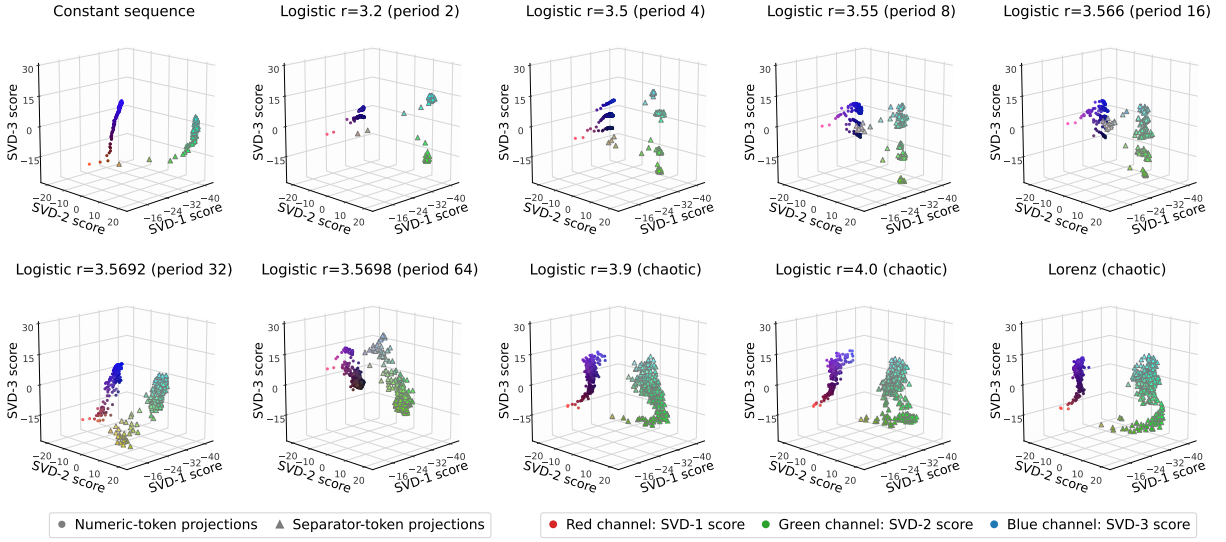


Figure 11: Last-layer hidden-state projections for Llama-3.2-3B across the main numerical input families at $N = 500$. Each panel shows token hidden states projected onto the leading three uncentered SVD coordinates, which define the RGB node colors used in the attention-induced graph visualizations.

energy captured by the first q components is

$$\eta_q^{(\ell)} = \frac{\sum_{k=1}^q (\sigma_{X,k}^{(\ell)})^2}{\sum_{k=1}^r (\sigma_{X,k}^{(\ell)})^2} = \frac{\sum_{k=1}^q (\sigma_{X,k}^{(\ell)})^2}{\|X^{(\ell)}\|_F^2}.$$

Here, $r = \min(N, d)$ is the number of singular values. At $N = 500$, averaging first across the main numerical input families within each layer and then across the analyzed layers gives $\eta_3 = 0.75$. Fig. 11 shows the resulting final-layer projections for all ten families, while Fig. 12 includes the layer-averaged cumulative SVD-energy diagnostic and its variation across layers.

Layout computation. We initialize node i on a token-ordered spiral by setting $t_i = (i - 1)/(N - 1)$, $\theta_i = -\pi/2 + 2\pi K_N t_i$, and

$$\mathbf{p}_i = (0.15 + 0.85t_i) \begin{bmatrix} \cos \theta_i \\ \sin \theta_i \end{bmatrix},$$

where $K_N = 2.25$ for $N < 80$ and $K_N = 2.75$ otherwise. We center the resulting coordinates and rescale them to unit maximum radius. This construction produces the initialization shown in the leftmost panels of Fig. 13. Because it depends only on N and token order, all inputs of the same length begin from identical node positions; their input-specific attention matrices $\bar{W}^{(\ell)}$ then determine the attractive forces that drive the subsequent reorganization. We compute the resulting positions using

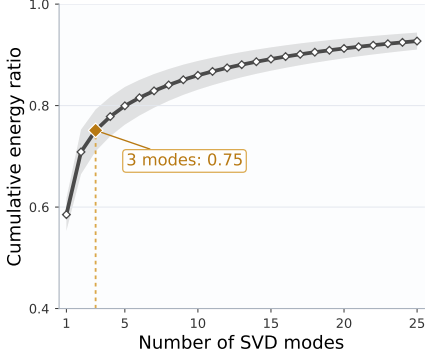


Figure 12: Cumulative hidden-state SVD-energy fraction for Llama-3.2-3B at $N = 500$, averaged first across the main numerical input families within each layer and then across the analyzed layers. The leading three components capture 75% of the hidden-state energy on average, supporting their use as RGB coordinates for token-node colors.

the implementation of the FR algorithm provided by the NetworkX Python library (Fruchterman and Reingold, 1991; Hagberg et al., 2008), with all off-diagonal entries of $\bar{W}^{(\ell)}$ used as edge weights. We run 100 iterations rather than the implementation’s default of 50 to allow further relaxation toward a stable configuration.

Layout evolution and iteration count. Figure 13 shows the evolution of the FR layout from the shared token-ordered spiral for representative constant and chaotic logistic map inputs. Most visible layout reorganization occurs by 50 iterations, with only minor changes thereafter. We therefore use the 100-iteration layouts throughout the token-graph visualizations.

Initialization robustness. To evaluate sensitivity to the exact initialization, we add independent Gaussian perturbations $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I_2)$ to the normalized spiral coordinates, reapply the same normalization, and rerun the 100-iteration layout. Fig. 14 compares $\sigma \in \{0.01, 0.05, 0.10\}$ for representative constant and chaotic logistic map inputs. Their broad organization remains stable across these perturbation magnitudes, indicating that the qualitative separation is not specific to the initial spiral coordinates.

Edge-display threshold sensitivity. Fig. 15 evaluates sensitivity to the multiplier κ in the edge-display threshold $\bar{W}_{ij}^{(\ell)} > \kappa/N$. Because this threshold affects only edge rendering, node positions, hidden-state colors, and graph-spectral di-

agnostics remain unchanged across panels. For both the representative constant and chaotic logistic map inputs, the qualitative structure of the displayed edge patterns remains consistent across $\kappa \in \{2, 4, 8\}$ and with the default visualization at $\kappa = 2.5$.

F Attention-Induced Token-Graph Visualizations at Additional Transformer Layers

Section 4.2 focuses on last-layer attention graphs because the final layer is closest to the prediction-facing representation used by the model. Fig. 16 uses layers 1 and 12 as illustrative snapshots of early and middle depth, respectively. In the layer-1 example, differentiation across input families is comparatively limited: although the graphs are not identical, their layouts share broad organization associated with numeric-versus-separator token roles and local sequence structure. In the layer-12 example, input-dependent differences are more pronounced and broadly track dynamical complexity: constant and lower-period inputs retain relatively regular, globally integrated layouts, whereas higher-period, chaotic logistic map, and Lorenz inputs exhibit less homogeneous organization and more pronounced mesoscale partitioning into localized subnetworks. These two layer-specific snapshots illustrate how cross-input differentiation can become more pronounced with depth; they do not establish a universal transition point or imply that all early and middle layers follow the same progression.

G Graph-Spectral Diagnostics Across Model Families, Scales, and Instruction Tuning

We investigate whether the graph-spectral behavior observed in the main analysis persists across model families and scales, and how instruction tuning affects this behavior. We first examine variation across model scale and instruction tuning within the Llama 3 family. We then extend the analysis to Phi-4 and SmoLM3, including a comparison between the base and instruction-tuned SmoLM3 variants.

Variation across model scales within the Llama 3 family. Using the same experimental setup as Section 4.1, Fig. 17(a)–(b) compares the layer-averaged diagnostics for Llama-3.2-1B and Llama-3.1-8B with the primary Llama-3.2-3B results. The

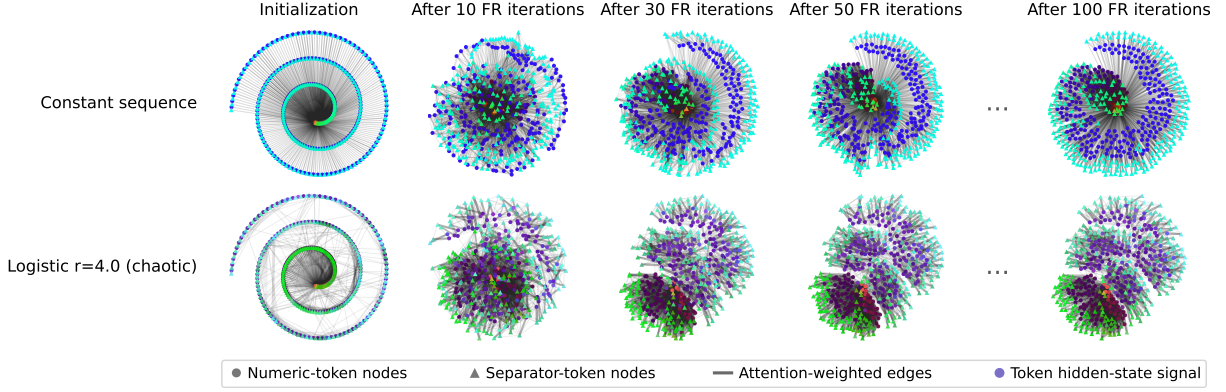


Figure 13: Evolution of attention-based node positions for Llama-3.2-3B at $N = 500$. Representative constant and chaotic logistic map inputs are shown at initialization and after 10, 30, 50, and 100 FR iterations. Both begin from the same token-ordered spiral, so subsequent differences reflect their input-specific attention weights. Most visible reorganization occurs by 50 iterations, with only minor changes thereafter; we use the 100-iteration layouts for subsequent token-graph visualizations.

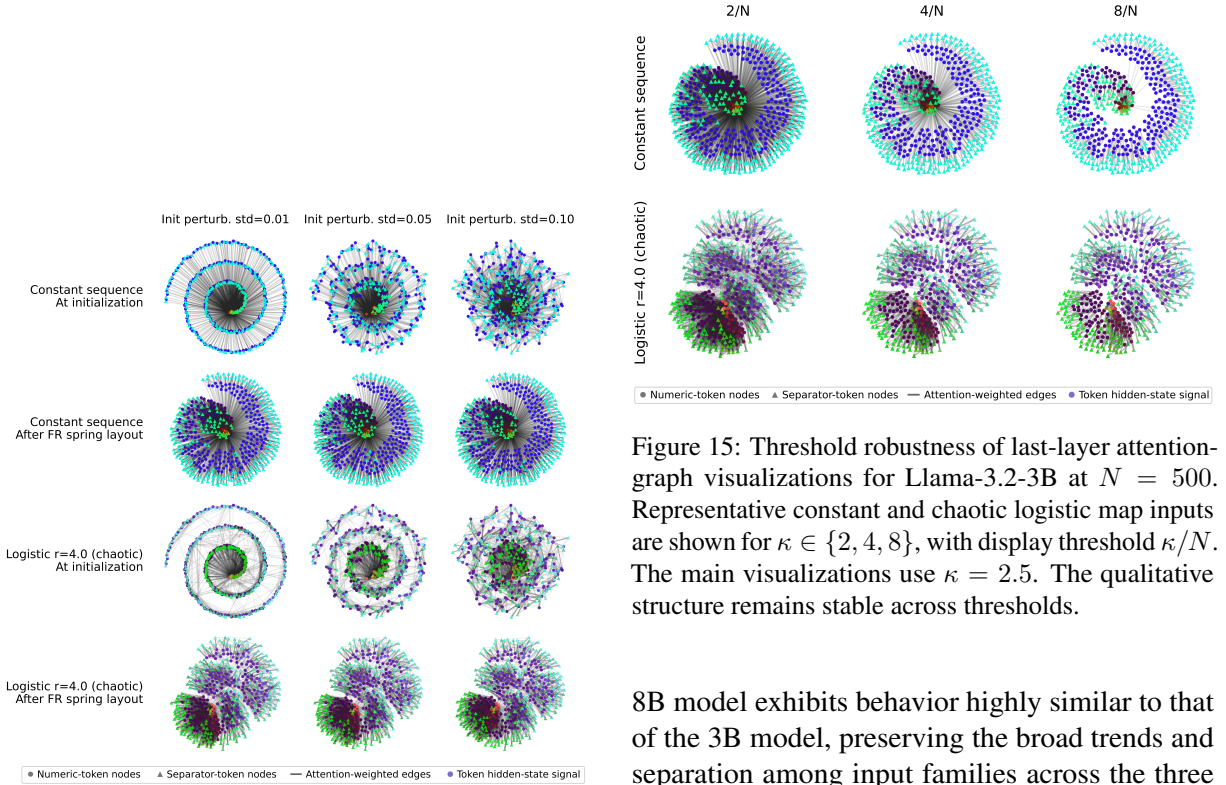


Figure 14: Layout-initialization robustness of last-layer attention-graph visualizations for Llama-3.2-3B at $N = 500$. Constant and chaotic logistic map examples are initialized from noisy spiral coordinates with perturbation standard deviations 0.01, 0.05, and 0.10, then compared before and after FR layout relaxation. The qualitative organization is preserved across perturbation scales.

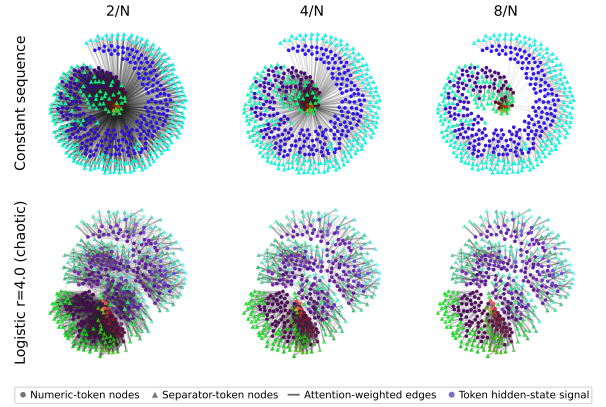
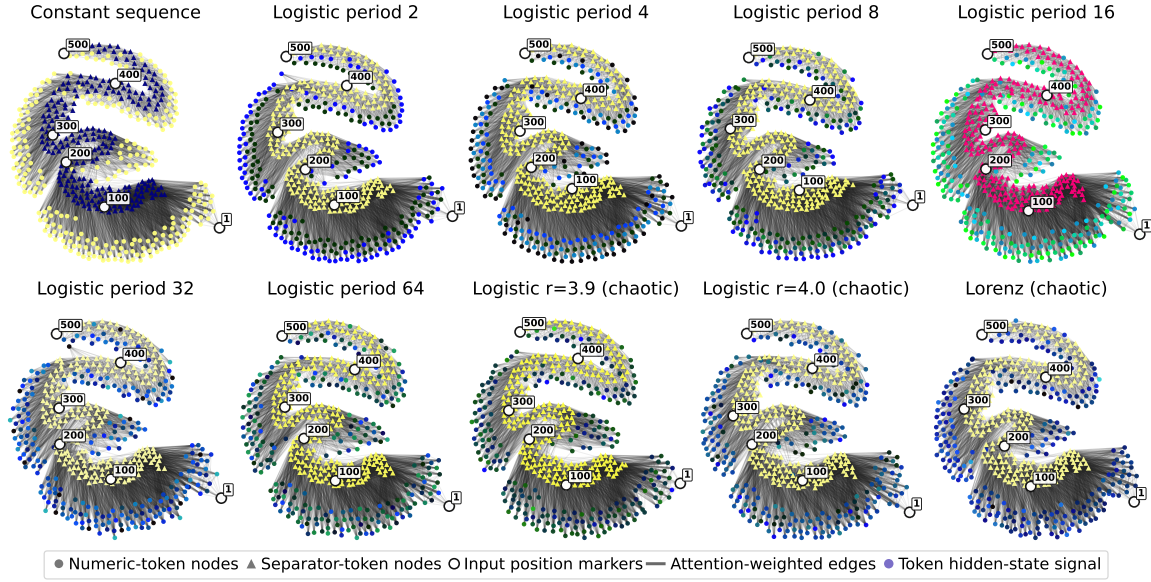


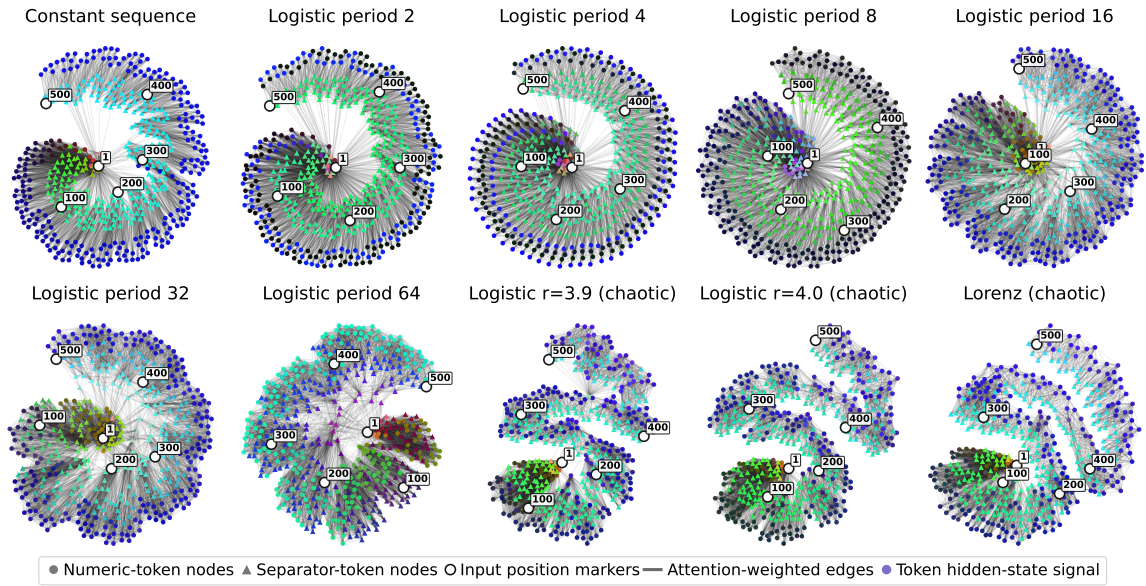
Figure 15: Threshold robustness of last-layer attention-graph visualizations for Llama-3.2-3B at $N = 500$. Representative constant and chaotic logistic map inputs are shown for $\kappa \in \{2, 4, 8\}$, with display threshold κ/N . The main visualizations use $\kappa = 2.5$. The qualitative structure remains stable across thresholds.

8B model exhibits behavior highly similar to that of the 3B model, preserving the broad trends and separation among input families across the three diagnostics. The 1B model, however, distinguishes the input families less clearly. In particular, its λ_2 and $\exp(\text{SE})$ results show less clear separation between the chaotic and high-period periodic inputs, while its HFER exhibits a less distinct complexity-ordered progression among the lower-period inputs. These results suggest that the smaller 1B model retains the broad context-dependent graph-spectral trends but less consistently resolves fine-grained differences between dynamical regimes.

Instruction-tuned model comparison. We repeat the experimental setup from Section 4.1 with



(a) Transformer layer 1



(b) Transformer layer 12

Figure 16: Layerwise attention-induced token graphs at $N = 500$ for the main numerical input families examined in Section 4.2. **(a)** Layer 1 provides an illustrative early-depth snapshot in which cross-input differentiation in graph layout is comparatively limited and the layouts retain broadly similar local organization. **(b)** Layer 12 provides an illustrative middle-depth snapshot in which cross-input differentiation is more pronounced, with graph organization tending to become less uniform as input dynamical complexity increases.

the instruction-tuned variants of the 3B model used in the main analysis and the 8B model evaluated in the preceding scale comparison: Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct. Llama-3 Instruct models undergo alignment-focused post-training to support helpful and safe instruction following (Grattafiori et al., 2024). We focus on the 3B and 8B models because the preceding scale comparison shows weaker input-family separation for the 1B base model. As shown in Fig. 17(c)–(d), both instruction-tuned models retain the principal context- and complexity-dependent behavior of their base counterparts, although some fine-grained distinctions between high-period periodic and chaotic inputs become less pronounced, particularly in exp(SE). This weakening is consistent with output-level evidence that alignment-related post-training can reduce extrapolation accuracy for time series (Gruver et al., 2023) and spatiotemporal dynamics (Bao et al., 2026).

Additional model families. Using the same experimental setup as Section 4.1, Fig. 18 presents the layer-averaged diagnostics for Phi-4 and the base and instruction-tuned SmoLM3-3B variants. At the time of our experiments, only the post-trained, chat-optimized Phi-4 checkpoint was publicly available; unlike Llama 3 and SmoLM3, Phi-4 had no corresponding publicly released base-model checkpoint. Across all three models, the broad context- and complexity-dependent behavior remains consistent with the Llama 3 results. SmoLM3-3B-Base closely reproduces the ordered input-family behavior of the 3B and 8B Llama base models, whereas Phi-4 and instruction-tuned SmoLM3-3B show less clear separation between high-period periodic and chaotic inputs, similar to the instruction-tuned Llama variants. The absolute diagnostic magnitudes and rates of change with context differ across models, indicating that the detailed graph-spectral signatures are model-dependent even though their overall ordering and trends remain broadly consistent.

H Attention-Only and Hidden-State-Only Non-Graph Baselines

We further show that the broad context- and complexity-dependent trends captured by the graph-spectral diagnostics are not artifacts of the graph-signal construction. Analogous summaries derived from attention alone or hidden states alone exhibit similar behavior as context length increases

and across input families, although they provide less fine-grained separation among families. As non-graph controls, we consider two hidden-state-only summaries based on fixed frequencies over serialized token position and one attention-only summary based on attention concentration. Following the notation of Section 3.2, let $A^{(\ell,h)} \in \mathbb{R}^{N \times N}$ denote the post-softmax attention matrix for layer ℓ and head h , and let $X^{(\ell)} \in \mathbb{R}^{N \times d}$ denote the hidden-state matrix entering layer ℓ . The positional controls use only $X^{(\ell)}$, whereas the attention control uses only $A^{(\ell,h)}$.

Normalized attention Gini coefficient. The Gini coefficient measures the concentration of a nonnegative vector (Hurley and Rickard, 2009). For a nonzero vector $\mathbf{z} \in \mathbb{R}_{\geq 0}^m$, define

$$g(\mathbf{z}) = \frac{\sum_{r=1}^m \sum_{s=1}^m |z_r - z_s|}{2m \sum_{r=1}^m z_r}.$$

This coefficient satisfies $g(\mathbf{z}) \in [0, 1 - 1/m]$: it equals 0 if and only if the entries of $\mathbf{z}$ are uniform, and reaches $1 - 1/m$ when all mass is concentrated on a single entry. For query position i and head h , let $\mathbf{a}_i^{(\ell,h)} = (A_{i1}^{(\ell,h)}, \dots, A_{ii}^{(\ell,h)})$ contain its causally available attention weights. We normalize by the maximum attainable for each row length and average across query positions and heads:

$$\text{Gini}^{(\ell)} = \frac{1}{H(N-1)} \sum_{h=1}^H \sum_{i=2}^N \frac{g(\mathbf{a}_i^{(\ell,h)})}{1 - 1/i}.$$

The normalization makes rows with different numbers of available keys comparable and gives $\text{Gini}^{(\ell)} \in [0, 1]$; the first position is omitted because it has only one available key. A value of 0 means that every included query distributes attention uniformly over its available keys, whereas a value of 1 means that every query concentrates all attention on a single key. Intermediate values quantify the average degree of attention concentration. In this sense, attention Gini is an attention-only control related to the normalized Fiedler value λ_2 from our graph-spectral diagnostics: greater rowwise concentration may accompany the more localized attention-graph structure indicated by smaller λ_2 , but Gini does not measure how those attention links combine into a globally connected graph.

Fixed positional-frequency representation. The two hidden-state controls measure how hidden states vary across token positions in the

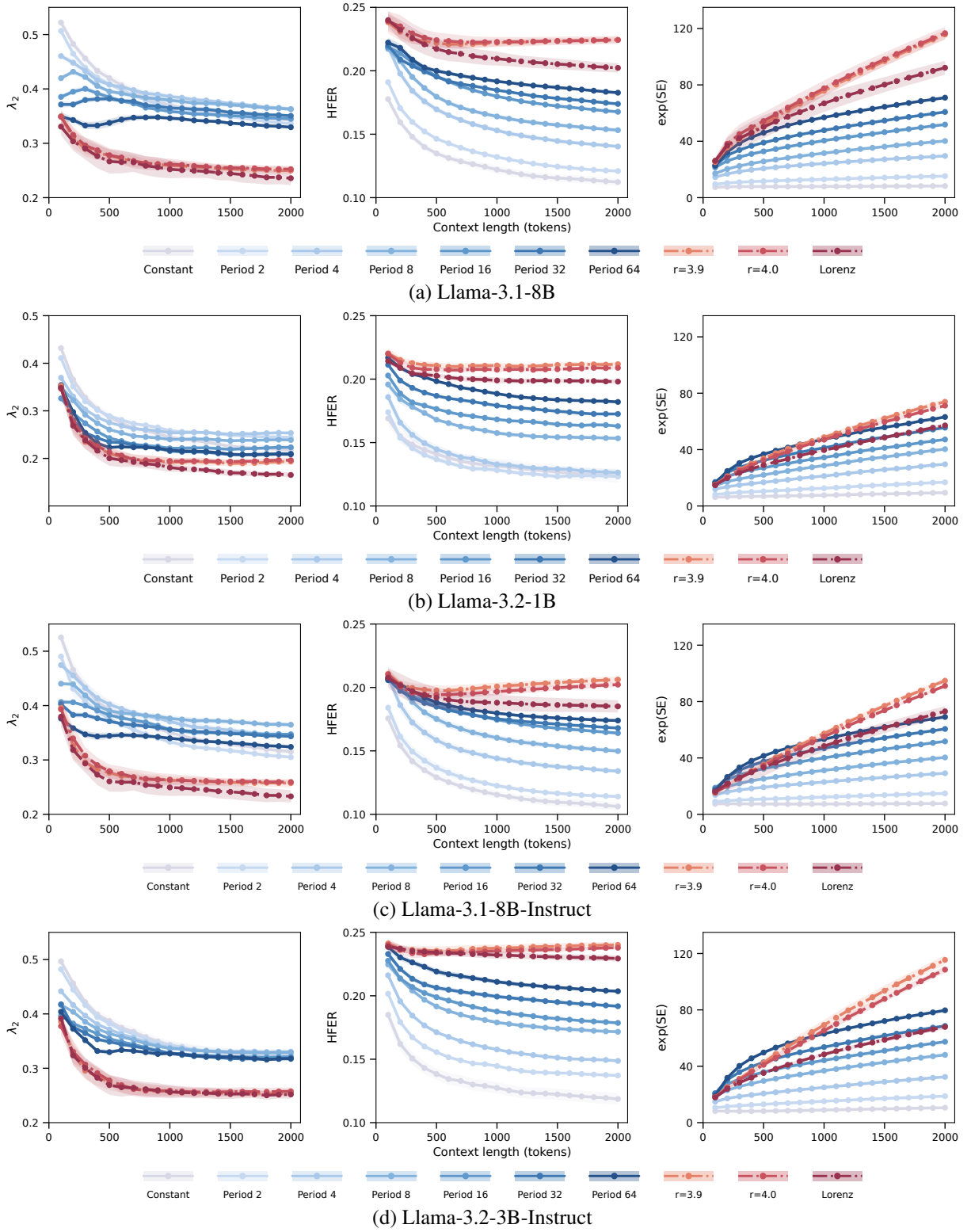


Figure 17: Layer-averaged graph-spectral diagnostics across base-model scales and instruction-tuned variants within the Llama 3 family, under the same experimental setup as Section 4.1. From left to right, the columns report normalized Fiedler value λ_2 , HFER, and $\exp(\text{SE})$ as functions of context length. Compared with the Llama-3.2-3B results in Fig. 3, the 8B base model exhibits highly similar behavior, whereas the 1B base model less clearly distinguishes the dynamical regimes. Both instruction-tuned models retain the broad behavior of their base counterparts, but some finer distinctions among input families become less pronounced with instruction tuning.

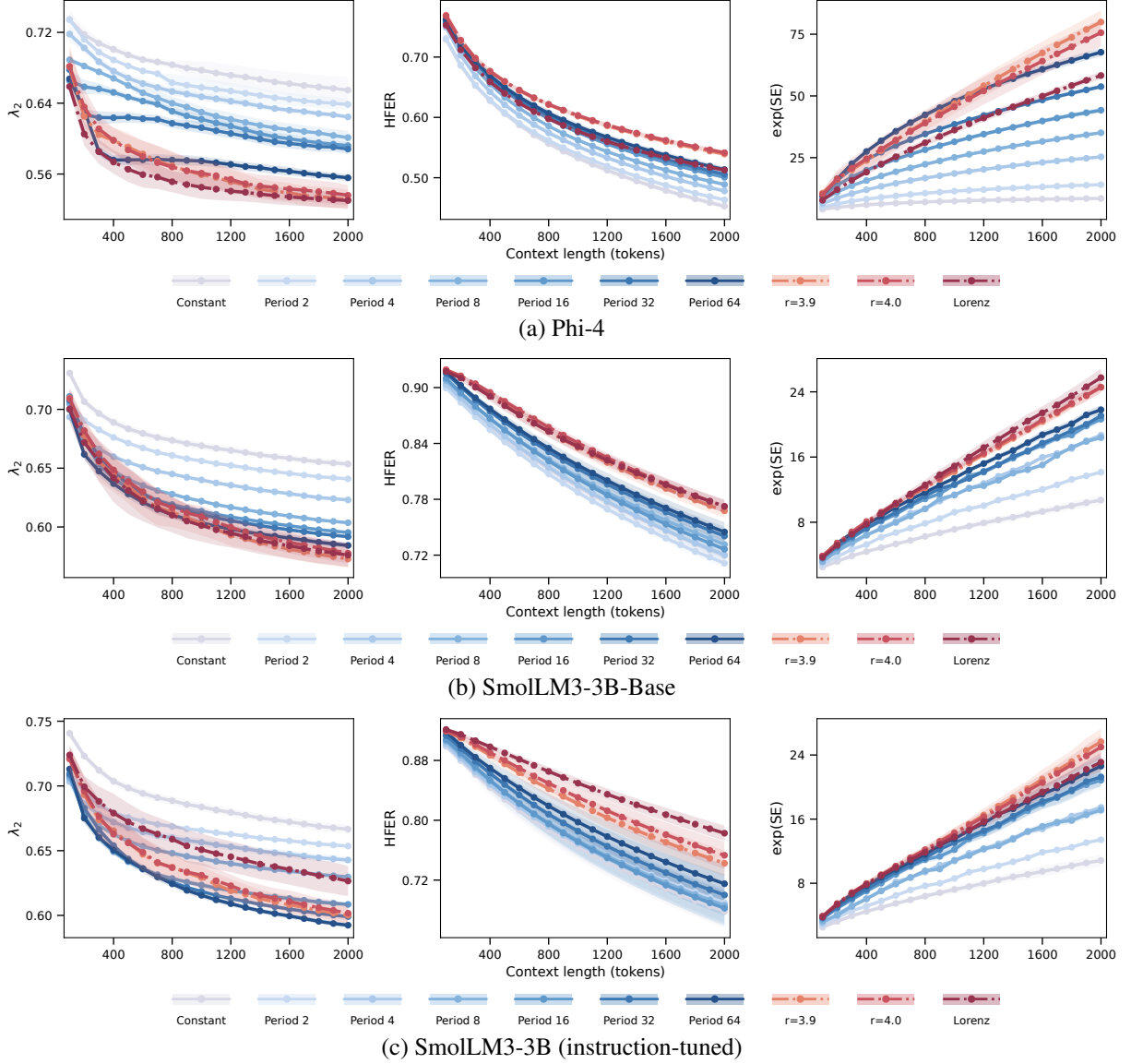


Figure 18: Layer-averaged graph-spectral diagnostics for additional model families under the same experimental setup as Section 4.1. From left to right, the columns report normalized Fiedler value λ_2 , HFER, and $\exp(\text{SE})$ as functions of context length. SmolLM3-3B-Base reproduces the broad context- and complexity-dependent behavior observed in the Llama 3 models. Phi-4, a post-trained chat model, and instruction-tuned SmolLM3-3B also retain the overall trends but exhibit less consistent fine-grained separation among high-period and chaotic inputs. Diagnostic magnitudes and rates of change with context length differ across models.

serialized input, without using attention. We construct this positional-frequency representation using an orthonormal type-II discrete cosine transform (DCT-II) (Ahmed et al., 1974). Unlike a standard discrete Fourier transform, the DCT-II does not impose periodic wraparound between the first and last token positions, while its orthonormality preserves total squared hidden-state energy. In implementation, we apply SciPy’s `scipy.fft.dct` with `type=2`, `norm="ortho"`, and `axis=0`, transforming the token dimension independently for each hidden-state coordinate. We write the resulting coefficients as

$$\tilde{X}_{\text{pos}}^{(\ell)} = \text{DCT}_{\text{II}}\left(X^{(\ell)}\right).$$

Let $\tilde{\mathbf{x}}_{\text{pos},k}^{(\ell)} \in \mathbb{R}^d$ denote the coefficient vector for positional mode k . These modes are ordered from slow to rapid variation in token hidden states across serialized positions, rather than by temporal frequency in the underlying numerical trajectory. The fixed DCT basis thus provides a hidden-state-only counterpart to the graph Fourier basis used in our main diagnostics: both decompose hidden states into frequency-ordered modes, but the DCT basis is determined only by context length and serialized token order, whereas the graph Fourier basis is determined by the attention-induced token geometry.

Positional HFER. Let $p_{\text{pos},k}^{(\ell)} = \|\tilde{\mathbf{x}}_{\text{pos},k}^{(\ell)}\|_2^2 / \sum_{r=1}^N \|\tilde{\mathbf{x}}_{\text{pos},r}^{(\ell)}\|_2^2$ denote the normalized positional-modal energy. Using the same default cutoff $K = \lfloor N/2 \rfloor$ as graph HFER, we define positional HFER as the fraction of hidden-state energy assigned to positional-frequency modes above the cutoff:

$$\text{HFER}_{\text{pos}}^{(\ell)} = \sum_{k=K+1}^N p_{\text{pos},k}^{(\ell)}.$$

This quantity lies in $[0, 1]$: it is 0 when all energy lies in modes at or below the cutoff and 1 when all energy lies in modes above the cutoff. Larger values indicate that a greater fraction of hidden-state energy is associated with rapid variation across serialized token positions. By comparison, graph HFER quantifies hidden-state variation over the attention-induced token geometry.

Positional spectral entropy (SE). Analogously to graph SE, positional SE measures how broadly hidden-state energy is distributed across the

positional-frequency modes. Using the normalized positional-modal energies $p_{\text{pos},k}^{(\ell)}$, we define

$$\text{SE}_{\text{pos}}^{(\ell)} = - \sum_{k=1}^N p_{\text{pos},k}^{(\ell)} \log p_{\text{pos},k}^{(\ell)}.$$

Following the main analysis, we report $\exp(\text{SE}_{\text{pos}}^{(\ell)})$ as the effective number of contributing positional modes. This quantity lies in $[1, N]$: it equals 1 when all energy is concentrated in one mode and N when energy is distributed uniformly across all modes. Larger values indicate that hidden-state energy is distributed across a broader set of positional-frequency modes. Thus, $\exp(\text{SE}_{\text{pos}}^{(\ell)})$ and graph-spectral $\exp(\text{SE}^{(\ell)})$ quantify analogous notions of spectral breadth, but in the fixed positional DCT basis and the attention-induced graph Fourier basis, respectively.

Comparison with graph-spectral diagnostics.

Under the same evaluation setup as Section 4.1, Fig. 19 compares the three non-graph controls, each computed at every transformer layer and then averaged across the L layers, with the main graph-spectral diagnostics across context lengths. These controls recover the broad context- and complexity-dependent trends across the main numerical input families, showing that the broad qualitative findings are also visible in attention- or hidden-state-only summaries and are therefore not artifacts of graph construction. When plotted with matched y-axis limits, positional HFER and $\exp(\text{SE}_{\text{pos}})$ exhibit narrower dynamic ranges and less separation among several input families. One notable difference between positional and graph HFER is that the former assigns the period-2 family greater high-frequency energy than periods 4–64 at every context length, reflecting the sensitivity of fixed positional modes to rapid alternation in serialized token order. Graph HFER instead measures hidden-state variation relative to the attention-induced token geometry, incorporating attention structure rather than relying on token order alone. The graph-spectral formulation therefore provides a joint characterization of attention-induced token connectivity and hidden-state variation and more consistently resolves the finer family- and context-specific behaviors discussed in Section 4.1.

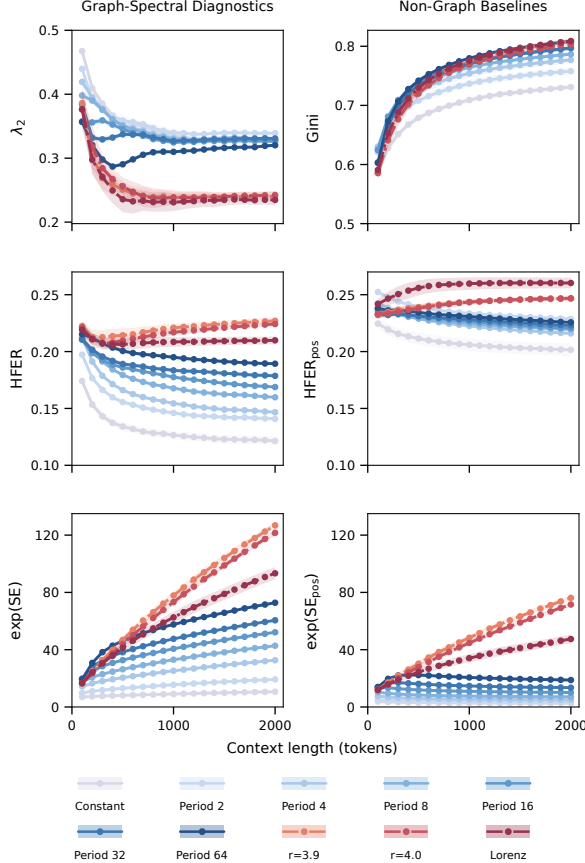


Figure 19: Comparison of graph-spectral diagnostics (left) and corresponding non-graph controls (right) across context lengths under the same evaluation setup as Section 4.1. The controls reproduce the broad context- and complexity-dependent trends. The top-row control is attention-only: increasing attention Gini indicates more concentrated attention. The lower two controls are hidden-state-only and measure high-frequency energy and effective spectral support in a fixed positional-frequency basis. The positional controls and corresponding graph-spectral diagnostics capture analogous spectral properties, but the latter show clearer separation among input families on matched y-axis scales, highlighting the added value of jointly analyzing hidden-state variation and attention-induced graph structure.

I Attention-Induced Token Graphs for Additional Structured Numerical Inputs

Beyond the main numerical input suite introduced in Section 3.1, we consider several additional structured numerical sequences representing abrupt distributional shifts, smooth periodic variation, and stochastic evolution. Each full sequence contains 500 numeric tokens and 500 separator tokens, yielding a 1000-token reference context. Fig. 20 displays the corresponding last-layer attention-induced token graphs at 10%, 30%, 50%, 70%, and 90% of the reference context, using the same graph construction and visualization procedure described in Section 3.3.

The Gaussian-switch inputs are piecewise Gaussian sequences with i.i.d. samples within each segment and abrupt changes in mean at prescribed context fractions. This construction adapts the distribution-switching setup used by Sarfati et al. (2026) to study belief dynamics. In the one-switch case, samples are drawn from $\mathcal{N}(\mu = 300, \sigma^2 = 10)$ over the first half of the sequence and from $\mathcal{N}(\mu = 700, \sigma^2 = 10)$ over the second half. In the two-switch case, the mean is 300 over the first 35% of the sequence, 700 over the middle 30%, and 300 over the final 35%, with $\sigma^2 = 10$ throughout. Each switch therefore has magnitude $\Delta_\mu = 400$, which is large relative to the within-regime standard deviation $\sigma = \sqrt{10}$. The sampled values are rounded to the nearest integer and clipped to $[150, 850]$ before serialization.

We also include a smooth periodic trajectory generated from $\sin(x)$, sampled uniformly over three full periods and rescaled to the same integer range. Unlike the periodic logistic map orbits examined in the main text, whose successive iterates can move sharply between distinct orbit values, this sequence varies smoothly from one step to the next.

Finally, we include two bounded stochastic processes: a random walk and a random walk with drift. Both follow

$$z_{t+1} = R_{[150, 850]}(z_t + \mu + \epsilon_t), \quad \epsilon_t \sim \mathcal{N}(0, 18^2),$$

with $\mu = 0$ for the random walk and $\mu = 2.5$ for the drifted random walk. The zero-drift and drifted walks are initialized at $z_0 = 500$ and $z_0 = 350$, respectively. The reflection operator $R_{[150, 850]}$ maps boundary overshoots back into $[150, 850]$, repeating the reflection if necessary.

Fig. 20 suggests that attention-induced graphs reflect not only deterministic dynamical complex-

ity, but also distributional shifts, smooth recurrence, and stochastic evolution. In the Gaussian-switch examples, the switch-defined temporal segments appear as distinct clusters in the attention-induced graphs. This observation complements Sarfati et al. (2026), who show that abrupt shifts in the mean of a Gaussian input distribution produce structured belief-update trajectories in hidden-state space, accompanied by corresponding changes in the model’s predictive distributions over numerical outputs. Our attention-induced token-graph view provides a complementary perspective by showing that change-point structure is also reflected in attention topology as clusters aligned with the switch-defined temporal segments. The sine input produces a graph structure distinct from periodic logistic map orbits despite their shared periodicity. By contrast, the random-walk inputs produce less sharply organized graph structures, consistent with the path-dependent nature of their generating processes.

J Preliminary Graph-Spectral Comparison of Numerical, Natural-Language, and Code Inputs

Shared setup. We construct three input suites—multilingual text, source code, and English-domain text—each comprising six input types and 50 samples per type. Mirroring the numerical-input evaluation in Section 4.1, we compute the same graph-spectral diagnostics at 20 context lengths from 100 to 2000 tokens in increments of 100, using the corresponding prefixes of each 2000-token sample.

Multilingual text. The multilingual suite contains text in six languages—Arabic, Chinese, English, French, Russian, and Spanish—from MultiUN, a parallel corpus of translated United Nations documents (Eisele and Chen, 2010; Tiedemann, 2012). Using English as the pivot, we align corresponding records across the six languages and concatenate consecutive aligned record sets until all six language streams contain at least 2000 tokens; we then truncate each stream to exactly 2000 tokens. The samples therefore contain aligned source material, although differences in tokenization density mean that the cutoff may fall at different positions within the final aligned record.

Code inputs. The programming-language suite comprises Python, JavaScript, Java, C++, Go, and Rust files from the StarCoder training dataset (Ko-

cetkov et al., 2023; Li et al., 2023). For each language, we sample 50 contiguous 2000-token segments at random offsets from distinct source files. We remove encoding artifacts and invalid control characters while preserving whitespace, line breaks, indentation, comments, and other code formatting.

English-domain text. The English-domain suite comprises news from CNN/DailyMail (Hermann et al., 2015; See et al., 2017), fiction from Project Gutenberg (50 distinct books, one per sample) (Christou and Tsoumakas, 2025), English Wikipedia articles from a November 2023 snapshot (Wikimedia Foundation, 2023), scientific text from arXiv (Cohan et al., 2018), dialogue from DailyDialog (Li et al., 2017), and legal text from LexGLUE (Chalkidis et al., 2022). We apply light common preprocessing by decoding HTML entities, removing generic boilerplate-only lines, flattening internal line breaks, and collapsing repeated whitespace. When a dataset entry is shorter than 2000 tokens, we concatenate consecutive entries until reaching the target length.

Findings and open questions. Fig. 21 reveals a consistent separation of the natural-language and code inputs from the controlled numerical inputs across all three diagnostics. Relative to the numerical inputs, the natural-language and code inputs exhibit lower normalized Fiedler values, indicating less globally integrated and more localized attention topology, together with higher HFER, indicating sharper hidden-state variation over that topology. The separation is especially pronounced in effective spectral support, with $\exp(\text{SE})$ remaining substantially larger for the natural-language and code families than for the numerical families across all evaluated context lengths. This difference indicates that their hidden-state energy is distributed across a substantially larger effective number of graph Fourier modes. These results support a coarse representational distinction between controlled numerical sequences and natural-language or code inputs. More controlled comparisons are needed to determine which input and model properties contribute to the coarse separation of natural-language and code inputs from numerical inputs and whether meaningful graph-spectral orderings exist within the multilingual-text, source-code, and English-domain suites.

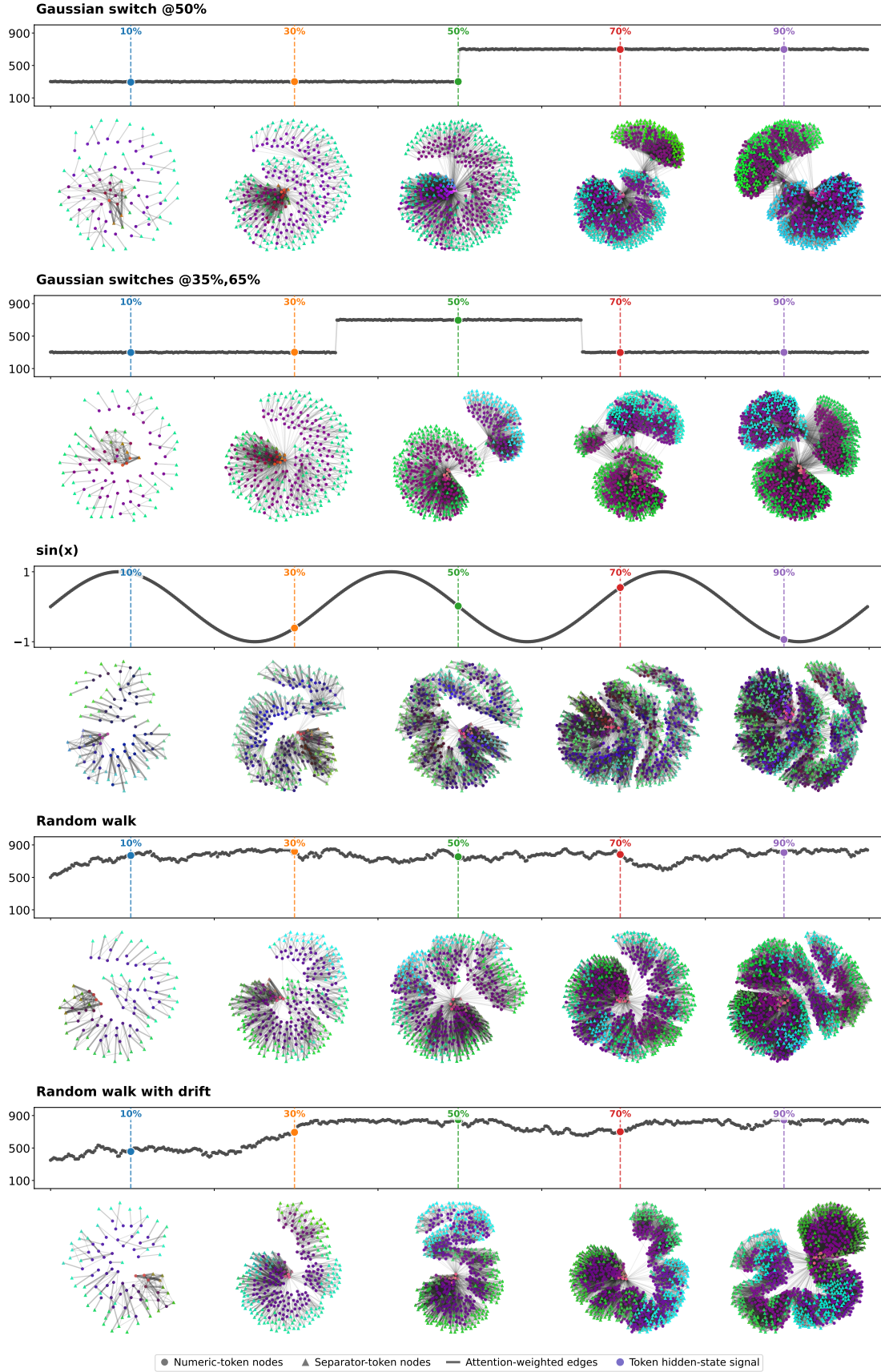
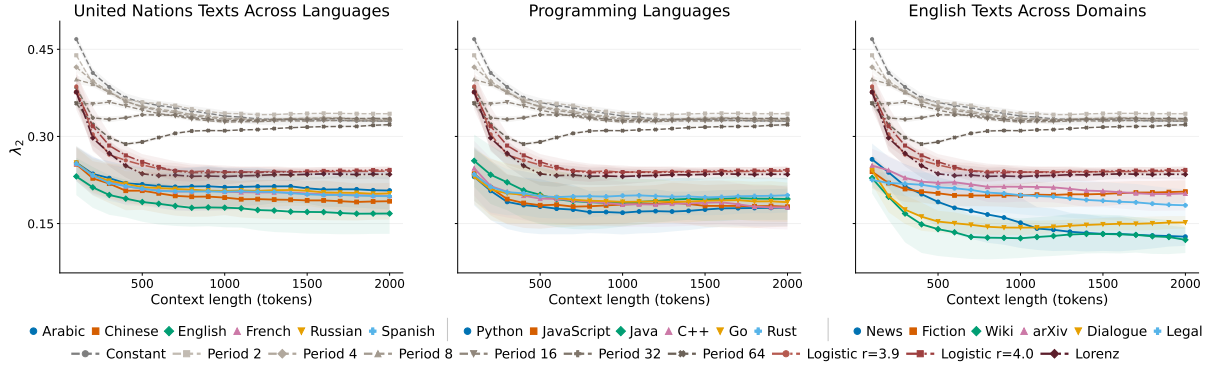
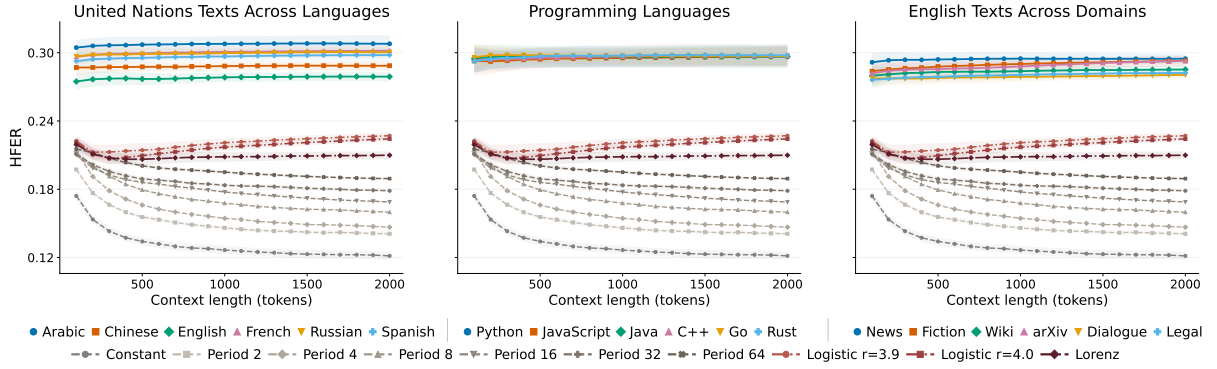


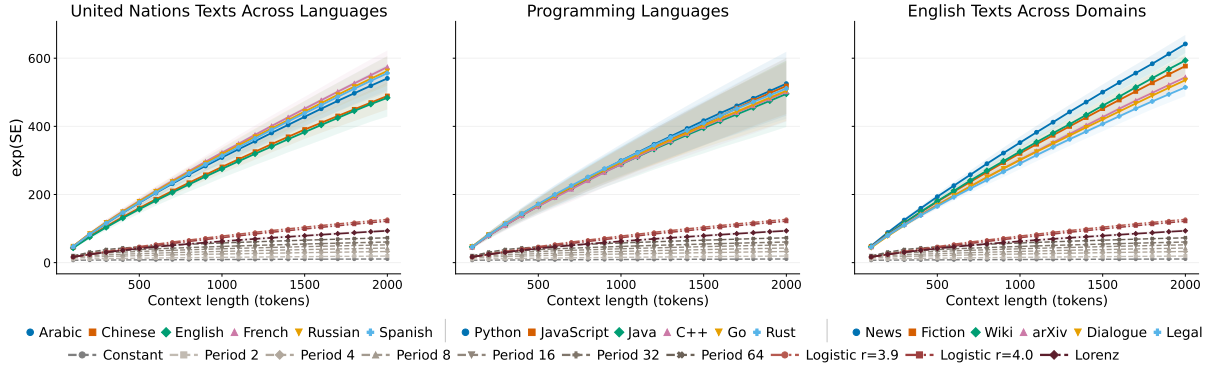
Figure 20: Last-layer attention-induced token graphs for additional structured numerical inputs with Llama-3.2-3B. Graphs are shown at 10%, 30%, 50%, 70%, and 90% of a 1000-token reference context. All examples use the same attention-induced token-graph construction and visualization procedure as Section 3.3. The visualizations illustrate qualitative graph-organization patterns beyond those observed for the main dynamical-system trajectories.



(a) Normalized Fiedler value λ_2



(b) High-frequency energy ratio (HFER)



(c) Effective spectral support $\exp(\text{SE})$

Figure 21: Layer-averaged graph-spectral diagnostics for natural-language and code inputs using Llama-3.2-3B, with the main numerical families shown as dashed matched-context references. Columns compare content-aligned United Nations translations across six languages, programming-language inputs, and English-domain inputs. Curves and shading denote the mean and ± 1 standard deviation across 50 samples per natural-language or code family and 20 realizations per numerical family. Across context lengths, the natural-language and code family means exhibit lower normalized Fiedler values, higher HFER, and broader effective spectral support than the numerical inputs; this coarse separation corresponds to weaker global attention connectivity, a larger high-frequency share of hidden-state energy, and a broader distribution of that energy across graph Fourier modes.