PARITY ECHOES IN A TRANSFORMER: PREDICTABLE FAR LENGTH RE-EMERGENCE AFTER SINGLE LENGTH TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how a Transformer trained to solve parity at a single sequence length behaves when evaluated on other lengths. Surprisingly, failure is not monotonic: the model can lose measurable parity signal at nearby unseen lengths and partially recover it much farther away in a regular pattern. We show that this behavior is closely related to how the model represents its inputs. Successful models organize sequences mainly by the number of ones they contain, with different counts arranged along an approximately one-dimensional structure. This representation makes the model's behavior at unseen lengths predictable: we derive where parity signal should reappear and how its strength changes with sequence length, and find close agreement with trained models. We also study how the difficulty of learning parity grows with sequence length. Empirically, the number of optimizer steps required to learn the task scales close to $L^{3/2}$, and a simplified spectral model yields the same scaling. These results connect two aspects of parity learning: how a trained model behaves away from its training length, and how the difficulty of learning the task grows as that length increases

1 INTRODUCTION

For AI safety, poor performance in a capability evaluation does not necessarily mean that the capability is absent: models can retain capabilities that remain latent under one evaluation condition and become accessible under another (van der Weij et al., 2025; Hofstätter et al., 2025). We study a simple setting where a related question can be examined under distribution shift. A causal Transformer trained from scratch on parity at one sequence length can lose measurable parity signal at nearby unseen lengths and recover it much farther away in a structured pattern. We ask why this happens and whether the lengths at which the signal returns can be predicted.

Our explanation starts from the model's internal representation. Successful models organize their post-attention states primarily by the number of ones in the input, known as its Hamming weight. Different Hamming weights are arranged along an approximately one-dimensional geometry. This structure makes the model's behavior at unseen lengths analytically tractable: we derive quantitative predictions for where parity signal should reappear and how strong it should be, and observe these predicted distant re-emergence events, which we call parity echoes, in the trained Transformer.

The same representation also provides a way to study how the difficulty of learning parity grows with sequence length. Empirically, the number of optimizer updates required to learn the task grows close to $L^{3/2}$. In a controlled lazy-learning model, we derive the same scaling from the parity-relevant kernel spectrum, providing a theoretical account of the observed training-time law.

Prior work does not, to our knowledge, report the predictable far-length re-emergence studied here. In standard direct-supervision settings, Transformers generally fail to extrapolate parity to longer sequences; strong length generalization has instead been obtained after introducing specially designed mechanisms such as recurrent working memory, scratchpad or intermediate-state supervision, or other task-specific architectural and training changes (Anil et al., 2022; Chi et al., 2023; Liu et al., 2023). Counting studies show that Transformers can encode token counts and organize them in low-dimensional representations, but do not connect that geometry to a quantitative prediction of parity behavior across unseen lengths (Behrens et al., 2025; Gurnee et al., 2026). Likewise, Transformer kernel analyses derive infinite-width spectra, context-length learnability, and spectral biases, but do

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

not identify the parity-carrying modes induced by a Hamming-weight representation or use them to derive an optimizer-time law in sequence length (Hron et al., 2020; Lavie et al., 2024; Vasudeva et al., 2025). The gap we address is therefore the structured far-length behavior of a standard single-length parity model, together with an analytical account of where the signal reappears and a separate spectral explanation for the observed $L^{3/2}$ learning-time scaling.

2 Experiments on the Causal Transformer

We first establish the two empirical phenomena studied in the paper. A Transformer trained only at $L = 90$ loses parity signal at nearby unseen lengths but recovers it near distant resonances, while a separate sweep over training lengths shows an optimizer-time scaling close to $L^{3/2}$. We then test the representation-level mechanism used in Section 3 to explain the length-dependent pattern.

2.1 Task, architecture, and signal measure

The task is binary parity: given a sequence of bits, the model predicts whether the number of ones is even or odd. Each sequence is followed by a separator token SEP, and the prediction is read from that final position.

The model is a single causal Transformer block. For a token embedding $\vec{e}_i \in \mathbb{R}^{d_m}$ at position i, one causal block computes

$$\begin{aligned} \vec{q}_i^{(h)} &= R_i W_Q^{(h)} \vec{e}_i, & \vec{k}_i^{(h)} &= R_i W_K^{(h)} \vec{e}_i, \quad \vec{v}_i^{(h)} = W_V^{(h)} \vec{e}_i, \\ \alpha_{ij}^{(h)} &= \mathrm{softmax}_{j \le i}\left(\frac{\vec{q}_i^{(h)\top} \vec{k}_j^{(h)}}{\sqrt{d_k}} \right), \\ \vec{z}_i &= \sum_{h=1}^{H} W_O^{(h)} \left(\sum_{j \le i} \alpha_{ij}^{(h)} \vec{v}_j^{(h)} \right), \\ \vec{h}_i &= W_5 \,\mathrm{ReLU}(W_4 \vec{z}_i), & \hat{y}_i &= W_d \vec{h}_i. \end{aligned} \tag{1}$$

Here R_i is the RoPE rotation (Su et al., 2024). The head outputs are first mapped back to model space by $W_O^{(h)}$ and are then summed across heads; there is no concatenation. The relevant dimensions are

$$\begin{gathered} W_Q^{(h)}, W_K^{(h)}, W_V^{(h)} \in \mathbb{R}^{d_k \times d_m}, \qquad W_O^{(h)} \in \mathbb{R}^{d_m \times d_k}, \\ W_4 \in \mathbb{R}^{d_{ff} \times d_m}, \qquad W_5 \in \mathbb{R}^{d_m \times d_{ff}}. \end{gathered} \tag{2}$$

Vectors are denoted by arrows, while scalars and matrices are left unmarked. We use $d_m = 64$, $H = 3$, and $d_k = 20$, with $\vec{z}_i \in \mathbb{R}^{d_m}$ and feed-forward width d_{ff}. Trainable weights use LeCun-normal initialization with standard deviation $1/\sqrt{\text{fan-in}}$, and all learned linear maps are bias-free. There are no residual connections, LayerNorm, or dropout. AdamW uses learning rate 5×10^{-4} and weight decay 1; $d_{ff} = 500$ for both the length-re-emergence experiment and the convergence sweep. For the $L = 90$ experiment, training stops when the training MSE first reaches 0.3. For run r evaluated at test length x, let $\mathrm{Acc}_r(x)$ be its classification accuracy and define

$$s_r(x) = 2\,\mathrm{Acc}_r(x) - 1, \qquad E_{\mathrm{emp}}(x) = \mathbb{E}_r\left[s_r(x)^2\right]. \tag{3}$$

Squaring makes parity and anti-parity contribute equally to the measured signal.

2.2 Single-length training reveals distant re-emergence

We train 30 random initializations at $L = 90$, of which 26 reach the stopping criterion, and evaluate every integer length $x = 2, \ldots, 550$. The clean comparison is the strict noiseless classifier derived in Section 3, averaged over the discrete Hamming-weight distribution:

$$\mathrm{Acc}_0(x) = \sum_{X=0}^{x} 2^{-x} \binom{x}{X} \mathbf{1}\left\{ \cos\left(\pi \frac{L-x}{x} X \right) > 0 \right\}, \qquad E_0(x) = \left(2\,\mathrm{Acc}_0(x) - 1\right)^2. \tag{4}$$

Because training is deliberately stopped before perfect interpolation, we scale the ideal curve once at the training length: $\widetilde{E}_0(x) = C E_0(x)$ with $C = E_{\mathrm{emp}}(L) = 0.735742$. There is no length-dependent fit.

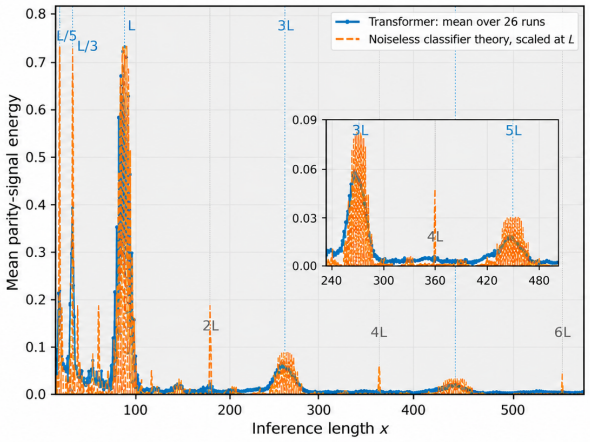

Figure 1: Parity-signal energy after training only at $L = 90$. Blue is the Transformer mean over 26 successful runs; orange is the noiseless classifier theory in Eq. (4), scaled once at L. Primary guides mark $L/5, L/3, L, 3L, 5L$; lighter guides mark selected $a = 1$ zero-margin contacts $2L, 4L, 6L$. More generally, zero-margin contacts occur whenever the reduced numerator b of $x/L = b/a$ is even. The main panel begins at $x = 15$ for readability; the poorly matched $x < 15$ regime is quantified separately in Appendix A.7.

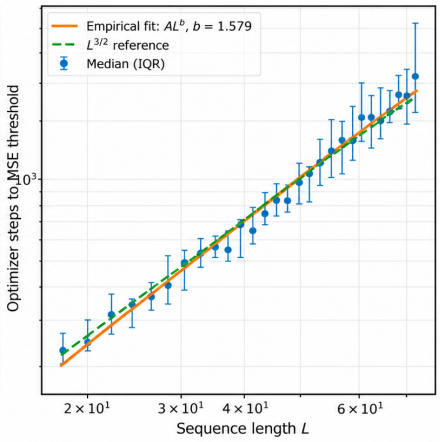

Figure 2: Optimizer updates required to reach training MSE < 0.2. Points show medians and interquartile ranges; the empirical fit has exponent 1.579, while the $L^{3/2}$ reference fits only its vertical prefactor.

Figure 1 compares the Transformer mean over 26 successful runs with the noiseless classifier theory scaled once at L. The blue guides mark the finite-margin family $L/5, L/3, L, 3L, 5L$, while the lighter gray guides mark selected $a = 1$ zero-margin contacts $2L, 4L, 6L$. More generally, contacts occur whenever the reduced numerator b of $x/L = b/a$ is even.

2.3 Optimizer time grows close to $L^{3/2}$

We train independent models at even lengths $L = 18, 20, \ldots, 68$, using 15 attempts per length and recording the first optimizer step at which training MSE falls below 0.2; more than 97% of attempts converge. Figure 2 reports the median over converged runs with the interquartile range. A log–log fit gives

$$T_{\text{med}}(L) = AL^b, \qquad b = 1.579, \qquad R^2 = 0.989, \tag{5}$$

close to the $L^{3/2}$ reference. Section 4 derives the $3/2$ exponent from a controlled frozen-feature spectrum; optimization and batch-schedule details are given in Appendix A.1.

2.4 Representation geometry and quantitative agreement

To test the mechanism behind the length-dependent curve, we examine the post-attention SEP representation before the feed-forward network. At each test length, examples with the same Hamming weight X form a *shell*; their mean representation is the shell center. We use three diagnostics: R_{collapse} measures within-shell spread relative to neighboring-shell separation, R^2_{line} measures how nearly the shell centers are collinear, and $\text{CV}_{\text{spacing}}$ measures the nonuniformity of adjacent gaps. Their ideal values are 0, 1, and 0, respectively; exact definitions and sampling details are in Appendix A.3–A.4.

The main geometric result is that training sharpens a pre-existing Hamming-weight organization. At $x = L = 90$, initialization already gives $R^2_{\text{line}} = 0.99176$, but the shells overlap strongly ($R_{\text{collapse}} = 1.705$) and their spacing is irregular ($\text{CV}_{\text{spacing}} = 0.435$). After training, $R_{\text{collapse}} = 0.0039$, $R^2_{\text{line}} = 0.999999965$, and $\text{CV}_{\text{spacing}} = 5.89 \times 10^{-4}$. The trained geometry also remains highly structured far beyond L; at $x = 550 \simeq 6.1L$, for example, $R_{\text{collapse}} = 0.0114$ and $R^2_{\text{line}} = 0.999999960$. Figure 3 gives a direct visualization for one successful run: the broad, Hamming-weight-ordered cloud present at initialization collapses during training onto an effectively one-dimensional sequence of Hamming-weight shells.

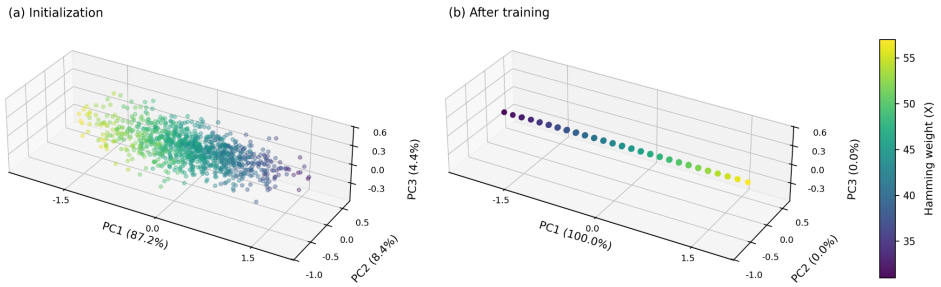

Figure 3: PCA visualization of the post-attention separator representation before and after training. Shown are the top three principal components of the d_m-dimensional post-attention separator representation $\vec{z}_{\mathrm{SEP}}$ for 1,000 test examples from one successful run trained at $L = 90$, colored by Hamming weight X. At initialization, Hamming weight already induces a dominant low-dimensional trend, while examples with the same Hamming weight retain substantial within-shell spread. After training, each visible point in the right panel corresponds effectively to one Hamming-weight shell, showing near-complete within-shell collapse and an almost perfectly one-dimensional organization.

To connect this geometry to the length-dependent classifier, we assign each test representation a scalar position along the affine Hamming-weight line fitted at L. Write this line as

$$\vec{\ell}(t) = \vec{a} + t\vec{v},$$

with the scale chosen so that the center of training shell X lies near $t = X$. Thus t measures position along the line in units of one training Hamming-weight step. For a test sequence of length x, Hamming weight X, and bit arrangement ω, let $\vec{z}^{(x)}_{\mathrm{SEP}}(X, \omega)$ denote its post-attention representation. We define its coordinate along the fitted line and its deviation from the ideal cross-length location by

$$u(x, X, \omega) = \frac{\vec{v}^{\top}\left(\vec{z}^{(x)}_{\mathrm{SEP}}(X, \omega) - \vec{a}\right)}{\|\vec{v}\|^2}, \qquad \delta u(x, X, \omega) = u(x, X, \omega) - \frac{LX}{x}. \tag{6}$$

The reference value LX/x corresponds to preserving normalized Hamming weight X/x when mapping a test example onto the training-length line.

These coordinate deviations remain small in shell units. Across successful runs, the median within-run standard deviation $\sigma_{\delta u}$ is 0.0016 at $x = L$ and about 0.020 at $x = 3L$. The mean displacement is also strongly correlated across test lengths within a run: $\mathrm{corr}[\mu_{\delta u}(270), \mu_{\delta u}(550)] = 0.980$. Section 3 shows that these small displacements are sufficient to modify the fine structure of the ideal length-dependent classifier, while leaving its broad distant resonances intact. Full definitions and additional diagnostics are given in Appendix A.5.

The width of those resonances provides a second quantitative test. Gaussian-envelope fits around $x_0 \in \{L, 3L, 5L\}$ give

$$W = 1.356\sqrt{x_0}, \qquad R^2 = 0.998, \tag{7}$$

with the same square-root scaling under a half-prominence estimator. Applying the same smoothing and Gaussian-plus-background width estimator to the full noiseless-classifier curve gives widths $12.70, 19.74, 27.61$, close to the measured $12.95, 21.89, 29.05$. Thus the clean theory reproduces the observed packet widths quantitatively. Section 3 derives the analytic scaling $W \propto \sqrt{x_0}/a$; the individual width measurements are reported in Appendix A.6.

3 From the SEP Geometry to the Energy-by-Length Law

Section 2 showed that the post-attention representations are already strongly organized by Hamming weight at initialization, and that training sharpens this tendency into tightly collapsed, nearly equally spaced shell centers that remain structured across unseen lengths. Here we show how this geometry produces the length-dependent parity signal in Figure 1. We first explain why the attention representation has a one-dimensional Hamming-weight tendency before training and why its central shell centers are approximately affine. We then show that an equally spaced Hamming-weight line,

together with the ReLU feed-forward network, gives a discrete cross-length classifier from which the locations and strengths of the parity echoes follow. Finally, a Fourier representation describes the widths of the echo packets and shows how the measured phase deviations from Section 2 modify the ideal prediction. The separate question of why learning parity takes approximately $L^{3/2}$ optimizer steps is addressed in Section 4.

3.1 RoPE scatter and locally affine Hamming-weight centers

Section 2 revealed three geometric features of the post-attention representation: examples with the same Hamming weight form shells, the shell centers lie close to a one-dimensional trajectory, and the central shell centers are nearly equally spaced. These features have different origins in the attention map, which we describe in turn.

First, consider the spread within a fixed Hamming-weight shell. For an input of length x and Hamming weight X, different bit arrangements place the same X ones at different RoPE positions and therefore receive slightly different attention weights. Let ω denote the arrangement. We write

$$\vec{z}^{(x)}_{\mathrm{SEP}}(X,\omega) = \vec{\mu}^{(x)}_X + \vec{\epsilon}^{(x)}_X(\omega), \tag{8}$$

where $\vec{\mu}^{(x)}_X$ is the shell center and $\vec{\epsilon}^{(x)}_X$ is the arrangement-dependent displacement around it. A large-x expansion shows that the typical within-shell displacement is $O(x^{-1/2})$; the full covariance calculation is given in Appendix B.2. Thus RoPE naturally produces a cloud of representations around each Hamming-weight-dependent center rather than destroying the Hamming-weight organization altogether. This is the architectural effect measured by R_{collapse} in Section 2.

Second, the shell centers themselves have a strong one-dimensional tendency. Writing $\rho = X/x$, their leading form is

$$\vec{\mu}^{(x)}_X = \sum_h \left[\vec{v}_{0,h} + \bar{A}_h(\rho;x)(\vec{v}_{1,h} - \vec{v}_{0,h})\right] + O(x^{-1}), \tag{9}$$

where $\vec{v}_{b,h} = W^{(h)}_O W^{(h)}_V \vec{e}_b$ and $\bar{A}_h(\rho;x)$ is a smooth scalar, at fixed sequence length x, determined by the average attention weights of head h. For a single head, varying ρ therefore moves the center along one fixed direction $\vec{v}_{1,h} - \vec{v}_{0,h}$. With several heads, different directions and different functions $\bar{A}_h(\rho;x)$ can bend the summed trajectory, but it remains smooth in normalized Hamming weight. Binomial concentration then restricts the initialization geometry to a locally affine central segment, explaining the high R^2_{line}.

The final question is why the part of this curve seen during training is not only nearly straight but also nearly equally spaced. At $x = L$, the binomial training distribution concentrates on

$$X - \frac{L}{2} = O(\sqrt{L}), \qquad \tau \equiv \frac{X - L/2}{L} = O(L^{-1/2}).$$

Because the shell center in Eq. (9) depends smoothly on the normalized Hamming weight X/L, we can expand it directly around $X = L/2$:

$$\begin{aligned} \vec{\mu}^{(L)}_X &= \vec{\mu}^{(L)}_{L/2} + \tau\vec{A} + \frac{1}{2}\tau^2\vec{B} + O(\tau^3) + O(L^{-1}), \\ \vec{A} &= \sum_h \bar{A}'_h\left(\frac{1}{2}\right)(\vec{v}_{1,h} - \vec{v}_{0,h}), \qquad \vec{B} = \sum_h \bar{A}''_h\left(\frac{1}{2}\right)(\vec{v}_{1,h} - \vec{v}_{0,h}). \end{aligned} \tag{10}$$

The linear displacement across the relevant window is $O(L^{-1/2})$, whereas the quadratic bending is only $O(L^{-1})$. Increasing X by one changes τ by $1/L$, so neighboring shell centers differ by $\vec{A}/L + \tau\vec{B}/L + O(L^{-2})$; across the central window, the variation of these gaps around $\vec{A}/L$ is therefore $O(L^{-3/2})$. The central shell centers therefore have an asymptotic tendency toward local straightness and regular spacing. These are the two properties measured by R^2_{line} and $\mathrm{CV}_{\mathrm{spacing}}$ in Section 2.

The attention geometry therefore supplies a strong Hamming-weight organization already at initialization: RoPE produces arrangement-dependent scatter around smooth Hamming-weight centers, while binomial concentration makes the central part of the center trajectory approximately affine and equally spaced. The empirical role of training is to sharpen this pre-existing structure dramatically, producing the nearly collapsed Hamming-weight line used in the remainder of the analysis.

3.2 From the Hamming-weight line to an idealized classifier

Given the approximately affine, equally spaced Hamming-weight geometry of Section 2, the shell centers can be described by the one-dimensional coordinate $u \in [0, L]$, with Hamming weight k lying near $u = k$. A two-layer ReLU network of width $L + 1$ can interpolate the alternating parity labels at these $L + 1$ locations with a piecewise-linear zigzag in u (Appendix B.1). Its decision boundaries lie halfway between consecutive integers, so the resulting classifier can be written as

$$\widehat{y}_0(u) = \operatorname{sgn} \cos(\pi u). \tag{11}$$

We use this midpoint construction as an idealized reference for the learned classifier.

At unseen lengths, we separately adopt the ideal fraction-preserving cross-length coordinate introduced in Section 2.4,

$$u_0(x, X) = \frac{LX}{x}. \tag{12}$$

Thus a test shell with Hamming weight X at length x is placed at the position corresponding to the same fraction of ones on the training-length coordinate. Evaluating the fixed classifier at u_0 and comparing its prediction with the true parity label gives the clean accuracy law in Eq. (4). This finite sum is the orange noiseless-classifier theory curve in Figure 1.

3.3 Residue classes give the arithmetic echo law

Equation (4) turns the distant re-emergence problem into arithmetic on Hamming-weight residues. Let $x/L = b/a$ in lowest terms. When a and b are both odd, the classifier depends only on X modulo b; at large length those residues are asymptotically uniform. Counting the positive residue classes gives

$$s_0\left(\frac{b}{a}L\right) \longrightarrow \frac{(-1)^{(b-1)/2}}{b}, \qquad E_0\left(\frac{b}{a}L\right) \longrightarrow \frac{1}{b^2}. \tag{13}$$

Thus, within the odd/odd branch, varying odd a moves the resonance to $x = (b/a)L$, while b fixes its asymptotic sign and strength. The $b = 1$ family contains exact shorter peaks such as $L/3$ and $L/5$; the prominent distant packets at $3L$ and $5L$ have $b = 3$ and $b = 5$. The other parity classes of the reduced ratio are classified next; the residue-counting proof is in Appendix B.3.

3.4 The remaining rational ratios: zero-margin contacts and silent cases

For the reduced ratio $x/L = b/a$, the complete asymptotic classification is

$$s_0\left(\frac{b}{a}L\right) \longrightarrow \begin{cases} (-1)^{(b-1)/2}/b, & a, b \text{ odd}, \\ -1/b, & b \text{ even}, \\ 0, & a \text{ even}, \end{cases} \qquad E_0 = s_0^2. \tag{14}$$

When b is even, asymptotic boundary mass $1/b$ lands exactly on the midpoint decision boundaries, producing a zero-margin contact under the strict > 0 convention. Thus $2L, 4L, 6L, \ldots$ are only the $a = 1$ members of a broader even-b family; for $L = 90$, for example, $x = 20$ has $x/L = 2/9$ and $x = 40$ has $x/L = 4/9$. These contacts are singular: arbitrarily small continuous phase perturbations split the boundary mass between the two classifier regions. When a is even the rational ratio is asymptotically silent. Moreover, for even training length L every odd test length is exactly silent in the ideal theory,

$$L \text{ even}, \ x \text{ odd} \quad \Longrightarrow \quad A_0(x) = \frac{1}{2}, \qquad s_0(x) = E_0(x) = 0, \tag{15}$$

which explains the alternating nearby-length comb. Details are in Appendix B.3.

3.5 Fourier representation: packet widths and phase sensitivity

Expanding the square-wave classifier in odd Fourier harmonics (Appendix B.3) gives a convenient description of the local packet shape. Near a finite-margin resonance $x_0 = (b/a)L$, with $x = x_0 + \Delta$, the resonant harmonic gives the energy envelope

$$E_{\text{env}}(x) \propto \exp\left[-\frac{a^2\pi^2\Delta^2}{4x_0}\right], \qquad W_{\text{th}} = \frac{4\sqrt{\ln 2}}{\pi}\frac{\sqrt{x_0}}{a}. \tag{16}$$

Thus the analytic envelope predicts the characteristic $\sqrt{x_0}/a$ packet-width scaling. The complete derivation is in Appendix B.3.

3.6 Measured phase deformation regularizes contacts and damps robust echoes

Finally, we ask whether the small departures from the ideal Hamming-weight coordinate measured in Section 2 explain the remaining mismatch with the clean theory. Write the measured coordinate as $u = LX/x + \delta u$. Using this Fourier representation, a one-Gaussian-per-length approximation with mean $\mu_{\delta u}(x)$ and standard deviation $\sigma_{\delta u}(x)$ multiplies the nth odd harmonic by

$$\exp\left(in\pi\mu_{\delta u}(x) - \frac{1}{2}n^2\pi^2\sigma^2_{\delta u}(x) \right). \tag{17}$$

The mean phase error shifts the harmonic, while the spread preferentially damps higher frequencies. Zero-margin even-b contacts are strongly regularized; $2L, 4L, \ldots$ are the most visible $a = 1$ members of this family. Finite-margin odd/odd packets such as $3L$ and $5L$ instead survive small noise and are attenuated smoothly. The full characteristic-function derivation and Gaussian-resummed accuracy formula are in Appendix B.3.

Using the measured $\mu_{\delta u}(x)$ and $\sigma_{\delta u}(x)$ substantially improves quantitative agreement at distant lengths. Over $150 \le x < 300$, RMSE falls by 61.0% and Pearson r rises from 0.480 to 0.917; over $300 \le x \le 500$, RMSE falls by 43.8% and r rises from 0.549 to 0.910. Full RMSE, MAE, and correlation results are in Appendix A.7.

4 Explaining the $L^{3/2}$ Learning-Time Law

Section 2 showed that the Transformer learns a representation in which nearby Hamming weights lie along an almost equally spaced line. This section asks how parity is learned once that geometry is fixed. For the analytical model, we take the post-attention SEP features themselves as given and assume that the central Hamming-weight shells lie at equally spaced points on an affine line. The feature-producing parameters are held fixed, while only the feed-forward and readout layers are trained. This makes the parity-relevant spectrum analytically solvable and leads to a learning time proportional to $L^{3/2}$. We then test the same spectral mechanism in a separate controlled model, using a frozen attention module to supply the Hamming-weight-based features and training only the feed-forward/readout layers.

4.1 From fixed Hamming-weight features to the effective kernel

We treat the post-attention SEP representation as fixed and study learning only in the feed-forward/readout tail. For Hamming-weight shell X, let

$$\vec{x}_X = \vec{V}_6 + t_X\vec{V}_5, \qquad t_X = \frac{X - L/2}{L}. \tag{18}$$

Thus neighboring shells are equally spaced in the affine coordinate. The feature-producing parameters are frozen and only W_4, W_5, W_d are trained. We use a uniformly weighted central window of $M = 2\sqrt{L}$ shells, so the corresponding population kernel operator is K/M. The complete initialization-averaged ReLU neural tangent kernel (NTK; Jacot et al., 2018) of this trainable tail is derived in Appendix C.1. For the solvable learning dynamics we set $f_0 = 0$, so the analytical prediction does not condition on a realized trainable-tail initialization; the finite-width networks in Section 4.4 are nevertheless trained from fresh random tail initializations.

Appendix C.2 shows that neighboring central feature directions have angular spacing

$$\delta = \frac{\gamma}{L} + O(L^{-2}), \tag{19}$$

where γ is fixed by the local affine geometry. Approximating this spacing as uniform, the linear angular cusp of the ReLU NTK gives the effective kernel

$$\boxed{K_0(i,j) = P_0\left[\frac{1}{2} - \frac{\delta}{2\pi}|i-j|\right], \qquad P_0 = \|\vec{x}(0)\|^2.} \tag{20}$$

Appendix C.3 gives the exact decomposition of the full NTK around K_0. We use this kernel below to isolate the parity-relevant spectral scale.

4.2 THE PARITY-RELEVANT MODES BECOME SLOW AS M^{-3}

The kernel in Eq. (20) can be diagonalized explicitly, and because parity alternates sign between neighboring Hamming-weight shells, its spectral weight is concentrated near the highest-frequency end of the K_0 spectrum. Let $y_i = (-1)^i$ and $\langle f, g\rangle_M = M^{-1}\sum_i f_i g_i$.
For the uniformly weighted operator $\mathcal{A}_0 = K_0/M$, the relevant cosine modes have components

$$[\vec{\psi}_k]_i = \sqrt{2}\cos\left[\frac{\pi k}{M}\left(i+\frac{1}{2}\right)\right], \tag{21}$$

and, in the parity sector,

$$\lambda_k = \frac{P_0\delta}{\pi M}\frac{1}{4\sin^2(\pi k/(2M))}, \qquad k \text{ odd}. \tag{22}$$

Near the highest-frequency edge, write $k = M - r$ with fixed odd r. Then

$$|\langle \vec{y}, \vec{\psi}_{M-r}\rangle_M|^2 \longrightarrow \frac{8}{\pi^2 r^2}, \qquad \lambda_{M-r} = \frac{P_0\gamma}{\pi M^3}[1+o(1)]. \tag{23}$$

Thus an order-one fraction of the parity target lies on modes whose eigenvalues shrink as M^{-3}. These are the modes that set the slow learning timescale.

4.3 FROM THE SPECTRUM TO THE $L^{3/2}$ LEARNING-TIME LAW

Under squared-loss gradient flow, a kernel eigenmode with eigenvalue λ decays on a timescale proportional to $1/\lambda$. Equation (23) therefore gives a parity-learning timescale of order M^3. Since the central Hamming-weight window contains $M = \Theta(\sqrt{L})$ shells,

$$M^3 = \Theta(L^{3/2}). \tag{24}$$

The exact asymptotic calculation, with η the gradient-flow learning-rate parameter, gives, for a fixed target MSE ϵ,

$$T_\epsilon(M) = \frac{\pi M^3}{4\eta P_0\gamma}\log\frac{1}{\epsilon}[1+o(1)], \qquad T_\epsilon(L) = \frac{2\pi}{\eta P_0\gamma}L^{3/2}\log\frac{1}{\epsilon}[1+o(1)]. \tag{25}$$

For the Jump-ReLU activation used in the controlled validation below, the same effective-kernel shape and Neumann spectrum hold with the cusp coefficient replaced by $c_0^J = d_m + d_{ff} + P_0$; Appendix C.9 gives the corresponding kernel decomposition, and Appendix D gives the exact finite-M prediction used in Figure 4.
The central prediction is therefore the exponent: the controlled model gives a parity-relevant learning time that grows as $L^{3/2}$, providing the spectral comparison to the end-to-end scaling measured in Section 2.

4.4 CONTROLLED VALIDATION

We next test whether the effective kernel K_0 captures the learning-time behavior of a finite trained network. We use a separate controlled model with one frozen attention head and no positional encoding in the active forward pass. Its initialization is filtered to produce a stable Hamming-weight-based feature map, after which only W_4, W_5, W_d are trained. The activation is

$$\phi_J(z) = z_+ + \mathbf{1}_{z>0}, \tag{26}$$

a Jump-ReLU variant that preserves the same M^{-3} spectral mechanism while increasing the strength of the A_0 cusp. For this activation the effective-kernel coefficient is

$$c_0^J = d_m + d_{ff} + P_0.$$

The finite-width model uses $d_m = 320$ and $d_{ff} = 10{,}000$ and is trained with full-batch SGD at learning rate 5×10^{-6}. For each sequence length L, training uses the uniformly weighted central shell set

$$I_L = \left\{k \in \{0,\dots,L\} : \left|k - \frac{L}{2}\right| \le \sqrt{L}\right\}, \qquad M(L) = |I_L|. \tag{27}$$

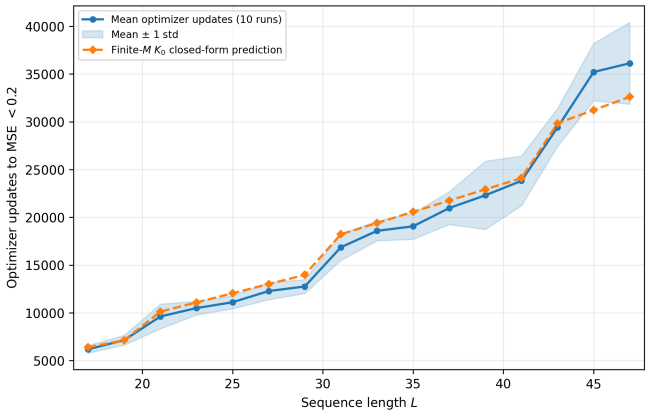

Figure 4: Controlled validation of the closed-form effective theory. Blue points show the mean optimizer updates required to reach training MSE < 0.2 over ten finite-width runs; shading denotes one standard deviation. The dashed closed-form curve is the exact finite-M Jump-ReLU K_0 prediction obtained from the Neumann eigensystem on the same central shell set, with its geometric parameters measured from the frozen features and no fit to the training times. Odd lengths $L = 17, 19, \ldots, 47$ are used so that the retained shell count M is even and the parity-sector formulas apply exactly.

We report odd lengths $L = 17, 19, \ldots, 47$, for which M is even and the exact parity-sector Neumann formulas apply. At each length we run ten finite-width trainings using the same frozen feature extractor and fresh W_4, W_5, W_d initializations, and record the first optimizer update at which the MSE falls below 0.2.

Figure 4 compares these training times with the exact finite-M K_0 prediction. For each L, the local frozen-feature geometry determines P_0 and the angular spacing δ; the resulting Neumann eigenvalues and parity spectral weights then determine the predicted time to MSE 0.2. No multiplicative timing constant is fitted to the finite-width runs. The complete finite-M formula and its numerical validation are given in Appendix D.

Across $17 \leq L \leq 47$, the closed-form prediction closely follows the absolute learning-time scale while retaining the discrete steps caused by the finite shell count $M(L)$. This provides a finite-width test of the same M^{-3} parity-edge mechanism that gives the $L^{3/2}$ scaling in Section 4.3.

5 Conclusion

We showed that a Transformer trained on parity at a single sequence length can lose measurable parity signal at nearby unseen lengths and recover it much farther away in a predictable pattern. This behavior is explained by the model's Hamming-weight representation, while a separate spectral analysis accounts for the observed $L^{3/2}$ growth of learning time with sequence length. A natural next question is how broadly such non-monotone behavior under distribution shift occurs beyond parity. In particular, it would be interesting to revisit existing length-generalization experiments—and to study non-algebraic or weakly structured tasks—to ask whether apparent failure at intermediate shifts can similarly coexist with structured recovery farther from the training distribution.

REFERENCES

Anil, C., Wu, Y., Andreassen, A., Lewkowycz, A., Misra, V., Ramasesh, V., Slone, A., Gur-Ari, G., Dyer, E., and Neyshabur, B. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 2022.

Behrens, F., Biggio, L., and Zdeborová, L. Counting in small Transformers: The delicate interplay between attention and feed-forward layers. *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267, 3500–3532, 2025.

Chi, T.-C., Fan, T.-H., Rudnicky, A. I., and Ramadge, P. J. Transformer working memory enables regular language reasoning and natural language length extrapolation. *Findings of EMNLP*, 2023.

Gurnee, W., Ameisen, E., Kauvar, I., Tarng, J., Pearce, A., Olah, C., and Batson, J. When models manipulate manifolds: The geometry of a counting task. *arXiv:2601.04480*, 2026.

Hofstätter, F., van der Weij, T., Teoh, J., Djoneva, R., Bartsch, H., and Ward, F. R. The elicitation game: Evaluating capability elicitation techniques. *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267, 23330–23356, 2025.

Hron, J., Bahri, Y., Sohl-Dickstein, J., and Novak, R. Infinite attention: NNGP and NTK for deep attention networks. *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119, 4376–4386, 2020.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Lavie, I., Gur-Ari, G., and Ringel, Z. Towards understanding inductive bias in Transformers: A view from infinity. *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, 26043–26069, 2024.

Liu, B., Ash, J. T., Goel, S., Krishnamurthy, A., and Zhang, C. Transformers learn shortcuts to automata. *International Conference on Learning Representations*, 2023.

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., and Liu, Y. RoFormer: Enhanced Transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., and Ward, F. R. AI sandbagging: Language models can strategically underperform on evaluations. *International Conference on Learning Representations*, 2025.

Vasudeva, B., Fu, D., Zhou, T., Kau, E., Huang, Y., and Sharan, V. Transformers learn low sensitivity functions: Investigations and implications. *International Conference on Learning Representations*, 2025.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and editing, LaTeX formatting, and discussion and checking of parts of the mathematical derivations and analysis. The authors independently checked the mathematical statements, experimental results, figures, and citations and take responsibility for the final content.

A EXPERIMENTAL AND DIAGNOSTIC DETAILS FOR SECTION 2

This appendix gives the implementation and analysis details behind the two empirical results in Section 2: the length-dependent parity-signal curve obtained after training only at $L = 90$, and the optimizer-time scaling measured when parity is trained separately at different lengths. We also define the representation-geometry and phase-deformation diagnostics used to connect the first phenomenon to the post-attention SEP representation.

A.1 End-to-end Transformer experiments

Task and data. Each input is a binary sequence $(x_1, \ldots, x_L)$ with independently sampled bits $x_i \sim \mathrm{Bernoulli}(1/2)$, followed by a separator token SEP. The target is parity, represented during training as a scalar in $\{-1, +1\}$. Only the output at SEP is supervised. For a sequence of length L, we write its Hamming weight as

$$X = \sum_{i=1}^{L} x_i.$$

All train, validation/test, and inference examples are sampled from the same Bernoulli product measure unless a diagnostic below explicitly states otherwise.

Architecture. The model is the single-block causal decoder defined in Section 2.1. It uses $d_m = 64$, three attention heads with $d_k = 20$, RoPE on queries and keys, summed head outputs, and no residual connection, LayerNorm, or dropout. The feed-forward nonlinearity is ReLU. Both the length-re-emergence experiment and the learning-time sweep use $d_{ff} = 500$. All learned linear maps use LeCun-normal initialization with standard deviation $1/\sqrt{\text{fan-in}}$ and zero bias. Training uses AdamW with learning rate 5×10^{-4} and weight decay 1.

Single-length re-emergence experiment. For the experiment in Fig. 1, every model is trained only at $L = 90$. We attempted 30 independent random initializations and retained the 26 runs satisfying the experiment's success criterion. Training stops when the training MSE first falls below 0.3, with a maximum budget of 15,000 optimizer updates. The effective training batch at $L = 90$ is $55{,}902$ examples, obtained from the schedule

$$B(L) = \mathrm{round}\Big(25{,}000\sqrt{L/18}\Big).$$

For each run, the training batch is sampled once and held fixed throughout training; each random initialization uses a freshly sampled batch. A held-out set of 15,000 examples is used for test-loss monitoring. After training, every successful model is evaluated at every integer test length $x = 2, \ldots, 550$, using 5,000 fresh random binary strings at each length.

Learning-time sweep. For Fig. 2, models are trained independently at each even length

$$L = 18, 20, \ldots, 68,$$

with 15 attempts per length. We record the first optimizer update at which the training MSE falls below 0.2, with a maximum budget of 20,000 optimizer updates. More than 97% of attempted runs cross the threshold within the budget. The points in Fig. 2 are the median crossing times and the error bars are the interquartile range across runs at each L. The empirical power law is obtained by ordinary least squares on the log–log relation between the per-length medians and L,

$$\log T_{\mathrm{med}}(L) = \log A + b \log L.$$

This gives $b = 1.579$ and $R^2 = 0.989$. Throughout the paper, "optimizer steps" means completed parameter updates; it is not wall-clock time.

A.2 Parity-signal energy and the noiseless comparison curve

For a trained run r evaluated at length x, let $\mathrm{Acc}_r(x)$ be its parity accuracy. We define

$$s_r(x) = 2\mathrm{Acc}_r(x) - 1, \qquad E_r(x) = s_r(x)^2, \qquad E_{\mathrm{emp}}(x) = \frac{1}{R}\sum_{r=1}^{R} E_r(x),$$

with $R = 26$ successful runs in the $L = 90$ experiment. Squaring removes the orientation of the parity relation: a run with strong anti-parity behavior contributes the same signal energy as a run with equally strong parity behavior. This is useful because small run-dependent shifts of the microscopic decision phase can reverse parity and anti-parity while preserving the broader resonance packet.

The clean comparison curve is evaluated directly from the strict discrete classifier used in Eq. (4) of the main text,

$$\mathrm{Acc}_0(x) = \sum_{X=0}^{x} 2^{-x} \binom{x}{X} \mathbf{1}\left\{\cos\left(\pi \frac{L-x}{x} X\right) > 0\right\}, \qquad E_0(x) = \big(2\mathrm{Acc}_0(x) - 1\big)^2.$$

The strict inequality is important at zero-margin contact resonances, including $2L, 4L, 6L, \ldots$, where finite binomial mass can land exactly on a decision boundary. These contacts therefore remain present in the orange noiseless-classifier curve.

Because the trained network does not interpolate perfectly even at the training length, the clean theory is given one global amplitude normalization,

$$\widetilde{E}_0(x) = C E_0(x), \qquad C = \frac{E_{\mathrm{emp}}(L)}{E_0(L)}.$$

For the strict classifier $E_0(L) = 1$, and the measured value is

$$C = E_{\mathrm{emp}}(90) = 0.735742.$$

No additional length-dependent parameter is fitted to the clean curve.

For visualization, the main panel in Fig. 1 begins at $x = 15$. The omitted very-short region $2 \le x < 15$ is not discarded from the analysis; it is treated as a separate quantitative regime in Appendix A.7. Primary guide lines mark the finite-margin resonance family $L/5, L/3, L, 3L, 5L$, while lighter guides mark selected $a = 1$ zero-margin contacts $2L, 4L, 6L$.

A.3 Sampling for representation diagnostics

All representation diagnostics use the post-attention SEP vector

$$z = z_{\mathrm{SEP}} \in \mathbb{R}^{64},$$

namely the sum of the three attention-head outputs after each head has been mapped back to model space and immediately before the first feed-forward matrix W_4.

The natural test distribution concentrates sharply around $X \simeq x/2$, which makes rare Hamming-weight shells difficult to characterize directly. We therefore estimate shell geometry using a balanced diagnostic bank: at each length x, the bank contains the same number of examples from every shell $X = 0, \ldots, x$. At the training length, 64 examples per shell are used for the detailed before/after comparison. For the all-length diagnostic, 16 examples per shell are used at every x. For a given x, exactly the same deterministic shell bank is passed through the pre-training and post-training model, so changes cannot be attributed to different sampled bit arrangements.

Balanced sampling is used only to estimate shell centers and within-shell spread accurately. Aggregate quantities are reweighted to the natural test measure

$$p_X = \Pr\{X\} = 2^{-x} \binom{x}{X}.$$

Thus rare shells do not receive artificial statistical importance merely because they were oversampled in the diagnostic bank.

For shell X, let

$$\mu_X = \mathbb{E}[z \mid X]$$

be its empirical center and define its RMS radius by

$$r_X = \sqrt{\mathbb{E}[\|z - \mu_X\|_2^2 \mid X]}.$$

For quantities associated with the gap $X \to X + 1$, we use the normalized weights

$$g_X \propto \frac{p_X + p_{X+1}}{2}, \qquad \sum_{X=0}^{x-1} g_X = 1.$$

A.4 COLLAPSE, COLLINEARITY, AND EQUAL SPACING

The three geometry diagnostics in Section 2.4 deliberately separate different properties of the representation.

Within-shell collapse. We compare the typical within-shell radius with the typical separation of neighboring shell centers,

$$R_{\mathrm{collapse}} = \frac{\sum_X p_X r_X}{\sum_X g_X \|\mu_{X+1} - \mu_X\|_2}.$$

A small value means that different bit arrangements with the same Hamming weight collapse tightly compared with the distance between neighboring Hamming weights.

One-dimensionality. Let

$$\bar{\mu} = \sum_X p_X \mu_X$$

be the binomially weighted center. We compute the first principal component q of the weighted shell-center cloud, equivalently the first right singular vector of the rows

$$\sqrt{p_X}(\mu_X - \bar{\mu}).$$

The fraction of shell-center variance explained by this direction is

$$R^2_{\mathrm{line}} = \frac{\sum_X p_X \left[q^\top(\mu_X - \bar{\mu})\right]^2}{\sum_X p_X \|\mu_X - \bar{\mu}\|_2^2}.$$

The PCA sign is oriented so that the projected coordinate increases with Hamming weight.

Equal spacing. Write

$$s_X = q^\top(\mu_X - \bar{\mu}), \qquad d_X = s_{X+1} - s_X, \qquad \bar{d} = \sum_X g_X d_X.$$

We quantify nonuniform spacing by

$$\mathrm{CV}_{\mathrm{spacing}} = \frac{\sqrt{\sum_X g_X (d_X - \bar{d})^2}}{|\bar{d}|}.$$

This diagnostic is intentionally separate from R^2_{line}: shell centers can lie perfectly on one line while being very nonuniformly spaced along it.

Observed sharpening at $L = 90$. The main text reports the across-run central tendency of these diagnostics. Already at initialization the shell centers have a strong one-dimensional tendency, with approximately

$$R^2_{\mathrm{line}} = 0.99176, \qquad R_{\mathrm{collapse}} = 1.705, \qquad \mathrm{CV}_{\mathrm{spacing}} = 0.435.$$

After successful training at $L = 90$ they become

$$R^2_{\mathrm{line}} = 0.999999965, \qquad R_{\mathrm{collapse}} = 0.0039, \qquad \mathrm{CV}_{\mathrm{spacing}} = 5.89 \times 10^{-4}.$$

The geometry remains highly structured far beyond the training length: at $x = 550 \simeq 6.1L$, for example,

$$R_{\mathrm{collapse}} = 0.0114, \qquad R^2_{\mathrm{line}} = 0.999999960.$$

These numbers support the interpretation used in the main text: training does not create one-dimensionality from nothing; rather, it strongly sharpens a pre-existing Hamming-weight organization by collapsing same-X permutations and regularizing the shell spacing.

The three scale-free geometry metrics should be interpreted as representation diagnostics, not as sufficient conditions for task success. In particular, an almost perfectly straight and uniformly spaced line could in principle have too little absolute separation, or be paired with an unsuitable downstream readout. The phase diagnostic below therefore uses the learned training-length coordinate itself rather than only intrinsic line quality.

A.5 FIXED TRAINING-LENGTH PHASE COORDINATE

The intrinsic geometry metrics above fit a fresh weighted PCA direction separately at every test length. This answers the local question "does a good Hamming-weight line exist at this length?" The phase-deformation diagnostic asks a different question: "where does the test representation lie relative to the coordinate system learned at $L = 90$?" For this second question the training-length coordinate is fitted once after training and then frozen.

At $x = L$, let q and $\bar{\mu}$ be the weighted PCA direction and origin defined above. Project the training shell centers onto that line,

$$s_X = q^\top(\mu_X - \bar{\mu}),$$

and perform a binomially weighted scalar least-squares fit

$$s_X \simeq \alpha + \beta X.$$

PCA determines the geometric affine line; this scalar fit calibrates distance along the line into Hamming-weight units. For an arbitrary SEP representation z, define

$$u(z) = \frac{q^\top(z - \bar{\mu}) - \alpha}{\beta}.$$

Equivalently, setting $\mathbf{v} = \beta q$ and $\mathbf{a} = \bar{\mu} + \alpha q$ gives the form used in Eq. (6),

$$u(z) = \frac{\mathbf{v}^\top(z - \mathbf{a})}{\|\mathbf{v}\|^2}.$$

For an example of length x and Hamming weight X, the ideal fraction-preserving cross-length coordinate is

$$u_0(x, X) = \frac{LX}{x}.$$

The phase displacement is therefore

$$\delta u(z, X, x) = u(z) - \frac{LX}{x}.$$

At every test length we evaluate δu on the same balanced Hamming-weight bank used for the geometry diagnostic, but compute its moments with the natural binomial weights. The resulting mean and standard deviation are denoted

$$\mu_{\delta u}(x), \qquad \sigma_{\delta u}(x).$$

At the training length, the median within-run spread is only

$$\sigma_{\delta u}(L) \simeq 0.0016.$$

By $x = 3L$ it grows to about 0.020. The run-dependent mean shift is also coherent across distant lengths: across successful runs,

$$\mathrm{corr}[\mu_{\delta u}(270), \mu_{\delta u}(550)] = 0.980.$$

Thus the representation remains extremely close to a one-dimensional Hamming-weight manifold while its phase relative to the frozen training-length coordinate develops a small but systematic displacement and spread.

A.6 EMPIRICAL RESONANCE-PACKET WIDTHS

The raw energy curve has a strong integer-length comb inside each broad resonance packet. To estimate the packet envelope rather than the microscopic alternation, we first smooth the empirical mean energy by averaging adjacent lengths. Around each of the three finite-margin packets

$$x_0 \in \{L, 3L, 5L\} = \{90, 270, 450\},$$

we fit a Gaussian plus constant background over a window of width proportional to $\sqrt{x_0}$,

$$E_{\mathrm{env}}(x) = b + a \exp\left[-\frac{(x - x_c)^2}{2\sigma^2}\right], \qquad |x - x_0| \le 4\sqrt{x_0}.$$

The reported width is the full width at half maximum,

$$W = 2\sqrt{2 \ln 2}\, \sigma.$$

The fitted widths are

$$W(90) = 12.95, \qquad W(270) = 21.89, \qquad W(450) = 29.05.$$

Applying the identical smoothing and Gaussian-plus-background fit to the exact noiseless curve $E_0(x)$ gives $W_0 = (12.70, 19.74, 27.61)$ at $x_0 = (90, 270, 450)$, respectively. A one-parameter fit to $W = a\sqrt{x_0}$ gives

$$W = 1.356\sqrt{x_0}, \qquad R^2 = 0.998.$$

As a check that this scaling is not an artifact of the Gaussian fit, a direct half-prominence estimator gives the closely related coefficient 1.324. Zero-margin contact resonances, including $2L, 4L, 6L$, are excluded from this width analysis because they are discrete contacts rather than broad finite-margin packets.

A.7 Quantitative agreement of clean and phase-damped theory

The main comparison in Section 2 concerns the mean empirical energy curve. Accordingly, the reported RMSE, MAE, and Pearson correlation are computed *after* averaging the empirical energy over the 26 successful runs. They are not obtained by computing a metric separately for each run and then averaging those metrics.

For the phase-damped theory, the measured $\mu_{\delta u,r}(x)$ and $\sigma_{\delta u,r}(x)$ are inserted into the Gaussian phase formula separately for each successful run r. Energy is formed at the run level and then averaged over runs. This preserves run-to-run phase variation and avoids replacing the distribution of phase parameters by their mean before applying the nonlinear theory. The same single normalization constant $C = 0.735742$ fixed at $x = L$ is then applied to both the clean and phase-damped curves. Thus adding the phase correction introduces no additional fitted amplitude parameter.

For a set of test lengths $\mathcal{X}$, define

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \left(E_{\mathrm{pred}}(x) - E_{\mathrm{emp}}(x)\right)^2},$$

$$\mathrm{MAE} = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \left|E_{\mathrm{pred}}(x) - E_{\mathrm{emp}}(x)\right|,$$

and let r be the Pearson correlation between the predicted and empirical energy values over the same set of lengths.

A single global score can be misleading here because many lengths have nearly zero signal energy. We therefore report separate regimes. The very-short interval $x < 15$ is isolated explicitly rather than being hidden by the cropped main panel. The remaining ranges are chosen to separate the near/training-length structure from the two distant regimes emphasized in the main text.

Range	RMSE_0	MAE_0	r_0	$\mathrm{RMSE}_{\mathrm{damp}}$	$\mathrm{MAE}_{\mathrm{damp}}$	r_{damp}
$x < 15$	0.30560	0.25425	0.234	0.23005	0.20912	0.086
$15 \le x < 150$	0.08480	0.04434	0.876	0.06139	0.03731	0.943
$150 \le x < 300$	0.02217	0.01282	0.480	0.00865	0.00697	0.917
$300 \le x \le 500$	0.00688	0.00503	0.549	0.00386	0.00345	0.910

Table 1: Agreement between the run-averaged empirical energy and the normalized noiseless classifier theory (0), versus the same normalization with the measured phase-damping correction.

In the two distant regimes, adding the independently measured phase deformation substantially improves both amplitude agreement and curve shape. For $150 \le x < 300$, RMSE falls by 61.0% and MAE by 45.7%, while Pearson correlation rises from 0.480 to 0.917. For $300 \le x \le 500$, RMSE

falls by 43.8% and MAE by 31.4%, while Pearson correlation rises from 0.549 to 0.910. In the intermediate $15 \le x < 150$ regime the clean theory already captures the main structure, and damping gives a smaller improvement. The $x < 15$ regime is qualitatively different: although damping reduces RMSE and MAE, its Pearson correlation is poor, which is why this region is displayed and reported separately rather than being used to visually support the main agreement claim.

The final tail $500 < x \le 550$ is retained in the main figure because it includes the $6L$ contact at $x = 540$, but it is not treated as a separate broad-range agreement bin. In that tail the energy is mostly small and the number of remaining lengths is limited, making correlation increasingly sensitive to individual contact features.

A.8 Relation between the diagnostics and the main claims

The diagnostics are intended to separate three logically distinct statements.

First, R_{collapse}, R^2_{line}, and $\mathrm{CV}_{\mathrm{spacing}}$ establish that successful models organize the post-attention SEP representation primarily by Hamming weight, and that training turns this organization into a tightly collapsed, nearly affine count coordinate. These are intrinsic geometric statements and do not depend on the downstream classifier.

Second, δu, $\mu_{\delta u}$, and $\sigma_{\delta u}$ compare unseen-length representations to the *fixed* coordinate learned at the training length. They therefore quantify cross-length phase deformation rather than simply asking whether each unseen-length representation can be well fit by some new line.

Third, the energy comparison asks whether the measured deformation is large enough to explain the observed deviations from the clean noiseless classifier. The strong reduction in distant-range RMSE/MAE and the large increase in Pearson correlation support that connection. The analysis does not claim that the three intrinsic geometry metrics alone are sufficient to predict whether an arbitrary training run succeeds; they are used to characterize the representation of the successful models whose extrapolation behavior is analyzed in Section 2.

B Derivations for Section 3

B.1 From the affine Hamming-weight line to an ideal zigzag classifier

This appendix supplies the construction used in Section 3.2. Two statements are useful to keep separate. First, if the post-attention SEP representations for the $L+1$ training Hamming weights lie on one affine line, then a two-layer ReLU network of width $L+1$ can interpolate alternating parity labels on those points. Second, if the shell centers are also equally spaced along that line, the resulting minimum-knot interpolation has decision boundaries at the half-integers of a natural scalar coordinate. This second, stronger statement yields the closed-form success-by-length law used in the main text.

Affine coordinate. Let the training-length shell centers be

$$\bar{\vec{z}}_L(k) = \bar{\vec{a}} + k\bar{\vec{v}}, \qquad k = 0, 1, \ldots, L, \tag{28}$$

with $\bar{\vec{v}} \neq 0$. We use the scalar coordinate u defined by

$$\vec{z}_{\mathrm{line}}(u) = \bar{\vec{a}} + u\bar{\vec{v}}, \tag{29}$$

so that the training shells sit at $u = k$. The parity targets are

$$y_k = (-1)^k. \tag{30}$$

The model used in the experiments has no explicit biases. Along an affine line, however, a bias-free linear form can realize an affine function of u. Assuming $\bar{\vec{a}}$ and $\bar{\vec{v}}$ are linearly independent, choose vectors $\vec{p}, \vec{q}$ in representation space such that

$$\vec{p}^{\top}\bar{\vec{a}} = 1, \qquad \vec{p}^{\top}\bar{\vec{v}} = 0, \qquad \vec{q}^{\top}\bar{\vec{a}} = 0, \qquad \vec{q}^{\top}\bar{\vec{v}} = 1. \tag{31}$$

Then along the line

$$\vec{p}^{\top}\vec{z}_{\mathrm{line}}(u) = 1, \qquad \vec{q}^{\top}\vec{z}_{\mathrm{line}}(u) = u, \qquad (\vec{q} - j\vec{p})^{\top}\vec{z}_{\mathrm{line}}(u) = u - j. \tag{32}$$

Thus the bias-free ReLU layer can produce a constant unit, the linear coordinate u on $u \ge 0$, and hinges $\mathrm{ReLU}(u - j)$. If explicit biases are present, the same construction is immediate and the linear-independence condition on $\bar{\vec{a}}, \bar{\vec{v}}$ is unnecessary.

Width-$L+1$ zigzag interpolation. Consider

$$f_L(u) = 1 - 2u + 4\sum_{j=1}^{L-1}(-1)^{j+1}\,\mathrm{ReLU}(u-j). \tag{33}$$

Equation (33) uses one constant hidden unit, one unit for u, and $L-1$ hinge units, for a total width $L+1$. Its slope on the first interval $(0,1)$ is -2. When u passes the knot j, the slope changes by

$$4(-1)^{j+1}. \tag{34}$$

Hence the slope on $(k, k+1)$ is

$$s_k = -2(-1)^k. \tag{35}$$

Starting from $f_L(0) = 1$, the change over one unit interval is therefore

$$f_L(k+1) - f_L(k) = s_k = -2(-1)^k, \tag{36}$$

which gives by induction

$$f_L(k) = (-1)^k, \qquad k = 0, \ldots, L. \tag{37}$$

More explicitly, for $u \in [k, k+1]$,

$$f_L(u) = (-1)^k\big[1 - 2(u-k)\big]. \tag{38}$$

Therefore the unique zero in that interval lies at

$$c_k = k + \frac{1}{2}, \tag{39}$$

and the classifier alternates sign once per training cell. On $0 \le u \le L$, away from the zeros,

$$\operatorname{sgn} f_L(u) = \operatorname{sgn}\cos(\pi u). \tag{40}$$

The cosine is only a convenient sign-equivalent representation of the piecewise-linear classifier; the ReLU output itself is the zigzag in Eq. (38).

Cross-length coordinate and the strict noiseless classifier. For a test sequence of length x and Hamming weight X, the ideal representation hypothesis preserves normalized Hamming weight X/x. Rescaling to the training coordinate therefore gives

$$u_0(x, X) = L\frac{X}{x}. \tag{41}$$

The ideal classifier predicts

$$\widehat{y}_0(x, X) = \operatorname{sgn}\cos\left(\pi\frac{L}{x}X\right). \tag{42}$$

The true parity label is $y(X) = (-1)^X = \cos(\pi X)$. Correct classification is therefore equivalent to

$$(-1)^X \cos\left(\pi\frac{L}{x}X\right) > 0. \tag{43}$$

Because X is an integer, $\sin(\pi X) = 0$. Using $\cos(A-B) = \cos A\cos B + \sin A\sin B$ with $A = \pi LX/x$ and $B = \pi X$ gives

$$\cos(\pi X)\cos\left(\pi\frac{L}{x}X\right) = \cos\left(\pi\frac{L-x}{x}X\right). \tag{44}$$

Hence the strict noiseless correctness rule is

$$\text{correct at } (x, X) \quad\Longleftrightarrow\quad \cos\left(\pi\frac{L-x}{x}X\right) > 0. \tag{45}$$

For uniformly random binary strings, $X \sim \mathrm{Binomial}(x, 1/2)$, with mass

$$p_x(X) = 2^{-x}\binom{x}{X}. \tag{46}$$

Therefore the exact strict noiseless accuracy is

$$A_0(x) = \sum_{X=0}^{x} p_x(X)\mathbf{1}\left\{\cos\left(\pi\frac{L-x}{x}X\right) > 0\right\}. \tag{47}$$

The signed parity signal and energy used in the paper are

$$s_0(x) = 2A_0(x) - 1, \qquad E_0(x) = s_0(x)^2. \tag{48}$$

The strict inequality in Eq. (47) matters at the special lengths where a nonzero amount of the discrete binomial mass lands exactly on a decision boundary. These zero-margin contact resonances are derived below.

B.2 Large-x expansion of the RoPE shell geometry

This appendix derives the representation-level statements used in Section 3.1. We retain the paper's notation: x is the input length, X is its Hamming weight, and $\vec{z}^{(x)}_{\mathrm{SEP}} \in \mathbb{R}^{d_m}$ is the post-attention representation after each head has been mapped back to model space by $W_O^{(h)}$ and the heads have been summed. The goal is to make precise three claims: at fixed Hamming weight, RoPE produces a well-defined shell center plus arrangement-dependent scatter; the absolute width of that scatter scales as $x^{-1/2}$ at fixed head count in the regular regime; and at the training length $x = L$, the $O(\sqrt{L})$ Hamming-weight window carrying most of the binomial mass samples a locally affine and locally equally spaced portion of the shell-center trajectory.

For a binary input, let

$$S \subseteq \{1, \dots, x\}, \qquad |S| = X, \qquad \rho = \frac{X}{x}, \tag{49}$$

denote the positions containing token 1. Conditioning on Hamming weight X means that S is sampled uniformly from the $\binom{x}{X}$ subsets of size X.

Exact one-head RoPE representation. The network definition in the main paper uses the head-space value vector $\vec{v}_i^{(h)} = W_V^{(h)}\vec{e}_i$. For the geometry below it is more convenient to work directly in model space, so define

$$\vec{u}_{b,h} \equiv W_O^{(h)}W_V^{(h)}\vec{e}_b \in \mathbb{R}^{d_m}, \qquad b \in \{0, 1, \mathrm{SEP}\}. \tag{50}$$

This is the same projected value denoted $\vec{v}_{b,h}$ in the main text. Let the SEP token occupy position $s = x + 1$, and define the unrotated SEP query and token keys

$$\widetilde{\vec{q}}_h = W_Q^{(h)}\vec{e}_{\mathrm{SEP}}, \qquad \widetilde{\vec{k}}_{b,h} = W_K^{(h)}\vec{e}_b. \tag{51}$$

Because RoPE rotates the query and key at positions s and p by R_s and R_p,

$$\vec{q}_s^{(h)\top}\vec{k}_p^{(h)} = \widetilde{\vec{q}}_h^{\top}R_s^{\top}R_p\widetilde{\vec{k}}_{b,h} = \widetilde{\vec{q}}_h^{\top}R_{p-s}\widetilde{\vec{k}}_{b,h}. \tag{52}$$

Define the positive unnormalized attention factors for tokens 1 and 0 at bit position p,

$$a_{p,h} = \exp\left(\frac{\widetilde{\vec{q}}_h^{\top}R_{p-s}\widetilde{\vec{k}}_{1,h}}{\sqrt{d_k}}\right), \qquad b_{p,h} = \exp\left(\frac{\widetilde{\vec{q}}_h^{\top}R_{p-s}\widetilde{\vec{k}}_{0,h}}{\sqrt{d_k}}\right). \tag{53}$$

The SEP self-attention factor is

$$F_{\mathrm{SEP},h} = \exp\left(\frac{\widetilde{\vec{q}}_h^{\top}\widetilde{\vec{k}}_{\mathrm{SEP},h}}{\sqrt{d_k}}\right), \tag{54}$$

where $R_s^\top R_s = I$ was used. For an arrangement S, write

$$U_{1,h}(S) = \sum_{p \in S} a_{p,h}, \qquad U_{0,h}(S) = \sum_{p \notin S} b_{p,h}, \qquad D_h(S) = U_{1,h}(S) + U_{0,h}(S) + F_{\mathrm{SEP},h}. \tag{55}$$

The exact model-space contribution of head h to the SEP representation is then

$$\vec{z}_h^{(x)}(S) = \frac{U_{1,h}(S)\vec{u}_{1,h} + U_{0,h}(S)\vec{u}_{0,h} + F_{\mathrm{SEP},h}\vec{u}_{\mathrm{SEP},h}}{D_h(S)}. \tag{56}$$

Equivalently,

$$\vec{z}_h^{(x)}(S) = \vec{u}_{0,h} + A_h(S)\Delta\vec{u}_h + C_h(S)\Delta\vec{u}_{\mathrm{SEP},h}, \tag{57}$$

where

$$A_h(S) = \frac{U_{1,h}(S)}{D_h(S)}, \qquad C_h(S) = \frac{F_{\mathrm{SEP},h}}{D_h(S)}, \tag{58}$$

and

$$\Delta\vec{u}_h = \vec{u}_{1,h} - \vec{u}_{0,h}, \qquad \Delta\vec{u}_{\mathrm{SEP},h} = \vec{u}_{\mathrm{SEP},h} - \vec{u}_{0,h}. \tag{59}$$

Without the SEP self term, $C_h = 0$ and every one-head representation lies exactly on the affine line $\vec{u}_{0,h} + t\Delta\vec{u}_h$; RoPE only changes the scalar coordinate $t = A_h(S)$. SEP self-attention introduces a second possible direction, but its arrangement-dependent contribution is parametrically smaller at large x.

Sampling uniformly within a Hamming-weight shell. Introduce the indicator variables

$$z_p = \mathbf{1}\{p \in S\}, \qquad \sum_{p=1}^{x} z_p = X. \tag{60}$$

At fixed X,

$$\mathbb{E}[z_p] = \rho, \qquad \mathrm{Var}(z_p) = \rho(1-\rho), \qquad \mathrm{Cov}(z_p, z_q) = -\frac{\rho(1-\rho)}{x-1} \quad (p \neq q). \tag{61}$$

The negative off-diagonal covariance is the finite-population correction: once one position is assigned token 1, the remaining positions are slightly less likely to receive token 1 because the total number of ones is fixed.

Define the positional means

$$\bar{a}_h = \frac{1}{x}\sum_{p=1}^{x} a_{p,h}, \qquad \bar{b}_h = \frac{1}{x}\sum_{p=1}^{x} b_{p,h}, \qquad d_h(\rho) = \rho\bar{a}_h + (1-\rho)\bar{b}_h. \tag{62}$$

We work in the regular regime in which $d_h(\rho) = \Theta(1)$ and no single position carries a macroscopic fraction of the total attention mass. At fixed weights this is natural because RoPE preserves norms, so bounded query and key norms imply bounded attention scores and hence $a_{p,h}, b_{p,h} = O(1)$.

Write

$$\xi_{a,h} = \sum_{p=1}^{x}(z_p - \rho)a_{p,h}, \tag{63}$$

$$\xi_{b,h} = \sum_{p=1}^{x}(z_p - \rho)b_{p,h}, \tag{64}$$

$$\xi_{c,h} = \xi_{a,h} - \xi_{b,h} = \sum_{p=1}^{x}(z_p - \rho)(a_{p,h} - b_{p,h}). \tag{65}$$

Then

$$U_{1,h} = x\rho\bar{a}_h + \xi_{a,h}, \qquad U_{0,h} = x(1-\rho)\bar{b}_h - \xi_{b,h}, \tag{66}$$

and

$$D_h = xd_h(\rho) + \xi_{c,h} + F_{\mathrm{SEP},h}. \tag{67}$$

Equation (61) gives

$$\xi_{a,h}, \xi_{b,h}, \xi_{c,h} = O_p(\sqrt{x}). \tag{68}$$

Leading shell center and within-shell fluctuation. Define the scalar leading attention coordinate

$$\bar{A}_h(\rho) = \frac{\rho \bar{a}_h}{d_h(\rho)}. \tag{69}$$

Expanding $A_h(S) = U_{1,h}/D_h$ in the $O_p(\sqrt{x})$ fluctuations gives

$$A_h(S) = \bar{A}_h(\rho) + \frac{1}{x}\sum_{p=1}^{x}(z_p - \rho)\alpha_{p,h}(\rho) + O_p(x^{-1}), \tag{70}$$

where the scalar positional response is

$$\alpha_{p,h}(\rho) = \frac{[1 - \bar{A}_h(\rho)]a_{p,h} + \bar{A}_h(\rho)b_{p,h}}{d_h(\rho)}. \tag{71}$$

The first-order random term has conditional mean zero and typical size $O_p(x^{-1/2})$. The SEP coefficient obeys

$$C_h(S) = \frac{F_{\mathrm{SEP},h}}{x d_h(\rho)} - \frac{F_{\mathrm{SEP},h}\xi_{c,h}}{x^2 d_h(\rho)^2} + O_p(x^{-2}). \tag{72}$$

Thus SEP self-attention moves the shell center by $O(x^{-1})$, while its arrangement-dependent fluctuation is only $O_p(x^{-3/2})$.

For the full network define the exact shell center and the individual within-shell displacement by

$$\vec{\mu}_X^{(x)} = \mathbb{E}_S[\vec{z}_{\mathrm{SEP}}^{(x)}(S) \mid X], \qquad \vec{\epsilon}_X^{(x)}(S) = \vec{z}_{\mathrm{SEP}}^{(x)}(S) - \vec{\mu}_X^{(x)}. \tag{73}$$

By definition,

$$\mathbb{E}_S[\vec{\epsilon}_X^{(x)}(S) \mid X] = \vec{0}. \tag{74}$$

The two vectors in Eq. (73) describe different geometric objects: $\vec{\mu}_X^{(x)}$ is the location of the whole Hamming-weight shell, whereas $\vec{\epsilon}_X^{(x)}(S)$ is the arrangement-dependent displacement of one example from that location.

Summing Eq. (70) over heads gives

$$\vec{\mu}_X^{(x)} = \sum_{h=1}^{H}\left[\vec{u}_{0,h} + \bar{A}_h(\rho)\Delta\vec{u}_h\right] + \vec{r}_X^{(x)}, \qquad \|\vec{r}_X^{(x)}\| = O(x^{-1}), \tag{75}$$

and

$$\vec{\epsilon}_X^{(x)}(S) = \frac{1}{x}\sum_{p=1}^{x}(z_p - \rho)\vec{G}_p(\rho) + O_p(x^{-1}), \tag{76}$$

where

$$\vec{G}_p(\rho) = \sum_{h=1}^{H}\alpha_{p,h}(\rho)\Delta\vec{u}_h. \tag{77}$$

Any variance-preserving head scaling is already contained in the projected vectors $\Delta\vec{u}_h$. The two remainder statements above should not be conflated. The vector $\vec{r}_X^{(x)}$ is a deterministic finite-x correction to the mean shell location: it translates the whole cloud. By contrast, $\vec{\epsilon}_X^{(x)}(S)$ changes from one arrangement to another and sets the width and shape of the cloud. Equivalently,

$$\vec{z}_{\mathrm{SEP}}^{(x)}(S) = \underbrace{\sum_h[\vec{u}_{0,h} + \bar{A}_h(\rho)\Delta\vec{u}_h]}_{\text{leading analytical center}} + \underbrace{\vec{r}_X^{(x)}}_{\text{finite-}x\text{ center correction}} + \underbrace{\vec{\epsilon}_X^{(x)}(S)}_{\text{within-shell noise}}. \tag{78}$$

Where the $O(x^{-1})$ center correction comes from. The $O(x^{-1})$ term in Eq. (75) has two sources: SEP self-attention and the usual second-order bias produced by taking the expectation of a ratio. Let

$$c_{p,h} = a_{p,h} - b_{p,h}, \tag{79}$$

and for scalar position-dependent sequences f_p, g_p define

$$\Gamma_h(f, g) = \frac{1}{x}\sum_{p=1}^{x}(f_p - \bar{f})(g_p - \bar{g}). \tag{80}$$

A second-order ratio expansion gives

$$\mathbb{E}_S[A_h(S) \mid X] = \bar{A}_h - \frac{\bar{A}_h F_{\mathrm{SEP},h}}{x d_h} + \frac{\rho(1-\rho)}{(x-1)d_h^2}\left[\bar{A}_h \Gamma_h(c, c) - \Gamma_h(a, c)\right] + O(x^{-2}), \tag{81}$$

$$\mathbb{E}_S[C_h(S) \mid X] = \frac{F_{\mathrm{SEP},h}}{x d_h} + O(x^{-2}). \tag{82}$$

Both corrections are deterministic once the network, x, and X are fixed. They therefore change where the shell is centered; they are not part of the random scatter around that center.

Within-shell covariance and the $x^{-1/2}$ scale. Because $\sum_{p=1}^{x}(z_p - \rho) = 0$, only the position-centered part of $\vec{G}_p$ contributes. Define

$$\bar{\vec{G}}(\rho) = \frac{1}{x}\sum_{p=1}^{x}\vec{G}_p(\rho), \tag{83}$$

and

$$C_G(\rho) = \frac{1}{x}\sum_{p=1}^{x}[\vec{G}_p - \bar{\vec{G}}][\vec{G}_p - \bar{\vec{G}}]^\top. \tag{84}$$

Using Eq. (61) in Eq. (76) gives

$$\mathrm{Cov}_S(\vec{\epsilon}_X^{(x)} \mid X) = \frac{\rho(1-\rho)}{x-1}C_G(\rho) + o(x^{-1}). \tag{85}$$

This matrix equation is the precise meaning of the statement that the within-shell noise has scale $x^{-1/2}$. For every fixed unit vector $\vec{d} \in \mathbb{R}^{d_m}$,

$$\mathrm{std}_S\left(\vec{d}^\top \vec{\epsilon}_X^{(x)} \mid X\right) = \sqrt{\frac{\rho(1-\rho)}{x-1}\vec{d}^\top C_G(\rho)\vec{d}} + o(x^{-1/2}). \tag{86}$$

Thus each fixed directional projection has standard deviation $O(x^{-1/2})$ whenever $C_G(\rho) = O(1)$. For fixed model dimension d_m, the RMS Euclidean shell radius is

$$\sqrt{\mathbb{E}_S[\|\vec{\epsilon}_X^{(x)}\|^2 \mid X]} = \sqrt{\frac{\rho(1-\rho)}{x-1}\,\mathrm{tr}\, C_G(\rho)} + o(x^{-1/2}) = O(x^{-1/2}). \tag{87}$$

No ambiguous scalar "standard deviation of a vector" is needed: the first expression gives directional standard deviations, while the second gives the RMS radius of the shell. The universal finite-population factor $\rho(1-\rho)$ vanishes at $X = 0$ and $X = x$ and is maximal at $\rho = 1/2$. The full covariance need not be strictly monotone in X, because $C_G(\rho)$ can also depend on ρ.

One head gives line-like noise; many heads give a cloud. For one head,

$$\vec{G}_p(\rho) = \alpha_{p,h}(\rho)\Delta\vec{u}_h, \tag{88}$$

so the leading covariance in Eq. (85) is rank one:

$$\mathrm{Cov}_S(\vec{\epsilon}_X^{(x)} \mid X) = \frac{\rho(1-\rho)}{x-1}C_{\alpha,h}(\rho)\Delta\vec{u}_h\Delta\vec{u}_h^\top + o(x^{-1}), \tag{89}$$

where

$$C_{\alpha,h}(\rho) = \frac{1}{x}\sum_{p=1}^{x}[\alpha_{p,h}(\rho) - \bar{\alpha}_h(\rho)]^2. \tag{90}$$

Thus the leading one-head within-shell fluctuation is along the single value-difference direction $\Delta\vec{u}_h$. The SEP self term adds only a smaller $O_p(x^{-3/2})$ transverse fluctuation.
With multiple heads,

$$C_G(\rho) = \sum_{h,g} C_{\alpha,hg}(\rho)\Delta\vec{u}_h\Delta\vec{u}_g^{\top}, \tag{91}$$

where

$$C_{\alpha,hg}(\rho) = \frac{1}{x}\sum_{p=1}^{x}[\alpha_{p,h} - \bar{\alpha}_h][\alpha_{p,g} - \bar{\alpha}_g]. \tag{92}$$

The same arrangement S drives all heads, so the headwise fluctuations are not independent samples that automatically average away. Under a variance-preserving head scaling, increasing H therefore does not generically yield an additional $H^{-1/2}$ reduction of the shell radius. Its robust geometric effect is instead to enlarge the span of possible fluctuations from the one-head direction $\Delta\vec{u}_h$ to

$$\mathrm{span}\{\Delta\vec{u}_1, \ldots, \Delta\vec{u}_H\}, \tag{93}$$

producing a generally anisotropic multidimensional cloud.

The central shell-center trajectory is locally affine. We now specialize to the training length $x = L$. The leading shell-center trajectory is

$$\vec{m}(\rho) = \sum_{h=1}^{H}[\vec{u}_{0,h} + \bar{A}_h(\rho)\Delta\vec{u}_h], \qquad \bar{A}_h(\rho) = \frac{\rho\bar{a}_h}{\rho\bar{a}_h + (1-\rho)\bar{b}_h}. \tag{94}$$

For one head, this trajectory lies on one exact affine line: only the scalar $\bar{A}_h(\rho)$ changes. For several heads, different directions $\Delta\vec{u}_h$ are multiplied by different scalar functions $\bar{A}_h(\rho)$, so the summed trajectory can be curved over the full Hamming-weight range.
The derivatives of the one-head scalar coordinate are

$$\bar{A}'_h(\rho) = \frac{\bar{a}_h\bar{b}_h}{[\rho\bar{a}_h + (1-\rho)\bar{b}_h]^2}, \tag{95}$$

and

$$\bar{A}''_h(\rho) = -\frac{2\bar{a}_h\bar{b}_h(\bar{a}_h - \bar{b}_h)}{[\rho\bar{a}_h + (1-\rho)\bar{b}_h]^3}. \tag{96}$$

At $\rho = 1/2$,

$$\bar{A}'_h\left(\frac{1}{2}\right) = \frac{4\bar{a}_h\bar{b}_h}{(\bar{a}_h + \bar{b}_h)^2}, \qquad \bar{A}''_h\left(\frac{1}{2}\right) = -\frac{16\bar{a}_h\bar{b}_h(\bar{a}_h - \bar{b}_h)}{(\bar{a}_h + \bar{b}_h)^3}. \tag{97}$$

The exact shell centers differ from Eq. (94) by the finite-L correction derived above. In the regular central regime this correction is smooth in normalized Hamming weight, so a smooth interpolation of the exact centers may be written

$$\vec{\mu}_L(\rho) = \vec{m}(\rho) + \frac{1}{L}\vec{R}_L(\rho) + O(L^{-2}), \tag{98}$$

with $\vec{R}_L$ and its first few derivatives uniformly $O(1)$ near $\rho = 1/2$. Although the absolute correction is $O(L^{-1})$, the same size as one neighboring-shell gap, it is a smooth bias of the center curve rather than an independent error attached to each shell. Consequently,

$$\frac{1}{L}\left[\vec{R}_L\left(\rho + \frac{1}{L}\right) - \vec{R}_L(\rho)\right] = O(L^{-2}), \tag{99}$$

so the finite-L center correction does not spoil the leading spacing law.

Set

$$\tau = \rho - \frac{1}{2} = \frac{X - L/2}{L}. \tag{100}$$

Taylor-expanding the exact center interpolation gives

$$\vec{\mu}_L\left(\frac{1}{2} + \tau\right) = \vec{\mu}_L\left(\frac{1}{2}\right) + \tau\vec{A}_L + \frac{1}{2}\tau^2\vec{B}_L + O(\tau^3), \tag{101}$$

where

$$\vec{A}_L = \left.\frac{\partial\vec{\mu}_L}{\partial\rho}\right|_{\rho=1/2}, \qquad \vec{B}_L = \left.\frac{\partial^2\vec{\mu}_L}{\partial\rho^2}\right|_{\rho=1/2}. \tag{102}$$

To leading order,

$$\vec{A}_L = \sum_h \bar{A}'_h\left(\frac{1}{2}\right)\Delta\vec{u}_h + O(L^{-1}), \qquad \vec{B}_L = \sum_h \bar{A}''_h\left(\frac{1}{2}\right)\Delta\vec{u}_h + O(L^{-1}). \tag{103}$$

Under the training distribution $X \sim \mathrm{Binomial}(L, 1/2)$, the Hamming weights carrying non-negligible probability satisfy

$$X - \frac{L}{2} = O(\sqrt{L}), \qquad \tau = O(L^{-1/2}). \tag{104}$$

Across this window, the tangent displacement $\tau\vec{A}_L$ is $O(L^{-1/2})$, whereas the quadratic bending $\frac{1}{2}\tau^2\vec{B}_L$ is $O(L^{-1})$. Thus the bending is smaller than the tangent displacement by a relative factor $O(L^{-1/2})$. This is the sense in which the binomially relevant portion of the multi-head shell-center trajectory becomes locally affine as L grows. No large-head limit is required.

Adjacent shell spacing. Increasing X by one changes τ by $1/L$. Therefore

$$\vec{\mu}^{(L)}_{X+1} - \vec{\mu}^{(L)}_X = \frac{\vec{A}_L}{L} + \left(\frac{\tau}{L} + \frac{1}{2L^2}\right)\vec{B}_L + O\left(\frac{\tau^2}{L} + \frac{|\tau|}{L^2} + L^{-3}\right). \tag{105}$$

For $|\tau| = O(L^{-1/2})$ this reduces to

$$\vec{\mu}^{(L)}_{X+1} - \vec{\mu}^{(L)}_X = \frac{\vec{A}_L}{L} + O(L^{-3/2}). \tag{106}$$

At the exact center $\tau = 0$, the first correction is only $O(L^{-2})$. The $O(L^{-3/2})$ term is the uniform correction across the full $O(\sqrt{L})$ binomial window: the local slope changes by $O(L^{-1/2})$ across that window, and one Hamming-weight step has size $1/L$ in normalized coordinates. Since the leading gap is $O(L^{-1})$, its relative variation across the central window is $O(L^{-1/2})$. Thus the same central-window limit that makes the shell-center trajectory locally straight also makes neighboring shell centers locally equally spaced.

Absolute shell width versus neighboring-shell separation. Two large-L scales should be kept separate. The covariance calculation above gives an absolute within-shell fluctuation scale $O(L^{-1/2})$, whereas Eq. (106) gives a neighboring-center separation $O(L^{-1})$. Thus increasing L makes the RoPE cloud thinner in absolute model-space units, but does not by itself imply that adjacent Hamming-weight shells become better separated. In units of one neighboring-shell gap, a generic $O(L^{-1/2})$ cloud width corresponds to $O(\sqrt{L})$ shell spacings unless training changes the prefactors. This distinction explains how the representation can have an extremely high shell-center R^2_{line} while individual examples still show substantial within-shell overlap, as observed at initialization in Section 2. Linearity of the shell-center sequence and collapse of individual examples are different geometric properties.

Summary. The derivation establishes the representation-level claims used in Section 3.1. At fixed Hamming weight, RoPE produces a well-defined shell center plus arrangement-dependent scatter; the deterministic $O(x^{-1})$ correction shifts the center and is not part of the random cloud. For fixed head count, the conditional covariance of the within-shell residual is $O(x^{-1})$, so every fixed directional standard deviation and the RMS shell radius at fixed d_m scale as $O(x^{-1/2})$. For one head, the leading arrangement scatter is rank one and lies along $\Delta\vec{u}_h$; summing heterogeneous heads can produce a higher-rank anisotropic cloud without a generic additional $H^{-1/2}$ suppression. Finally, at the training length, the $O(\sqrt{L})$ Hamming-weight window carrying most of the binomial mass samples only a locally affine portion of the smooth shell-center trajectory, with neighboring central shell centers differing by $\vec{A}_L/L$ up to $O(L^{-3/2})$ variation.

B.3 Arithmetic echoes, contact resonances, Fourier widths, and measured phase deformation

This appendix collects the derivations used in Sections 3.3–3.6. The arithmetic locations and asymptotic strengths of the robust echoes follow directly from the strict classifier in Eq. (45). Fourier analysis is not needed to locate those echoes; its main role is to describe their local packet shape and to make the effect of measured phase perturbations analytically transparent.

B.3.1 Odd/odd rational ratios and the residue-class echo law

Write the test/train ratio in lowest terms as

$$\frac{x}{L} = \frac{b}{a}, \qquad \gcd(a, b) = 1. \tag{107}$$

Then

$$\frac{L - x}{x} = \frac{L}{x} - 1 = \frac{a - b}{b}, \tag{108}$$

so Eq. (45) becomes

$$\cos\left(\pi \frac{a - b}{b} X\right) > 0. \tag{109}$$

Assume first that a and b are both odd. Then $a - b$ is even, so write

$$a - b = 2c. \tag{110}$$

Since $\gcd(a, b) = 1$ and b is odd,

$$\gcd(c, b) = 1. \tag{111}$$

The correctness condition is therefore

$$\cos\left(2\pi \frac{cX}{b}\right) > 0. \tag{112}$$

It depends only on $X \bmod b$. Multiplication by c permutes the b residue classes because c is invertible modulo b. Consequently the asymptotic accuracy is obtained by counting the residues $r \in \{0, \ldots, b - 1\}$ for which

$$\cos\left(2\pi \frac{r}{b}\right) > 0. \tag{113}$$

It remains to justify that the binomial mass becomes uniform over residues modulo fixed b. The standard roots-of-unity filter gives

$$\Pr(X \equiv r \pmod b) = \frac{1}{b} \sum_{j=0}^{b-1} e^{-2\pi i j r/b}\, \mathbb{E}[e^{2\pi i j X/b}] \tag{114}$$

$$= \frac{1}{b} \sum_{j=0}^{b-1} e^{-2\pi i j r/b} \left(\frac{1 + e^{2\pi i j/b}}{2}\right)^x. \tag{115}$$

For $j = 0$ the term equals 1. For $j \neq 0$,

$$\left|\frac{1 + e^{2\pi i j/b}}{2}\right| = \left|\cos \frac{\pi j}{b}\right| < 1. \tag{116}$$

Thus, at fixed b and $x \to \infty$,

$$\Pr(X \equiv r \pmod b) = \frac{1}{b} + O(\rho_b^x), \qquad \rho_b = \max_{1 \le j \le b-1} \left|\cos \frac{\pi j}{b}\right| < 1. \tag{117}$$

So the limiting accuracy is exactly the fraction of a regular b-gon lying in the positive-cosine half-plane.

For odd b there are two cases. If $b = 4q + 1$, the positive residues are

$$r = 0, 1, \ldots, q \qquad \text{and} \qquad r = b - q, \ldots, b - 1, \tag{118}$$

for a total of $2q + 1 = (b + 1)/2$. If $b = 4q + 3$, the same count gives $2q + 1 = (b - 1)/2$. Both cases can be written as

$$N_+(b) = \frac{b + (-1)^{(b-1)/2}}{2}. \tag{119}$$

Therefore

$$A_0\left(\frac{b}{a}L\right) \longrightarrow \frac{1}{2} + \frac{(-1)^{(b-1)/2}}{2b}, \qquad a, b \text{ odd}, \tag{120}$$

which implies

$$s_0\left(\frac{b}{a}L\right) \longrightarrow \frac{(-1)^{(b-1)/2}}{b}, \qquad E_0\left(\frac{b}{a}L\right) \longrightarrow \frac{1}{b^2}. \tag{121}$$

This is the arithmetic echo law quoted in Section 3.3. The numerator b fixes the asymptotic sign and strength, while the odd denominator a moves the resonance to $x = (b/a)L$ without changing that limiting strength.

The exact shorter peaks are the special case $b = 1$. If $x = L/q$ with odd integer q, then $a = q$, $b = 1$, and directly from Eq. (45),

$$\cos(\pi(q - 1)X) = 1 \tag{122}$$

for every integer X, because $q - 1$ is even. Hence $A_0(L/q) = 1$ exactly whenever q is an odd divisor of L. Conversely, $b = 3$ gives an asymptotic anti-parity signal $s_0 \to -1/3$ and energy $1/9$, while $b = 5$ gives $s_0 \to +1/5$ and energy $1/25$.

B.3.2 The remaining rational classes: zero-margin contacts and silent ratios

The odd/odd case above gives the finite-margin echo family. Because a and b are coprime, the only remaining possibilities are b even (and therefore a odd), or a even (and therefore b odd). Recall that at the reduced ratio $x/L = b/a$ the strict correctness condition is

$$\cos\left(\pi \frac{a - b}{b} X\right) > 0. \tag{123}$$

Write $d = a - b$. Since $\gcd(a, b) = 1$, also $\gcd(d, b) = 1$.

Case 1: b even—zero-margin contact resonances. If b is even, then a and d are odd, so $\gcd(d, 2b) = 1$. Multiplication by d therefore permutes the $2b$ residue classes modulo $2b$. The phase is periodic in X modulo $2b$, and asymptotically it is enough to count the signs of

$$\cos\left(\frac{\pi r}{b}\right), \qquad r = 0, 1, \ldots, 2b - 1. \tag{124}$$

The same roots-of-unity argument as above gives asymptotically uniform residue mass modulo $2b$. Because b is even, there are two exact zeros, at $r = b/2$ and $r = 3b/2$; among the remaining residues, $b - 1$ have positive cosine and $b - 1$ negative cosine. Under the strict rule, the two zeros count as failures. Hence

$$A_0\left(\frac{b}{a}L\right) \longrightarrow \frac{b - 1}{2b} = \frac{1}{2} - \frac{1}{2b}, \qquad b \text{ even}, \tag{125}$$

so

$$s_0\left(\frac{b}{a}L\right) \longrightarrow -\frac{1}{b}, \qquad E_0\left(\frac{b}{a}L\right) \longrightarrow \frac{1}{b^2}, \qquad b \text{ even}. \tag{126}$$

The total probability mass on exact decision boundaries is asymptotically $1/b$. Thus the familiar spikes at $2L, 4L, 6L, \ldots$ are the $a = 1$ members of the general even-b contact family. For example, at $L = 90$, $20/90 = 2/9$, $36/90 = 2/5$, and $60/90 = 2/3$ are all $b = 2$ contacts, while $40/90 = 4/9$, $72/90 = 4/5$, and $120/90 = 4/3$ are $b = 4$ contacts. At finite lengths the residue masses are not exactly uniform, so these limiting energies are asymptotic in general.

Case 2: a even—asymptotically silent rational ratios. If a is even, then b and $d = a - b$ are odd. Again multiplication by d permutes the $2b$ residue classes, but now $\cos(\pi r/b) = 0$ has no integer solution. Exactly b residues have positive cosine and b have negative cosine, giving

$$A_0\left(\frac{b}{a}L\right) \longrightarrow \frac{1}{2}, \qquad s_0\left(\frac{b}{a}L\right) \longrightarrow 0, \qquad E_0\left(\frac{b}{a}L\right) \longrightarrow 0, \qquad a \text{ even}. \tag{127}$$

Together with Eq. (121), these cases exhaust all reduced rational ratios.

Exact silence at every odd test length when L is even. For the experimental setting $L = 90$, the asymptotic a-even statement has a useful exact strengthening. Suppose L is even and x is odd. Pair Hamming weights X and $x - X$. Since L is even,

$$\cos\left(\pi\frac{L}{x}(x - X)\right) = \cos\left(\pi L - \pi\frac{L}{x}X\right) = \cos\left(\pi\frac{L}{x}X\right), \tag{128}$$

so the ideal classifier prediction is unchanged, while the true parity label flips because x is odd. The binomial weights are equal and there is no unpaired central shell. No shell lies exactly on a boundary: $(L - x)X/x = k + 1/2$ would imply an even integer equals an odd integer. Hence each pair contains exactly one correct and one incorrect shell, and

$$L \text{ even}, \quad x \text{ odd} \quad \Longrightarrow \quad A_0(x) = \frac{1}{2}, \qquad s_0(x) = 0, \qquad E_0(x) = 0. \tag{129}$$

This exact symmetry produces the alternating comb of silent odd lengths around an even training length.

Strict boundary convention versus the Fourier midpoint convention. Define the exact boundary mass

$$B_L(x) = \sum_{X=0}^{x} p_x(X)\mathbf{1}\left\{\cos\left(\pi\frac{L - x}{x}X\right) = 0\right\}. \tag{130}$$

Let $q(z) = \mathbf{1}\{\cos z > 0\}$. Its Fourier series is

$$q_F(z) = \frac{1}{2} + \frac{2}{\pi}\sum_{m=0}^{\infty}\frac{(-1)^m}{2m + 1}\cos((2m + 1)z). \tag{131}$$

For $\cos z \neq 0$, $q_F(z) = q(z)$; at a discontinuity the Fourier series converges to $1/2$, whereas the strict indicator equals 0. Therefore

$$A_{\text{strict}}(x) = A_F(x) - \frac{1}{2}B_L(x). \tag{132}$$

For the even-b family, $B_L \to 1/b$, so the Fourier midpoint convention removes the singular strict contact contribution. With a continuous symmetric displacement ϵ of a clean boundary phase z_*, $\Pr[\cos(z_* + \epsilon) > 0] \to 1/2$ as the noise scale tends to zero through positive values. Thus

$$\lim_{\sigma \to 0^+} A_{\text{noisy}}(x; \sigma) = A_{\text{strict}}(x) + \frac{1}{2}B_L(x) = A_F(x). \tag{133}$$

The even-b contacts are therefore singular zero-margin predictions of the exact noiseless lattice: infinitesimal continuous perturbations regularize their boundary mass, whereas finite-margin odd/odd echoes survive and are attenuated smoothly.

Complete reduced-rational classification. For fixed reduced $x/L = b/a$ and increasing common scale,

$$s_0\left(\frac{b}{a}L\right) \longrightarrow \begin{cases} (-1)^{(b-1)/2}/b, & a, b \text{ odd}, \\ -1/b, & b \text{ even}, \\ 0, & a \text{ even}, \end{cases} \qquad E_0 = s_0^2. \tag{134}$$

The three rows correspond respectively to finite-margin echoes, zero-margin contacts, and asymptotically silent rational ratios.

B.3.3 Fourier representation and the general packet-width law

Fourier analysis is not required to obtain the resonance locations or residue-class amplitudes above. Its main use is to describe the local shape of a finite-margin resonance packet and to express phase-noise damping through a characteristic function. Averaging Eq. (131) over $X \sim \mathrm{Binomial}(x, 1/2)$ uses

$$\mathbb{E}[e^{i\alpha X}] = \left(\frac{1+e^{i\alpha}}{2}\right)^x = e^{i\alpha x/2}\cos^x\left(\frac{\alpha}{2}\right), \tag{135}$$

and gives

$$A_F(x) = \frac{1}{2} + \frac{2}{\pi}\sum_{m=0}^{\infty}\frac{(-1)^m}{n_m}\cos^x\left(\frac{n_m\pi(L-x)}{2x}\right)\cos\left(\frac{n_m\pi(L-x)}{2}\right), \qquad n_m = 2m+1. \tag{136}$$

Away from boundary contacts, this agrees with the strict accuracy.
Consider a finite-margin odd/odd resonance

$$x_0 = \frac{b}{a}L, \qquad a, b \text{ odd}, \qquad \gcd(a,b) = 1, \tag{137}$$

and write $x = x_0 + \Delta$. The first resonant odd Fourier harmonic is $n = b$, with phase

$$\theta_b(x) = \frac{b\pi(L-x)}{2x} = \frac{b\pi}{2}\left(\frac{L}{x} - 1\right). \tag{138}$$

At x_0, $\theta_b(x_0) = \pi(a-b)/2$, an integer multiple of π, and

$$\theta_b'(x_0) = -\frac{a\pi}{2x_0}. \tag{139}$$

Thus

$$\theta_b(x) = \theta_b(x_0) - \frac{a\pi}{2x_0}\Delta + O\left(\frac{\Delta^2}{x_0^2}\right). \tag{140}$$

Using $\log|\cos\epsilon| = -\epsilon^2/2 + O(\epsilon^4)$ near zero, the leading signed-signal envelope is

$$|\cos\theta_b(x)|^x \simeq \exp\left(-\frac{a^2\pi^2\Delta^2}{8x_0}\right). \tag{141}$$

Squaring gives the energy envelope

$$E_{\mathrm{env}}(x) \propto \exp\left(-\frac{a^2\pi^2(x-x_0)^2}{4x_0}\right), \tag{142}$$

and hence the full width at half maximum

$$W_{\mathrm{th}} = \frac{4\sqrt{\ln 2}}{\pi}\frac{\sqrt{x_0}}{a} \simeq 1.060\frac{\sqrt{x_0}}{a}. \tag{143}$$

The denominator a therefore narrows the packet linearly, not as $a^{-1/2}$. For the three packets fitted in the main text, $x_0 \in \{L, 3L, 5L\}$, one has $a = 1$, so Eq. (143) reduces to $W_{\mathrm{th}} \simeq 1.060\sqrt{x_0}$. Higher resonant harmonics change the finite packet shape and therefore its fitted width; applying the empirical width estimator to the full noiseless classifier gives much closer agreement with the measured packets (Appendix A.6), while Eq. (143) gives the leading $\sqrt{x_0}/a$ envelope scaling.

B.3.4 Measured phase deformation and the characteristic-function correction

Section 3.6 measures deviations of the post-attention SEP representation from the ideal coordinate in Eq. (41). Let the fitted training line be $\vec{z}_{\text{line}}(u) = \bar{\vec{a}} + u\bar{\vec{v}}$, and for an individual test example of length x, Hamming weight X, and bit arrangement ω, define

$$u(x, X, \omega) = \frac{\bar{\vec{v}}^{\top}(\vec{z}^{(x)}_{\text{SEP}}(X, \omega) - \bar{\vec{a}})}{\|\bar{\vec{v}}\|^2}, \qquad \delta u(x, X, \omega) = u(x, X, \omega) - \frac{LX}{x}. \tag{144}$$

Writing

$$m_x(X) = \mathbb{E}_{\omega}[\delta u \mid x, X], \qquad \eta_x(X, \omega) = \delta u - m_x(X), \tag{145}$$

separates shell-center misplacement from within-shell scatter, with $\mathbb{E}_{\omega}[\eta_x \mid x, X] = 0$.
With this measured deformation, while keeping the ideal midpoint boundary lattice fixed, correctness is exactly

$$\cos\left[\pi\left(\frac{L-x}{x}X + \delta u(x, X, \omega)\right)\right] > 0. \tag{146}$$

Hence the representation-corrected accuracy can be evaluated directly from measured samples,

$$A_{\text{meas}}(x) = \sum_{X=0}^{x} p_x(X)\, \mathbb{E}_{\omega|X}\mathbf{1}\left\{\cos\left[\pi\left(\frac{L-x}{x}X + \delta u\right)\right] > 0\right\}. \tag{147}$$

For analytic work define the conditional characteristic function

$$\phi_{x,X}(t) = \mathbb{E}_{\omega|X}[e^{it\delta u(x,X,\omega)}]. \tag{148}$$

For continuous phase noise, exact boundary hits have probability zero, and

$$A_{\phi}(x) = \frac{1}{2} + \frac{2}{\pi}\sum_{m=0}^{\infty}\frac{(-1)^m}{n_m}\sum_{X=0}^{x} p_x(X)\ \text{Re}\left[e^{in_m\pi(L-x)X/x}\phi_{x,X}(n_m\pi)\right], \qquad n_m = 2m+1. \tag{149}$$

This permits both the mean and spread of δu to depend on Hamming weight.
The one-Gaussian-per-length approximation used in the main text is

$$\delta u \mid x \approx \mathcal{N}(\mu_{\delta u}(x), \sigma^2_{\delta u}(x)), \tag{150}$$

whose characteristic function multiplies the nth odd Fourier harmonic by

$$\exp\left(in\pi\mu_{\delta u}(x) - \frac{1}{2}n^2\pi^2\sigma^2_{\delta u}(x)\right). \tag{151}$$

Thus its magnitude is damped by

$$D_n(x) = \exp\left[-\frac{1}{2}n^2\pi^2\sigma^2_{\delta u}(x)\right]. \tag{152}$$

Combining this with the binomial identity gives

$$A_G(x) = \frac{1}{2} + \frac{2}{\pi}\sum_{m=0}^{\infty}\frac{(-1)^m}{n_m}\exp\left[-\frac{1}{2}n_m^2\pi^2\sigma^2_{\delta u}(x)\right]\cos^x\left(\frac{n_m\pi(L-x)}{2x}\right) \times \cos\left(\frac{n_m\pi(L-x)}{2} + n_m\pi\mu_{\delta u}(x)\right), \qquad n_m = 2m+1. \tag{153}$$

The corresponding signal and energy are $s_G(x) = 2A_G(x) - 1$ and $E_G(x) = s_G(x)^2$. For several trained runs, the experimental statistic is reconstructed run by run,

$$E_{G,\text{mean}}(x) = \mathbb{E}_r[(2A_{G,r}(x) - 1)^2], \tag{154}$$

rather than by averaging accuracies first and squaring afterward.

Why the correction is especially informative at even-b contacts. At an odd/odd resonance, the relevant residue classes have finite margin from the ideal boundaries, so small phase noise changes their probabilities continuously and the Fourier harmonics are smoothly attenuated. At an even-b contact, the boundary mass has exactly zero margin; any nonzero symmetric continuous perturbation splits it between the adjacent classifier regions. Thus the strict contact contribution is regularized toward the Fourier-midpoint value, while robust odd/odd packets such as $3L$ and $5L$ persist and are only gradually damped. The conspicuous $2L, 4L, \ldots$ spikes are simply the $a = 1$ examples of this more general effect.

Scope of the correction. The variable δu is measured before the MLP and therefore quantifies representation-side departures from the ideal phase map. Equations (147)–(153) keep the midpoint decision lattice of Appendix B.1 fixed. A wide learned MLP can in principle shift those decision boundaries or create additional zero-crossing pairs; such classifier-side deviations are not contained in δu. Accordingly, the phase-deformation analysis should be interpreted as a quantitative test of how much of the clean-theory mismatch is accounted for by measured post-attention geometry under the ideal-classifier approximation, rather than as a unique causal decomposition of all errors in the trained network.

C Effective-Kernel Theory for the $L^{3/2}$ Learning-Time Law

C.1 From frozen Transformer features to the ReLU NTK

This appendix supplies the notation and NTK derivation omitted from Section 4.1. We first derive the complete initialization-averaged ReLU NTK conditional on the frozen shell features, then derive the local angular lattice and isolate the solvable cusp sector K_0. Section 4 studies parity learning after the feature-producing part of the Transformer has been frozen. For each retained Hamming-weight shell X, let

$$\vec{x}_X \in \mathbb{R}^{d_m} \tag{155}$$

denote the frozen post-attention SEP representation. Over the central $O(\sqrt{L})$ Hamming-weight window, these features are approximately affine in the normalized shell coordinate

$$t_X = \frac{X - L/2}{L}, \tag{156}$$

so that

$$\vec{x}_X = \vec{V}_6 + t_X \vec{V}_5. \tag{157}$$

No translation or centering of the feature vectors is performed here; $t_X = 0$ simply denotes the central Hamming weight $X = L/2$.

The trainable tail is

$$\vec{p} = W_4 \vec{x}, \qquad \vec{a} = \mathrm{ReLU}(\vec{p}), \qquad \vec{h} = W_5 \vec{a}, \qquad f(\vec{x}) = W_d \vec{h}, \tag{158}$$

with

$$W_4 \in \mathbb{R}^{d_{ff} \times d_m}, \qquad W_5 \in \mathbb{R}^{d_m \times d_{ff}}, \qquad W_d \in \mathbb{R}^{1 \times d_m}. \tag{159}$$

The entries are independently initialized with the LeCun variances

$$\mathrm{Var}(W_4) = \frac{1}{d_m}, \qquad \mathrm{Var}(W_5) = \frac{1}{d_{ff}}, \qquad \mathrm{Var}(W_d) = \frac{1}{d_m}. \tag{160}$$

We condition throughout on the frozen feature vectors $\{\vec{x}_i\}$ and average only over the initialization of W_4, W_5, W_d.

For two retained shells i, j, define

$$P_{ij} = \vec{x}_i^{\top} \vec{x}_j, \qquad q_i = \frac{\|\vec{x}_i\|}{\sqrt{d_m}}, \qquad \cos\theta_{ij} = \frac{P_{ij}}{\|\vec{x}_i\| \, \|\vec{x}_j\|}. \tag{161}$$

For a single row $\vec{w}$ of W_4,

$$z_i = \vec{w}^{\top} \vec{x}_i, \qquad z_j = \vec{w}^{\top} \vec{x}_j \tag{162}$$

are jointly Gaussian, with variances q_i^2, q_j^2 and correlation $\cos\theta_{ij}$.
The two standard ReLU Gaussian averages needed below are

$$A_0(\theta) = \mathbb{E}[\mathbf{1}_{z_i>0}\mathbf{1}_{z_j>0}] = \frac{\pi-\theta}{2\pi}, \tag{163}$$

and

$$A_1(\theta) = \frac{\sin\theta + (\pi-\theta)\cos\theta}{2\pi}, \tag{164}$$

so that

$$\mathbb{E}[\mathrm{ReLU}(z_i)\,\mathrm{ReLU}(z_j)] = q_i q_j A_1(\theta_{ij}). \tag{165}$$

The NTK is the sum of the gradient inner products from the three trainable parameter blocks,

$$K^R = K_{W_d} + K_{W_5} + K_{W_4}. \tag{166}$$

For W_d, differentiating removes the final readout and leaves $\vec{h}$. Averaging over W_5 gives

$$K_{W_d}(i,j) = d_m q_i q_j A_1(\theta_{ij}). \tag{167}$$

For W_5, the gradient contains W_d and the ReLU activations. The LeCun scaling of W_d cancels the d_m output dimension, while summing over the d_{ff} hidden units gives

$$K_{W_5}(i,j) = d_{ff} q_i q_j A_1(\theta_{ij}). \tag{168}$$

Finally, differentiating through W_4 introduces the ReLU gate,

$$\frac{d}{dz}\,\mathrm{ReLU}(z) = \mathbf{1}_{z>0}, \tag{169}$$

and an inner product of the input features. With the initialization in Eq. (160),

$$K_{W_4}(i,j) = P_{ij} A_0(\theta_{ij}). \tag{170}$$

Combining Eqs. (167)–(170) gives the complete initialization-averaged ReLU NTK conditional on the frozen shell features,

$$K^R(i,j) = C_N q_i q_j A_1(\theta_{ij}) + P_{ij} A_0(\theta_{ij}), \qquad C_N = d_m + d_{ff}. \tag{171}$$

This is an initialization-averaged NTK conditional on the frozen SEP features. It is not the realized finite-width NTK of one particular initialization.

C.2 Local angular spacing of the affine feature line

Although the shell centers are equally spaced in the affine coordinate t, their polar angles need not be exactly equally spaced. Let

$$\vec{x}(t) = \vec{V}_6 + t\vec{V}_5 \tag{172}$$

and let $\phi(t)$ denote its polar angle inside the two-dimensional plane spanned by $\vec{V}_6$ and $\vec{V}_5$. The angular velocity is

$$|\phi'(t)| = \frac{\sqrt{\|\vec{x}(t)\|^2\|\vec{V}_5\|^2 - [\vec{x}(t)^\top \vec{V}_5]^2}}{\|\vec{x}(t)\|^2}. \tag{173}$$

At the central shell,

$$\gamma := |\phi'(0)| = \frac{\sqrt{\|\vec{V}_6\|^2\|\vec{V}_5\|^2 - (\vec{V}_6^\top \vec{V}_5)^2}}{\|\vec{V}_6\|^2}. \tag{174}$$

Adjacent Hamming weights differ by $\Delta t = 1/L$. Therefore at the center

$$\phi(1/L) - \phi(0) = \frac{\gamma}{L} + O(L^{-2}). \tag{175}$$

Across the entire $O(\sqrt{L})$ central window, $\phi'(t)$ itself varies smoothly by $O(L^{-1/2})$. Thus the exact angular grid has a slowly varying spacing, while the solvable reference problem replaces it by the uniform lattice

$$\phi_i = \phi_c + \left(i - \frac{M-1}{2}\right)\delta, \qquad \delta = \frac{\gamma}{L}. \tag{176}$$

Consequently,

$$\theta_{ij} = \delta|i-j| \tag{177}$$

in the reference model. Equal angular spacing is therefore an explicit local idealization, rather than an additional property imposed on the original feature vectors.

C.3 Exact decomposition of the ReLU kernel around K_0

We now isolate the part of Eq. (171) responsible for the linear angular cusp. At the center of the retained window define

$$P_0 = \|\vec{x}(0)\|^2 = \|\vec{V}_6\|^2. \tag{178}$$

Because

$$A_0(\theta) = \frac{1}{2} - \frac{\theta}{2\pi}, \tag{179}$$

the central part of the W_4 tangent kernel is

$$K_0(i,j) = P_0\left(\frac{1}{2} - \frac{\theta_{ij}}{2\pi}\right). \tag{180}$$

Under the equal-angle lattice,

$$K_0(i,j) = P_0\left[\frac{1}{2} - \frac{\delta}{2\pi}|i-j|\right]. \tag{181}$$

To see precisely what has been discarded, it is useful to rewrite A_1. Define

$$\Psi(\theta) = \sin\theta - \theta\cos\theta. \tag{182}$$

Then

$$A_1(\theta) = \frac{1}{2}\cos\theta + \frac{1}{2\pi}\Psi(\theta), \tag{183}$$

with

$$\Psi(\theta) = \frac{\theta^3}{3} + O(\theta^5). \tag{184}$$

Let

$$\kappa^2 = \frac{1}{d_m}, \qquad r_i = \|\vec{x}_i\|. \tag{185}$$

Since $q_i q_j \cos\theta_{ij} = \kappa^2 P_{ij}$, Eq. (171) can be written exactly as

$$K^R = K_0 + F + E_{\mathrm{bil}} + E_{\mathrm{sm}}, \tag{186}$$

where

$$F_{ij} = \frac{1}{2}\left[(1 + C_N\kappa^2)P_{ij} - P_0\right], \tag{187}$$

$$(E_{\mathrm{bil}})_{ij} = -\frac{P_{ij} - P_0}{2\pi}\,\theta_{ij}, \tag{188}$$

and

$$(E_{\mathrm{sm}})_{ij} = \frac{C_N\kappa^2}{2\pi} r_i r_j \Psi(\theta_{ij}). \tag{189}$$

This decomposition makes the special role of K_0 explicit. First, F is finite rank in the affine model. Since every feature has the form $\vec{x}_i = \vec{V}_6 + t_i\vec{V}_5$, the Gram matrix P has rank at most two. Adding the constant matrix proportional to P_0 increases the rank by at most one, and therefore

$$\operatorname{rank} F \le 3. \tag{190}$$

Second, E_{bil} contains the deviation $P_{ij} - P_0$. On a central window containing $M = O(\sqrt{L})$ shells,

$$t_i = O(M/L) = O(M^{-1}), \tag{191}$$

so generically $P_{ij} - P_0 = O(M^{-1})$. The largest angular separation inside the same window is

$$\theta_{ij} = O(M\delta) = O(M^{-1}), \tag{192}$$

because $\delta = O(M^{-2})$. Thus

$$(E_{\mathrm{bil}})_{ij} = O(M^{-2}). \tag{193}$$

Finally, E_{sm} begins at cubic order in the angular separation, $\Psi(\theta) = O(\theta^3)$, and therefore

$$(E_{\mathrm{sm}})_{ij} = O(M^{-3}) \tag{194}$$

across the central window.

The key qualitative distinction is therefore that K_0 contains an exact $|i-j|$ cusp, while the remaining non-low-rank ReLU contribution is smoother in the angular coordinate. For completeness, one can sharpen this observation at the operator level. After multiplying by the uniform probability weight $1/M$, the high-rank singular values of the two correction terms obey bounds of the form

$$s_j(E_{\mathrm{bil}}/M) \lesssim M^{-2} j^{-2}, \qquad s_j(E_{\mathrm{sm}}/M) \lesssim M^{-3} j^{-2}. \tag{195}$$

Together with the fixed rank of F, a min–max argument gives stability of the M^{-3} spectral scale throughout any proportional bulk window $k/M \in [c_1, c_2] \subset (0, 1)$.

This statement should not be confused with a mode-by-mode perturbation theorem at the last few Nyquist-edge eigenvectors. Their eigenvalue spacings are much smaller: for fixed r, s,

$$|\lambda^{(0)}_{M-r} - \lambda^{(0)}_{M-s}| = O(M^{-5}). \tag{196}$$

Consequently, small corrections can rotate neighboring edge eigenvectors substantially even when they preserve the parity-relevant spectral scale. This is why the controlled comparison in Section 4.4 is made at the level of the resulting learning dynamics rather than by requiring one-to-one equality of individual edge eigenvectors.

C.4 K_0 IS THE GREEN KERNEL OF A NEUMANN PATH LAPLACIAN

The reference kernel in Eq. (181) can be diagonalized explicitly. Let L_N denote the M-vertex Neumann path Laplacian,

$$L_N = \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}. \tag{197}$$

Also define

$$\vec{e} = \vec{e}_0 + \vec{e}_{M-1}, \qquad \alpha = \frac{1}{2\pi - (M-1)\delta}. \tag{198}$$

Provided the angular window occupies less than a full 2π turn,

$$(M-1)\delta < 2\pi, \tag{199}$$

one obtains the exact identity

$$\left(\delta^{-1} L_N + \alpha \vec{e}\vec{e}^{\top}\right) K_0 = \frac{P_0}{\pi} I. \tag{200}$$

The origin of this identity is elementary. Away from $i = j$, the discrete second difference of $|i-j|$ is zero. At $i = j$,

$$2K_0(i,i) - K_0(i-1,i) - K_0(i+1,i) = \frac{P_0 \delta}{\pi}. \tag{201}$$

The only remaining defects occur at the two endpoints; the rank-one term $\alpha\vec{e}\vec{e}^{\top}$ corrects them because

$$K_0(0,j) + K_0(M-1,j) = \frac{P_0}{2\pi}[2\pi - (M-1)\delta] \tag{202}$$

is independent of j.
Equation (200) also shows that K_0 is positive definite. L_N is positive semidefinite with the constant vector as its only null direction, while the positive rank-one endpoint term has nonzero overlap with that direction.

C.5 Exact parity sector for even M

Let reflection of the path be

$$(Rf)_i = f_{M-1-i}. \tag{203}$$

For even M, the alternating parity vector

$$y_i = (-1)^i \tag{204}$$

is reflection odd,

$$Ry = -y. \tag{205}$$

Any reflection-odd vector satisfies $f_{M-1} = -f_0$, and hence $\vec{e}^{\top} f = f_0 + f_{M-1} = 0$. The endpoint correction in Eq. (200) therefore vanishes identically in the parity sector. On that sector,

$$K_0 = \frac{P_0\delta}{\pi} L_N^{-1}. \tag{206}$$

The Neumann eigenfunctions, normalized in the population inner product

$$\langle f, g\rangle_M = \frac{1}{M}\sum_{i=0}^{M-1} f_i g_i, \tag{207}$$

are $\psi_0(i) = 1$ and, for $1 \le k \le M-1$,

$$\psi_k(i) = \sqrt{2}\cos\left[\frac{\pi k}{M}\left(i + \frac{1}{2}\right)\right]. \tag{208}$$

They satisfy

$$L_N\psi_k = \mu_k\psi_k, \qquad \mu_k = 4\sin^2\left(\frac{\pi k}{2M}\right). \tag{209}$$

Reflection acts as

$$R\psi_k = (-1)^k\psi_k, \tag{210}$$

so the reflection-odd sector consists precisely of odd k.
The learning operator for the uniform shell measure is

$$A_0 = \frac{1}{M}K_0. \tag{211}$$

Therefore, in the parity sector,

$$\lambda_k^{(0)} = \frac{P_0\delta}{4\pi M}\csc^2\left(\frac{\pi k}{2M}\right), \qquad k \text{ odd}. \tag{212}$$

This is the parity-sector eigenvalue formula used in the main text.
The exact squared parity coefficients are also available in closed form. Direct summation gives

$$|\langle y, \psi_k\rangle_M|^2 = \frac{2}{M^2}\sec^2\left(\frac{\pi k}{2M}\right), \qquad k \text{ odd}, \tag{213}$$

and zero for even k. The coefficients satisfy

$$\sum_{\substack{1\le k\le M-1\\ k \text{ odd}}} |\langle y, \psi_k\rangle_M|^2 = 1. \tag{214}$$

Thus the parity target is represented exactly by the odd Neumann modes.

C.6 THE NYQUIST EDGE CARRIES ORDER-ONE PARITY MASS

To study the most rapidly alternating modes, write $k = M - r$, with fixed odd r. Equation (213) becomes

$$|\langle y, \psi_{M-r}\rangle_M|^2 = \frac{2}{M^2}\csc^2\Big(\frac{\pi r}{2M}\Big). \tag{215}$$

For fixed r and $M \to \infty$,

$$|\langle y, \psi_{M-r}\rangle_M|^2 \longrightarrow \frac{8}{\pi^2 r^2}. \tag{216}$$

In particular, the highest-frequency mode $r = 1$ alone carries asymptotic mass

$$\frac{8}{\pi^2} \simeq 0.811. \tag{217}$$

The important point, however, is not that all parity mass collapses onto a single eigenvector. Rather,

$$\sum_{\substack{r\geq 1\\ r\ \mathrm{odd}}} \frac{8}{\pi^2 r^2} = 1, \tag{218}$$

so an order-one fraction of the target remains distributed over a finite collection of Nyquist-edge modes.

Using Eq. (212),

$$\lambda^{(0)}_{M-r} = \frac{P_0\delta}{4\pi M}\sec^2\Big(\frac{\pi r}{2M}\Big). \tag{219}$$

For the central window used in the asymptotic calculation,

$$M = 2\sqrt{L}, \qquad \delta = \frac{\gamma}{L} = \frac{4\gamma}{M^2}. \tag{220}$$

Hence

$$\lambda^{(0)}_{M-r} = \frac{P_0\gamma}{\pi M^3}\sec^2\Big(\frac{\pi r}{2M}\Big) = \frac{P_0\gamma}{\pi M^3}[1 + O(M^{-2})]. \tag{221}$$

Thus every fixed-r mode carrying order-one asymptotic parity mass has eigenvalue of order M^{-3}. This is the spectral origin of the learning-time law.

C.7 EXACT FINITE-M LEARNING CURVE AND THE $L^{3/2}$ LIMIT

Under the $f_0 = 0$ analytical reference defined in Section 4.1, we use the squared-loss gradient-flow convention of Section 4. If A is the population kernel operator, the residual along an eigenmode of eigenvalue λ decays as $e^{-2\eta\lambda t}$, so its contribution to squared error decays as

$$e^{-4\eta\lambda t}. \tag{222}$$

For even M, substituting Eqs. (215) and (219) gives the exact finite-M K_0 learning curve

$$E_M(t) = \sum_{\substack{1\leq r\leq M-1\\ r\ \mathrm{odd}}} \frac{2}{M^2}\csc^2\Big(\frac{\pi r}{2M}\Big)\exp\Big[-\frac{\eta P_0\delta}{\pi M}t\sec^2\Big(\frac{\pi r}{2M}\Big)\Big]. \tag{223}$$

Using $\delta = 4\gamma/M^2$, introduce the dimensionless time

$$s = \frac{4\eta P_0\gamma}{\pi M^3}t. \tag{224}$$

For a desired error level $0 < \epsilon < 1$, define $s_{\epsilon,M}$ as the unique solution of

$$\sum_{\substack{1\leq r\leq M-1\\ r\ \mathrm{odd}}} \frac{2}{M^2}\csc^2\Big(\frac{\pi r}{2M}\Big)\exp\Big[-s_{\epsilon,M}\sec^2\Big(\frac{\pi r}{2M}\Big)\Big] = \epsilon. \tag{225}$$

The left-hand side decreases continuously from 1 to 0, so the solution exists and is unique. At finite M, $s_{\epsilon,M}$ is most conveniently obtained by one-dimensional bisection.

The exact threshold time is therefore

$$T_\epsilon(M) = \frac{\pi M^3}{4\eta P_0 \gamma} s_{\epsilon,M}. \tag{226}$$

For every fixed odd r, $\sec^2(\pi r/(2M)) \to 1$ and

$$\frac{2}{M^2} \csc^2\Big(\frac{\pi r}{2M}\Big) \longrightarrow \frac{8}{\pi^2 r^2}. \tag{227}$$

Because the limiting coefficients sum to one, Eq. (225) tends to $e^{-s_\epsilon} = \epsilon$. Hence

$$s_{\epsilon,M} \longrightarrow \log \frac{1}{\epsilon}. \tag{228}$$

Combining this with Eq. (226) gives

$$T_\epsilon(M) = \frac{\pi M^3}{4\eta P_0 \gamma} \log \frac{1}{\epsilon} [1 + o(1)]. \tag{229}$$

Since $M = 2\sqrt{L} + O(1)$, we have $M^3 = 8L^{3/2}[1 + o(1)]$, and therefore

$$T_\epsilon(L) = \frac{2\pi}{\eta P_0 \gamma} L^{3/2} \log \frac{1}{\epsilon} [1 + o(1)]. \tag{230}$$

This is Eq. (25) of the main text.

The logarithm in Eq. (230) does not require the approximation that all parity weight occupies only the single $k = M - 1$ eigenvector. Instead, all fixed-r edge modes become asymptotically degenerate after the common M^{-3} rescaling, while their limiting parity weights sum to one. The complete Nyquist packet therefore behaves asymptotically as one effective exponential.

C.8 Even M in the controlled closed-form sweep

The controlled comparison uses odd sequence lengths $L = 17, 19, \ldots, 47$ and the integer shell set

$$I_L = \left\{ k \in \{0, \ldots, L\} : \left| k - \frac{L}{2} \right| \le \sqrt{L} \right\}. \tag{231}$$

Since $L/2$ is then a half-integer, the retained shells occur in reflection pairs and

$$M(L) = 2 \left\lfloor \sqrt{L} + \frac{1}{2} \right\rfloor \tag{232}$$

is even. The alternating target is therefore reflection odd and the closed-form parity-sector formulas in Appendices C.5–C.7 apply directly. The finite K_0 matrix is also diagonalized numerically as an implementation check; this numerical prediction agrees with the closed-form prediction to machine precision and is not the curve displayed in Figure 4. Appendix D gives the exact finite-L implementation.

C.9 Why the controlled experiment uses Jump-ReLU

The analytical derivation above uses ordinary ReLU because it makes the origin of the A_0 cusp especially transparent. The controlled experiment of Section 4.4 uses instead

$$\phi_J(z) = z_+ + \mathbf{1}_{z>0}. \tag{233}$$

The derivative of the added step is zero almost everywhere, so the W_4 tangent contribution remains

$$P_{ij} A_0(\theta_{ij}). \tag{234}$$

The forward covariance changes. For jointly Gaussian preactivations,

$$\Sigma_J(i,j) = q_i q_j A_1(\theta_{ij}) + (q_i + q_j) A_{01}(\theta_{ij}) + A_0(\theta_{ij}), \tag{235}$$

where

$$A_{01}(\theta) = \frac{1+\cos\theta}{2\sqrt{2\pi}}. \tag{236}$$

The full initialization-averaged Jump-ReLU NTK is therefore

$$K^J(i,j) = C_N \Sigma_J(i,j) + P_{ij} A_0(\theta_{ij}). \tag{237}$$

Unlike ordinary ReLU, the forward covariance itself now contains an A_0 term. Consequently the coefficient of the solvable cusp becomes

$$c_0^J = C_N + P_0, \tag{238}$$

and the corresponding reference kernel is

$$K_0^J(i,j) = (C_N + P_0)\left[\frac{1}{2} - \frac{\delta}{2\pi}|i-j|\right]. \tag{239}$$

For ordinary ReLU,

$$c_0^R = P_0. \tag{240}$$

Thus the two activations have the same solvable Green-kernel shape and the same M^{-3} parity-edge spectrum, but Jump-ReLU multiplies that sector by the larger coefficient $C_N + P_0$. The corresponding asymptotic times differ only through this coefficient:

$$T_\epsilon^R \sim \frac{\pi M^3}{4\eta P_0 \gamma} \log\frac{1}{\epsilon}, \qquad T_\epsilon^J \sim \frac{\pi M^3}{4\eta (C_N + P_0)\gamma} \log\frac{1}{\epsilon}. \tag{241}$$

For the same frozen geometry,

$$\frac{T_\epsilon^R}{T_\epsilon^J} \sim \frac{C_N + P_0}{P_0}. \tag{242}$$

This is the sense in which Jump-ReLU preserves the same spectral mechanism while making the controlled experiment substantially faster.

It also clarifies the parameter-block origin of K_0. With ordinary ReLU, the $P_0 A_0$ cusp comes entirely from the W_4 derivative. With Jump-ReLU, the solvable A_0 sector receives contributions from both the W_4 tangent term and the W_5/W_d forward-covariance blocks.

C.10 Controlled-validation protocol and prediction

The purpose of the controlled experiment is to isolate the spectral mechanism above in a finite-width trained network. The feature-producing module consists of a single frozen attention head with no positional encoding. Its initialization is filtered so that the attention factors for zero and one tokens are comparable and the corresponding value directions are nondegenerate. Once an accepted feature extractor is obtained, it is shared across the repeated runs, while only W_4, W_5, W_d are freshly initialized and trained.

The finite-width model uses

$$d_m = 320, \qquad d_{ff} = 10{,}000, \tag{243}$$

and Jump-ReLU activation. Training uses full-batch SGD with

$$\eta = 5 \times 10^{-6}. \tag{244}$$

For each sequence length L, training uses the uniformly weighted shell set

$$I_L = \left\{ k \in \{0,\dots,L\} : \left| k - \frac{L}{2} \right| \le \sqrt{L} \right\}, \qquad M(L) = |I_L|. \tag{245}$$

The main figure displays the odd lengths $L = 17, 19, \dots, 47$, for which

$$M(L) = 2\left\lfloor \sqrt{L} + \frac{1}{2} \right\rfloor \tag{246}$$

is even. Ten independent finite-width runs are performed at each displayed length, using the same frozen feature extractor and fresh trainable-tail initializations. Training stops when the MSE first falls below 0.2.

The effective-kernel prediction is obtained independently of the measured stopping times. From the two central frozen shell features, the finite-L geometry supplies P_0 and the local angular step δ; equivalently $\gamma = L\delta$. For Jump-ReLU,

$$c_0^J = C_N + P_0 = d_m + d_{ff} + P_0. \tag{247}$$

Because M is even, the displayed prediction uses the exact Neumann eigenvalues and parity projections,

$$\lambda_k = \frac{c_0^J \delta}{4\pi M} \csc^2\left(\frac{\pi k}{2M}\right), \qquad a_k^2 = \frac{2}{M^2} \sec^2\left(\frac{\pi k}{2M}\right), \qquad k \text{ odd}. \tag{248}$$

The predicted convergence time is the unique root of

$$\sum_{k \text{ odd}} a_k^2 e^{-4\eta\lambda_k t} = 0.2, \tag{249}$$

solved by one-dimensional bisection. Numerical diagonalization of the finite K_0 matrix is retained only as an independent consistency check and is not the source of the curve displayed in Figure 4. There is no fitted multiplicative time constant in the comparison. Appendix D gives the exact finite-M implementation and quantitative checks.

The Figure 4 comparison is reported over $17 \le L \le 47$. The sweep continues to larger lengths in Appendix D, where run-to-run variability grows and the finite-width means depart more strongly from the fixed-K_0 curve; the timing data alone do not establish the cause of that longer-length breakdown.

C.11 Scope of the effective-kernel claim

The reduction to K_0 is intentionally narrower than the statement

$$K^R \simeq K_0 \tag{250}$$

as full matrices. That stronger statement is neither required nor generally true.

The exact decomposition in Appendix C.3 shows that the full kernel contains finite-rank and smoother terms that can change individual eigenvectors and finite-size eigenvalues. In particular, the last few eigenvalues of K_0 become nearly degenerate: their separations are $O(M^{-5})$. Perturbations smaller than the $O(M^{-3})$ eigenvalue scale can therefore still rotate individual edge modes.

The quantity relevant to parity learning is instead the parity-weighted spectral measure,

$$d\nu_y(\lambda) = \sum_a |\langle y, v_a\rangle_M|^2 \, \delta(\lambda - \lambda_a), \tag{251}$$

because the learning curve is

$$E(t) = \int e^{-4\eta\lambda t} \, d\nu_y(\lambda). \tag{252}$$

The effective-kernel claim is consequently that the linear angular cusp captured by K_0 determines the dominant parity-relevant spectral scale,

$$\lambda_{\text{parity}} \sim M^{-3}, \tag{253}$$

and therefore the observed

$$T(L) \sim L^{3/2} \tag{254}$$

learning-time law. It does not require one-to-one agreement between every eigenvector of K_0 and every eigenvector of the full finite-width kernel.

This distinction is particularly important at the Nyquist edge, where substantial redistribution among neighboring eigenmodes can coexist with a nearly unchanged parity-weighted learning curve.

D CLOSED-FORM FINITE-M VALIDATION OF THE EFFECTIVE KERNEL

D.1 PURPOSE AND RELATION TO APPENDIX C

Appendix C derives the effective kernel K_0, its Neumann-path eigensystem, the parity spectral weights, and the asymptotic $L^{3/2}$ learning-time law. Figure 4 uses the closed-form finite-M spectrum itself rather than a numerical diagonalization of the K_0 matrix. This appendix records exactly how that finite-M prediction is evaluated in the controlled experiment and separates three objects: the finite-width trained network, the idealized finite-M effective kernel K_0, and the exact analytical eigensystem of K_0. No convergence-time data are used to fit the analytical prediction. The only length-dependent quantities entering the prediction are the shell count and the local geometry measured from the frozen feature extractor.

D.2 CENTRAL SHELL SET AND WHY THE DISPLAYED LENGTHS ARE ODD

For sequence length L, training uses a uniform distribution over the integer Hamming-weight shells

$$I_L = \left\{ k \in \{0, \dots, L\} : \left| k - \frac{L}{2} \right| \le \sqrt{L} \right\}, \qquad M(L) = |I_L|. \tag{255}$$

The main-text figure displays

$$L = 17, 19, \dots, 47. \tag{256}$$

For odd L, the center $L/2$ is a half-integer, so the retained integer shells occur in reflection pairs. In this case

$$M(L) = 2 \left\lfloor \sqrt{L} + \frac{1}{2} \right\rfloor, \tag{257}$$

which is always even. The alternating parity target is therefore reflection odd on the retained path, and the exact closed-form parity-sector formulas from Appendix C apply without an odd-M correction. Across the displayed range, M takes the values $8, 10, 12$, and 14. Its discrete changes produce the visible finite-size steps in the theoretical curve. Asymptotically,

$$M(L) = 2\sqrt{L} + O(1), \tag{258}$$

so the finite-size prediction has the same $L^{3/2}$ envelope derived in Appendix C.

D.3 GEOMETRY CALIBRATION FROM THE FROZEN FEATURE EXTRACTOR

Let $\vec{x}_k \in \mathbb{R}^{d_m}$ denote the frozen post-W_O SEP feature for Hamming-weight shell k. The controlled model uses one frozen attention head with no positional encoding in the active forward pass. The frozen feature extractor is shared across repeated runs, while the trainable tail W_4, W_5, W_d is freshly initialized for each run.

For the odd lengths used in Figure 4, define the two central shells

$$k_- = \frac{L-1}{2}, \qquad k_+ = \frac{L+1}{2}. \tag{259}$$

The code calibrates the central squared radius by the symmetric average

$$P_0 = \frac{1}{2} \left(\|\vec{x}_{k_-}\|^2 + \|\vec{x}_{k_+}\|^2 \right), \tag{260}$$

and measures the local angular step directly as

$$\delta = \arccos\left(\frac{\vec{x}_{k_-}^{\top} \vec{x}_{k_+}}{\|\vec{x}_{k_-}\| \, \|\vec{x}_{k_+}\|} \right). \tag{261}$$

Thus the quantity called γ in the asymptotic derivation is represented at finite L by

$$\gamma = L\delta. \tag{262}$$

This is a direct local finite-difference calibration of the frozen geometry, not a fit to optimizer times and not an affine-line regression.

The controlled experiment uses Jump-ReLU,

$$\phi_J(z) = z_+ + \mathbf{1}_{z>0}. \tag{263}$$

With frozen W_O, LeCun initialization of the trainable tail, and trainable W_4, W_5, W_d, the cusp coefficient is

$$c_0^J = d_m + d_{ff} + P_0. \tag{264}$$

For the experiment, $d_m = 320$ and $d_{ff} = 10{,}000$.

D.4 Exact finite-M spectral prediction

On the retained shell set, the idealized effective kernel is

$$K_0(i,j) = c_0^J \left[\frac{1}{2} - \frac{\delta}{2\pi}|i-j|\right], \qquad i,j = 0,\dots,M-1. \tag{265}$$

The population measure is uniform, so the learning operator is $A_0 = K_0/M$. Because M is even, parity couples only to the odd Neumann modes

$$k = 1, 3, \dots, M-1. \tag{266}$$

Their exact operator eigenvalues are

$$\lambda_k^{(0)} = \frac{c_0^J \delta}{4\pi M} \csc^2\left(\frac{\pi k}{2M}\right), \qquad k \text{ odd}, \tag{267}$$

and the exact squared parity projections are

$$a_k^2 = \frac{2}{M^2} \sec^2\left(\frac{\pi k}{2M}\right), \qquad k \text{ odd}, \tag{268}$$

with $\sum_{k \text{ odd}} a_k^2 = 1$.

Using the same squared-loss gradient-flow convention as Appendix C, the exact finite-M learning curve of the effective theory is

$$E_0(t) = \sum_{k \text{ odd}} a_k^2 \exp[-4\eta\lambda_k^{(0)} t]. \tag{269}$$

The plotted theoretical time is the unique solution of

$$E_0(T_{0.2}) = 0.2. \tag{270}$$

The eigensystem in Eqs. (267)–(268) is closed form. The final scalar root in Eq. (270) is found by monotone one-dimensional bisection because a finite sum of exponentials has no simpler elementary inverse in general. This procedure never constructs or diagonalizes the K_0 matrix.

Equivalently, with $k = M - r$ and odd r,

$$E_0(t) = \sum_{\substack{1 \le r \le M-1 \\ r \text{ odd}}} \frac{2}{M^2} \csc^2\left(\frac{\pi r}{2M}\right) \exp\left[-\frac{\eta c_0^J \delta}{\pi M} \sec^2\left(\frac{\pi r}{2M}\right) t\right]. \tag{271}$$

In the idealized limit $M = 2\sqrt{L}$ and $\delta = 4\gamma/M^2$, this reduces to the asymptotic result derived in Appendix C,

$$T_\epsilon(L) = \frac{2\pi}{\eta c_0^J \gamma} L^{3/2} \log\frac{1}{\epsilon}\,[1 + o(1)]. \tag{272}$$

The main-text curve uses the exact finite-M sum, not this asymptotic envelope.

D.5 Independent numerical check of the closed-form implementation

As a code-level validation, the same K_0 matrix in Eq. (265) was also constructed explicitly, weighted by the same uniform shell measure, and diagonalized numerically. This route is not used for the displayed theoretical curve; it is only an independent check. Across all displayed lengths $17 \le L \le 47$, the numerical-diagonalization time and the closed-form time agree to essentially machine precision. The maximum relative discrepancy is

$$\max_L \frac{|T_{0.2}^{\mathrm{num}} - T_{0.2}^{\mathrm{closed}}|}{T_{0.2}^{\mathrm{closed}}} \approx 2.0 \times 10^{-13}. \tag{273}$$

This verifies the finite-M eigenvalue normalization, parity projections, and $e^{-4\eta\lambda t}$ loss convention used in the closed-form implementation.

D.6 Quantitative comparison over the displayed range

Figure 4 compares the closed-form prediction to ten finite-width runs at each displayed odd length. The optimizer is full-batch SGD with learning rate

$$\eta = 5 \times 10^{-6}, \tag{274}$$

and training stops when the training MSE first falls below 0.2. The table reports the mean and run-to-run standard deviation together with the closed-form prediction.

L	M	runs	train mean	train s.d.	K_0 closed form
17	8	10	6,188	396	6,394
19	8	10	7,146	508	7,143
21	10	10	9,622	1,294	10,127
23	10	10	10,499	714	11,089
25	10	10	11,113	683	12,051
27	10	10	12,283	904	13,013
29	10	10	12,762	727	13,976
31	12	10	16,867	1,372	18,230
33	12	10	18,584	1,035	19,405
35	12	10	19,063	1,359	20,580
37	12	10	20,970	1,716	21,754
39	12	10	22,307	3,572	22,929
41	12	10	23,820	2,608	24,104
43	14	10	29,412	1,987	29,842
45	14	10	35,216	3,007	31,229
47	14	10	36,130	4,280	32,616

Over these 16 displayed lengths, the mean absolute percentage difference between the closed-form prediction and the empirical mean is 5.55%, the maximum absolute percentage difference is 11.32%, and the prediction lies within one empirical standard deviation of the mean at 12 of the 16 lengths. These are descriptive summaries of the controlled experiment, not fitted quantities.

D.7 Longer-length runs and scope of the main-text comparison

The sweep was continued beyond the range displayed in the main figure. Those later runs show sharply increased run-to-run variability and a growing discrepancy from the fixed-K_0 prediction. They are reported here so that the main-text cutoff is explicit rather than implicit.

L	M	runs	train mean	train s.d.	K_0 closed form
49	14	10	39,118	5,379	34,003
51	14	10	39,549	6,298	35,390
53	14	10	56,704	26,889	36,777
55	14	10	110,271	47,903	38,164
57	16	10	80,644	37,385	45,598
59	16	10	122,581	44,569	47,197
61	16	7	129,552	36,185	48,796

The present analysis does not identify the cause of this longer-length breakdown. Possibilities such as finite-width effects, departure from fixed-kernel dynamics, or other optimization effects require separate diagnostics and are not inferred from the timing data alone. Accordingly, Figure 4 is a controlled quantitative comparison over the displayed range $17 \leq L \leq 47$, not evidence that the fixed-K_0 approximation remains accurate for arbitrarily large L.

D.8 What Figure 4 establishes

The theoretical curve in Figure 4 is stronger than a fitted $L^{3/2}$ line: at each displayed length it uses the actual integer shell count $M(L)$, the locally measured frozen-feature geometry (P_0, δ), the architecture-fixed Jump-ReLU coefficient c_0^J, and the exact finite-M Neumann spectrum and parity weights. No multiplicative timing constant is fitted to the finite-width runs.

The comparison therefore tests whether the analytically isolated cusp sector predicts the absolute learning-time scale of the controlled finite network over the reported range. Its asymptotic significance remains the mechanism developed in Appendix C: parity places order-one spectral mass near the Nyquist edge, those effective-kernel eigenvalues scale as M^{-3}, and $M = \Theta(\sqrt{L})$ gives the $L^{3/2}$ learning-time law.