

Predicting Neural Scaling Laws without Training: A Data Manifold Oracle

Hanlin Zhang¹ Jikai Jin² Sham Kakade¹

¹Harvard University ²Stanford University

{hanlinzhang, sham}@g.harvard.edu

Abstract

Neural scaling laws provide a critical compass for large-scale pretraining, yet estimating data scaling exponents traditionally requires fitting dozens of compute-intensive proxy models, an unsustainable bottleneck for data curation and resource allocation. We propose the Data Manifold Oracle (DMO), a training-free framework that predicts neural data scaling trajectories directly from the static, lossless compression properties of raw text. By linking the sequence-length scaling behavior of DEFLATE compression to neural loss decay, DMO forecasts the data scaling exponent α_D before any model is trained on the candidate corpus. Moreover, we pair the empirical predictor with a rigorous theoretical boundary, proving that symbol-level compressibility cannot uniquely resolve the dimension of a continuous source, and showing that it is nonetheless an accurate empirical proxy for neural learning dynamics. A calibration frozen on 38 pretraining corpora predicts the exponents of five held-out corpus families to an MAE of 0.029 without GPU overhead, against 0.084 for a proxy model trained on the candidate corpus itself. By turning scaling law estimation from a multi-GPU training challenge into a data-profiling task, DMO provides a scalable foundation for lightweight data selection and pretraining optimization.

1 Introduction

A raw-text compression curve can predict how quickly a language model benefits from more training data. The Data Manifold Oracle (DMO) measures how compressed bits per character change with window length, then uses a fixed calibration to predict a corpus’s data-scaling exponent. The manifold in the name and the curve that does the predicting have separate jobs. Geometry says what a symbol-only measurement can never determine; the compression curve, calibrated once against trained grids, is what we test as a predictor of the fitted data exponent. On five evaluable new parent families measured before their training outcomes existed, this prediction reduces mean absolute error by 40% relative to a fixed-intercept baseline and 36% relative to a domain mean. Applying it requires only the candidate text.

The quantity being predicted has a direct training interpretation. Scaling-law experiments (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022) fit how loss changes with parameter count P and training-token budget D . At fixed model size, the data dependence takes the form

$$L(D) = E + AD^{-\alpha_D},$$

where E is the loss floor and A sets the amplitude. Doubling the data multiplies the excess loss $L - E$ by $2^{-\alpha_D}$: a larger α_D means faster improvement at a given floor and amplitude. Fitting these quantities means training a family of models at several scales: the Chinchilla exponents rest on over 400 models from 70M to 16B parameters (Hoffmann et al., 2022), and 512 proxy runs were needed to regress loss on data mixture weights (Liu et al., 2024). That cost is per corpus, which is why data is normally chosen by a heuristic and measured afterwards, if at all. DMO supplies a prediction before that sweep; one low-data training run then fixes the amplitude for loss extrapolation.

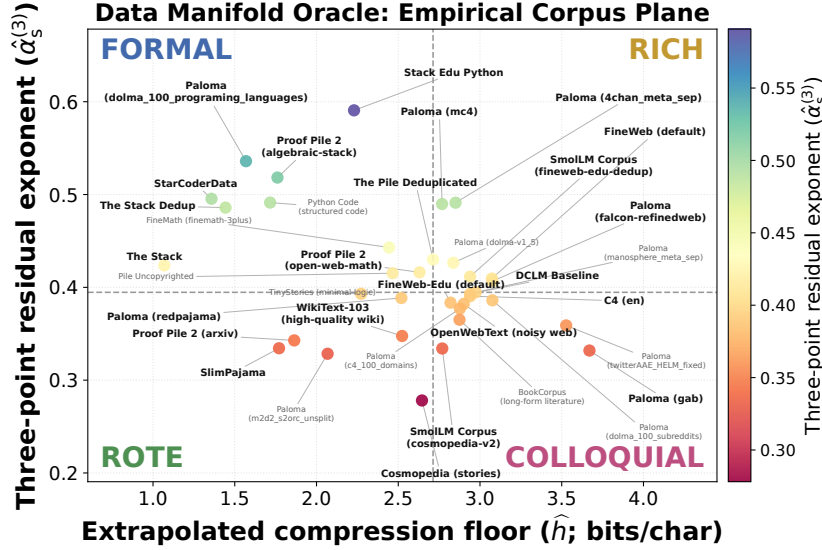


Figure 1: Explanatory three-point compression plane. Corpora placed by the three-point floor $\hat{h}$ and residual exponent $\hat{\alpha}_s^{(3)}$ of Figure 3, at 10,000 documents each; median splits define the four descriptive quadrants, whose corner names Section 5 reuses for its document-level split. The deployed map uses the eight-point slope $\hat{\alpha}_{LZ}$ in place of $\hat{\alpha}_s^{(3)}$ (Table 1).

What the manifold oracle reads. DMO asks which structure relevant to prediction can be recovered from symbolic observations. The source manifold describes the system generating the symbols; the predictive-state manifold describes a sufficient state used to predict the next symbol. Under the conformal assumptions of Assumption A.4, the entropy rate h_X and an independently measured expansion rate λ on the same clock recover the source dimension, $d_{\text{geo}} = h_X / \lambda$. A lower-dimensional regular predictive state admits faster approximation by a head that receives that state directly, a *direct head*: $\alpha_{\text{head}} \geq 2/d_{\text{pred}}$. This is a lower bound on the model axis; it does not order the fitted data-axis exponents. We therefore test the compression slope directly as a calibrated predictor of α_D .

The same frame fixes the limit. Two sources of different dimension can emit one symbol law, so no symbolic-only estimator is consistent for the dimension, and Theorem B.7 states that as a lower bound holding at every sample size. Raw text supplies an empirical α_D ordering under the fixed protocol, not a geometric scale. Section 3 develops the geometric result and its assumptions.

Reading structure across scales. Longer text windows give a compressor access to more repeated structure. We measure the mean DEFLATE rate on random windows and read two features from its dependence on length: a long-window level $\hat{h}$ and a log-log slope $\hat{\alpha}_{LZ}$. The level summarizes where the measured curve settles; the slope records how quickly the rate falls over the measured lengths. A frozen calibration maps that slope to the fitted data exponent. A gzip statistic has previously been related to scaling parameters (Pandey, 2024), and conditional-entropy and token-correlation statistics measured on trained models have been used to derive a data-scaling exponent (Cagnetta et al., 2026). DMO tests a multiscale compression measurement as a frozen predictor on new corpus families (Sections 2 and 4).

The predictive-information connection. An exact loss decomposition separates information outside the predictor’s context from information within that context that it has yet to learn. It gives a common language for the compression-scale measurements and the training outcomes, and identifies context-dependent prediction gain as the next quantity to measure directly (Section 3).

Table 1: **The fixed DMO-Corpus map.** The two rankers use independent corpus budgets and length grids; neither uses a trained-model quantity.

Output	Corpus budget	Character lengths N (draws per length)	Definition
$\hat{h}$	100,000 documents	128, 2048, 32768 (2000, 1000, 500)	$\max\{0, (C_{128}C_{32768} - C_{2048}^2) / (C_{128} + C_{32768} - 2C_{2048})\}$
$\hat{\alpha}_{LZ}$	10,000 documents	64, 128, ..., 8192 (300, 150, 100, 75, 60, 50, 42, 37)	$ \text{OLSslope}(\log C_N \text{ on } \log N) $

From prediction to intervention. On the 38-corpus calibration panel, $\hat{\alpha}_{LZ}$ ranks fitted data exponents at $\rho = 0.84$; the ranking also holds at individual model sizes from 16M to 1B parameters. The floor readout ranks 26 near-converged neural floors at $\rho = 0.75$. Neither is read as a dimension, which is what the obstruction forbids; the frozen map gives the exponent’s value, and the floor readout is used as a ranker. The prospective family test then evaluates the chosen predictor without refitting.

Changing sequence order tests the same connection within a corpus. We permute text blocks in code, web, and math, retain the strings inside each block, and retrain. Both $\hat{\alpha}_{LZ}$ and α_D move with block length. In code, the intervention shifts compression gains toward longer windows while slowing improvement with more training data. This gives the correlation a concrete interpretation: the manipulation changes the window at which sequential structure is available, and both readouts move with it. Combining the readouts with a 12M-token anchor run reduces median loss-prediction error from 32% to 14% across 882 larger-budget cells under leave-one-corpus-out calibration (Section 4).

2 DMO-Corpus: measuring a compression curve

What the curve is a measurement of. The measured object is the compression curve: a finite, compressor-specific statistic, not a latent state (Theorem B.7). That is why the map from curve to exponent is fit on trained grids instead of derived.

The compression curve. DMO-Corpus needs only the joined raw text X from a candidate corpus. At each character length N , it samples random starting positions and measures the raw-DEFLATE rate

$$c_N(x, i) := \frac{8 |\text{Deflate}_{9,-15}(\text{UTF8}(x_{i:i+N}))|}{N} \quad \text{bits/character},$$

where level 9 and wbits = −15 fix headerless DEFLATE, whose LZ77 window is 32,768 bytes. Let C_N be the mean over independent starting positions drawn with replacement, uniformly from $\{0, \dots, |X| - N - 1\}$; Table 1 gives the number of draws per length. DMO reads this curve in two deployed ways: its long-window level gives the floor ranker $\hat{h}$, while its log–log decay over shorter windows gives the rate-slope ranker $\hat{\alpha}_{LZ}$. With $\mathcal{I}$ denoting the sampled start indices, $Q_{\text{DMO}}(X; \mathcal{I}) = (\hat{h}(X; \mathcal{I}), \hat{\alpha}_{LZ}(X; \mathcal{I}))$. Figure 3 shows the shared compression measurement and its measured three-point plane; Table 1 specifies the separate eight-point slope used for deployment.

Floor ranker. The floor formula extrapolates three measured rates using the finite-range template

$$C(N) = H_0 + A_H N^{-\alpha_s}. \quad (1)$$

The three lengths are in geometric progression, so (1) can be solved for H_0 in closed form, and $\hat{h}$ is that solution clipped at zero (Table 1). It summarizes the curve’s long-window level and is evaluated as a corpus ranker against neural loss floors.

Rate-slope ranker. The slope uses the raw C_N values at eight lengths from 64 to 8192 characters; the 16k and 32k points are excluded because of compressor saturation. A slope of $\hat{\alpha}_{LZ} = 0.18$ means that, along the fitted line, doubling the window reduces bits per character by about 12%. This is a relative decline in the total compressed rate. Its relation to a conditional-entropy-tail exponent depends on the rate’s floor and the compressor; the exact distinction is derived in Section I.

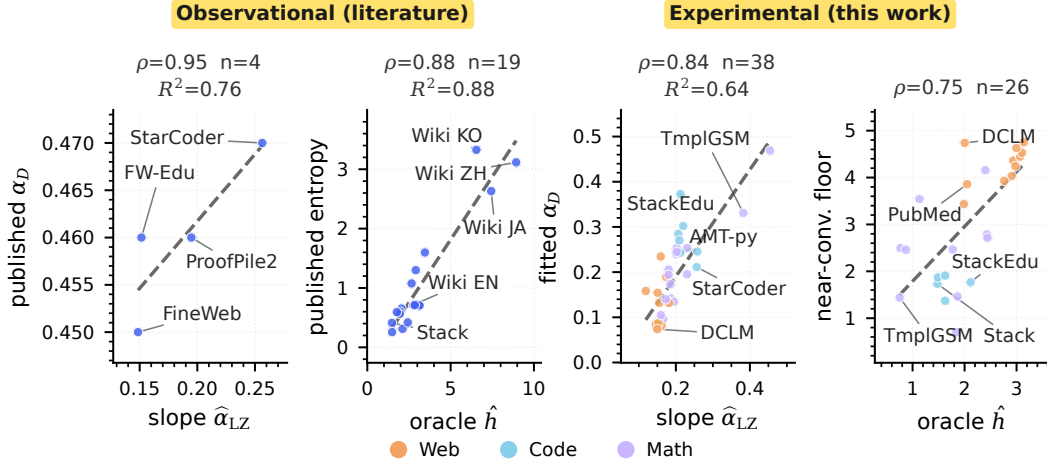


Figure 2: **Predictive raw-text readouts versus external references and fitted scaling quantities.** The left pair compares compression readouts with published scaling exponents and entropy references. The right pair shows our 38-corpus exponent calibration and 26 near-converged neural floor fits. Panels 1 and 3 use the slope $\hat{\alpha}_{LZ}$; panels 2 and 4 use the floor ranker $\hat{h}$ in bits/character, against entropy in bits/character and neural floors in nats/token respectively. Colors denote data type, and four or five corpora are named in each panel, each joined to its point by a leader line. Axis titles are shortened to fit four panels across one text width. Panel 3’s dashed line is the frozen map $\hat{\alpha}_D = 1.182\hat{\alpha}_{LZ} - 0.046$; the other dashed lines are ordinary least squares. Insets give Spearman ρ , R^2 where shown, and panel size. Some entropy references share same-language proxies; neural floor fits require a $D = 20P$ run of at least 11B tokens. Corpus values and fitting details are in Table 8 and Section J.3; prospective transfer is tested separately in Table 2.

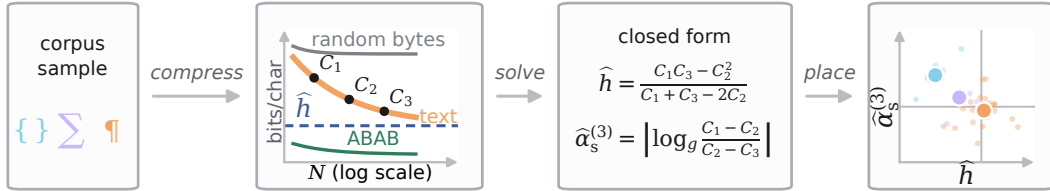


Figure 3: **Reading a corpus into two coordinates without GPUs.** The curve panel is schematic: one fixed compressor reads random bytes, repeated ABAB, and text across a geometric grid of window lengths. Random bytes remain near 8 bits per character, ABAB collapses toward zero, and text falls toward a nonzero level. On the text curve, C_1, C_2, C_3 denote rates at N, gN , and g^2N . The two closed forms give the floor $\hat{h} = (C_1C_3 - C_2^2)/(C_1 + C_3 - 2C_2)$ and the three-point residual exponent $\hat{\alpha}_s^{(3)} = |\log_g[(C_1 - C_2)/(C_2 - C_3)]|$. The final panel plots these recorded coordinates for 36 real corpora from the same three-point artifact used in Figure 7; faint points are corpora, filled markers are data-type centroids, and the crosshairs are coordinate-wise medians. The deployed slope $\hat{\alpha}_{LZ}$ is a separate statistic, an eight-length regression of the raw rate, and enters the frozen map $\hat{\alpha}_D = 1.182\hat{\alpha}_{LZ} - 0.046$ (Table 1). Computing the two plotted coordinates trains no model. The training grids of Section 4 are used only to fit and evaluate the separate $\hat{\alpha}_{LZ} \rightarrow \alpha_D$ map.

The fixed prediction rule. The fixed calibration maps this compressor-and-grid statistic to

$$\hat{\alpha}_D = 1.182\hat{\alpha}_{LZ} - 0.046. \quad (2)$$

For a new corpus, the evaluations in Table 1 suffice to compute the prediction. On the 38-corpus calibration panel, $\hat{\alpha}_{LZ}$ ranges from 0.119 to 0.454 (median 0.180). Rerunning the readout at five start-index seeds moves it by 0.0006–0.0022, which is its noise floor; a synthetic source with no long-range structure reads 0.152 under the same grid, so the bottom

of the panel’s range is where a structureless control reads (Section D.2). The compressor, grid, and calibration were chosen on that panel and fixed before the prospective family test in Section 4. Section S.1 and Table 9 reports the estimator comparisons; Section P records preprocessing, sampling, and numerical edge cases.

3 The data manifold oracle: information, geometry, and prediction

DMO asks what the structure of a data source tells us about the difficulty of learning from it. The theory starts with a latent source manifold and the symbols it generates. Entropy measures the information carried by those symbols; an independently measured dynamical scale converts that information into geometric dimension. For prediction, the relevant geometry is that of a sufficient state: a representation of the past that retains everything needed to predict the next symbol. The geometry settles two questions and leaves the paper’s main one to measurement. The first is the source dimension, which those two measurements recover and no reading of the symbols alone can. The second is how fast a head supplied with the state can be made to learn, which that dimension bounds from below. Neither result gives the corpus prediction, which we obtain by calibrating $\hat{\alpha}_{LZ}$ against trained grids (Sections 2 and 4).

3.1 Source dimension from entropy and expansion

Entropy and that dynamical scale combine into a single ratio. Under the conformal expanding-source assumptions of Assumption A.4, the source dimension is

$$d_{\text{geo}} = \frac{h_X}{\lambda}. \quad (3)$$

Here h_X is the entropy rate in bits per symbol step, and λ is the average base-2 logarithmic expansion per step. The numerator counts information produced; the denominator measures how quickly geometric resolution is lost. Their ratio recovers the dimension, provided both use the same clock. Ideal universal compression identifies h_X in its asymptotic regime (Theorem B.1 and Proposition B.6).

The distinction between information and geometric scale is exact. A fourfold expanding map on a circle and a coordinatewise twofold expanding map on a two-dimensional torus both generate independent, uniform four-symbol observations. Their entropy rates are both 2 bits per step, but their expansion rates are 2 and 1. Equation (3) therefore returns dimensions 1 and 2. The identical symbolic laws also give a sharp identification limit: for every observation length, any symbol-only dimension estimator has expected absolute error at least 1/2 on one of these sources (Theorems B.7 and H.1). Measuring expansion supplies exactly the scale that additional symbols cannot determine.

From a manifold to a prediction rate. With that scale in hand, dimension also fixes a guaranteed rate for a learner given the state. For P -parameter deep-ReLU logit heads given a smooth sufficient predictive state, Lipschitz conditional logits admit a best-in-class parameter-count approximation exponent $\alpha_{\text{head}} \geq 2/d_{\text{pred}}$: lower-dimensional predictive states admit a faster guaranteed rate. That is a statement on the model axis, about what the function class can reach; the corpus experiments measure α_D on the data axis and test the $\hat{\alpha}_{LZ} \rightarrow \alpha_D$ calibration directly. If the source manifold maps onto that predictive state through a sufficient γ -Hölder factor, the conditional composition in Theorem C.1 gives

$$\alpha_{\text{head}} \geq \frac{2}{d_{\text{pred}}} \geq \frac{2\gamma}{d_{\text{geo}}} = \frac{2\gamma\lambda}{h_X}. \quad (4)$$

This bound concerns a head supplied with the state. Proposition C.2 then proves that an expanding source never supplies the factor it needs, so an instance has to come from an invertible source instead. Sections A to C give the full state, regularity, and factor conditions. The ratio also has a finite-sample form: a one-sided entropy ceiling U_h and same-clock expansion lower bound $\lambda_L > 0$, with failure probabilities δ_h and δ_λ , yield $d_{\text{geo}} \leq U_h/\lambda_L$ with probability at least $1 - \delta_h - \delta_\lambda$ (Corollary C.3).

3.2 The information a learner can use

The practical readout starts from the numerator of this ratio, the entropy. Longer windows expose more structure to compression. To connect that scale dependence to prediction, consider a stationary ergodic process on a finite alphabet. Let p_m be the true next-symbol distribution conditional on the last m symbols, let $h_m = H(X_0 \mid X_{-m}^{-1})$, and let q_m be a predictor using that context. Its population log-loss satisfies

$$L(q_m) - h_X = \underbrace{h_m - h_X}_{\text{information beyond the context}} + \underbrace{\mathbb{E} \text{KL}_2(p_m \parallel q_m)}_{\text{prediction error within the context}}. \quad (5)$$

The first term is the best possible gain from access to the more distant past. It equals the conditional mutual information between the next symbol and that past, given the visible context (Proposition B.2). The second term is the gap between the predictor and the best prediction available from the same m symbols. The entropy chain rule and the cross-entropy–KL identity prove (5); the full argument is in Section N.1.

More training data reduces the second term: it improves the use of information already inside the context. In the horizon-limited model of Cagnetta et al. (2026), the growth of a usable prediction horizon links conditional-entropy decay to the data-scaling exponent. Their entropy-tail measurement uses trained-model losses. DMO measures the raw corpus first and tests its forecast against the subsequent training curve.

The empirical readout. What we measure on the raw corpus is a gain per doubling. The observed gain $G_{\text{comp}}(N) = C_N - C_{2N}$ records how many additional bits per character the compressor saves when its input window doubles. The slope $\hat{\alpha}_{1,Z}$ summarizes relative rate changes over eight lengths. These are measurements of the specified compressor’s access to structure; (5) describes an ideal prediction target. The empirical question is whether the compression readout predicts how quickly trained models approach their fitted floors.

Section 4 tests that connection, while (3) states the additional measurement that supports its geometric interpretation.

4 Predicting data scaling from raw text

DMO-Corpus predicts how quickly loss falls as training data grows. We first test the fixed map on new corpus families, then change sequence order within a corpus, and finally use one training point to predict the benefit of a larger data budget.

4.1 Transfer to new corpus families

The fixed map reduced prediction error by 40% against the intercept baseline and 36% against the domain mean on five new parent families, each one source, such as BigPatent or Hacker News, trained on its own 6×5 grid of model sizes and data budgets. Six families were measured before their scaling fits existed; all 180/180 training cells passed the technical checks, and five families passed the fit-quality threshold declared before training, a joint-fit $R^2 \geq 0.85$ on at least 16 cells; these are the evaluable families. On those five, mean absolute error (MAE) was 0.029, against 0.048 and 0.045 for the two baselines. The predicted and measured exponents had Spearman correlation $\rho = 0.800$, and all five fell within the fixed prediction intervals. These results satisfy every criterion of the frozen transfer rule (Table 2 and Section J.7; the appendix files the corpus-level experiments as Case 1). Neither the map nor the calibration panel was updated with these outcomes.

Spending training compute on the candidate corpus does not buy a better exponent.

The obvious alternative to reading the text is to train something small on it. Fitting $L = E + AD^{-\alpha}$ on the 16M-parameter row of the same grid, five budgets and 1.1B tokens per corpus, gives MAE 0.084 on these five families, against 0.029 for the raw-text map. Three budgets instead of five costs a sixth of the tokens and gives 0.085. The proxy fails by bias

Table 2: **Prospective transfer: recorded numerical precision.** Predictions use the fixed map without refitting. Residual is observed minus predicted α_D , and fit R^2 is computed separately within each family. All five residuals are inside intervals of fixed half-width 0.103.

Family	$\hat{\alpha}_{LZ}$	predicted α_D	observed α_D	residual	fit R^2
BigPatent descriptions	0.207	0.198	0.239	+0.041	0.930587
Hacker News	0.141	0.120	0.139	+0.019	0.950144
Dolmino FLAN	0.220	0.214	0.151	-0.063	0.947757
YouTube Commons	0.159	0.142	0.146	+0.003	0.954415
Amazon Books	0.149	0.130	0.111	-0.019	0.928724

Table 3: **Floor prediction has weaker transfer.** The first two rows are within-panel associations; the third is the prospective diagnostic, scored by leave-family-out prediction error. [Sections J.3, J.6](#) and [J.7](#) give the full record.

Evaluation	Rank correlation	Scope or prediction error
Published entropy references	$\rho = 0.88, n = 19$	Some assignments reuse same-language proxies
Long-run neural floor fits	$\rho = 0.75, n = 26$	Runs consumed ≥ 11 B tokens, against a gate of half the requested 20B; no separate convergence test
Matched-token diagnostic	$\rho = 0.30, n = 5$	Leave-family-out MAE 2.064 vs. intercept 1.414 bits/token

rather than by noise: one model size underestimates the joint exponent on every family, by two to three times. On ranking the two are indistinguishable at this sample size: the three-budget proxy orders the five families exactly ($\rho = 1.00$, exact $p = 0.008$) against $\rho = 0.80$ for $\hat{\alpha}_{LZ}$, and one adjacent swap moves ρ by 0.1 at $n = 5$. What the raw text supplies without training is the value of the exponent, not only its ordering ([Section J.8](#)).

Calibration and model-size transfer. The slope $\hat{\alpha}_{LZ}$ ranks fitted α_D at $\rho = 0.84$ on the 38 calibration corpora passing the declared scope and fit-quality filters ([Figure 2](#)). The ordering also holds at each model size: fitting $L(D) = E_p + A_p D^{-\alpha_{D,p}}$ separately with a free floor gives $\rho(\hat{\alpha}_{LZ}, \alpha_{D,p}) = 0.88\text{--}0.96$ from 16M to 1B parameters. After controlling for $\hat{h}$, the partial correlations remain 0.73–0.85 ([Section J.4](#)). Thus the association is visible within fixed model sizes as well as in the joint scaling fits.

Leaving out each of the calibration panel’s 27 parent families gives MAE 0.0514 and $Q^2 = 0.584$, compared with 0.0785 and -0.057 for an intercept. A domain-mean baseline has similar MAE (0.0510), with lower $Q^2 = 0.332$ and rank correlation 0.543, versus DMO’s 0.795. This is internal validation of an estimator selected on the panel; the frozen five-family test above measures prospective transfer. Compressor variants, common-grid refits, and measurement resampling are reported in [Sections J.3, J.7](#) and [S.1](#).

The floor coordinate. The floor ranker $\hat{h}$ orders 19 published entropy-reference assignments at $\rho = 0.88$ and 26 near-converged neural floors at $\rho = 0.75$ (permutation $p = 10^{-5}$; [Table 3](#)). These are within-panel rankings, and some references reuse same-language proxies. The slope adds information beyond this compression level, retaining partial $\rho = 0.74$ with α_D after controlling for $\hat{h}$ on the 38-corpus panel.

4.2 Moving structure across scales changes data scaling

Changing sequence order moves $\hat{\alpha}_{LZ}$ and the fitted data exponent together within code, web, and math corpora. We cut documents into ℓ -character blocks, shuffle the blocks, and retrain a fifteen-cell grid for each arm, using $\ell \in \{64, 256, 1024, 4096\}$ and an unshuffled control. From the 64-character shuffle to the original order, α_D rises from 0.032 to 0.055 on FineWeb-Edu ([Penedo et al., 2024](#)), 0.085 to 0.130 on StarCoder ([Li et al., 2023](#)), and 0.053 to 0.098 on

Proof Pile 2 arXiv (Azerbaiyev et al., 2024). Every tested compression-slope estimator tracks these changes; the selected slope gives a one-sided stratified exact permutation $p = 0.00013$ over the three within-corpus arm orderings (Section J.9). The manipulation retains block contents and document lengths while changing order and post-shuffle tokenization.

Where the compression gain occurs. The diagnostic compression curves extend to 32,768 characters and show which window lengths recover the structure affected by shuffling. Define

$$G_{\text{comp}}(N) := C_N - C_{2N}, \quad (6)$$

the bits saved per character when the compression window doubles. On code, shuffling at $\ell = 64$ raises $G_{\text{comp}}(16,384)$ from 0.024 to 0.128 bits/character, a $5.33\times$ increase, while α_D falls from 0.130 to 0.085. Compression gains have moved toward longer windows as learning from additional data becomes slower. Across the fifteen arms, gains at short windows are positively associated with α_D , whereas gains from 2048 characters onward are negatively associated after centering within each corpus. This locates the empirical connection at the scale of predictable structure; the compression window is a measured property of the text and compressor, not a measurement of the model’s effective context.

Matching the slope across different mixtures. The scalar slope also predicts two StarCoder mixtures that distribute the shuffle differently across documents. We shuffle a fraction of documents at $\ell = 64$ and leave the rest intact, choosing the fraction to match a uniformly shuffled arm’s slope. Both mixture grids completed 15/15 cells. Their measured exponents lie 0.0026 and 0.0035 from the matched-arm predictions, compared with 0.0105 and 0.0164 from a weighted average of the two ingredient exponents: the matched-slope predictions are $4.1\times$ and $4.7\times$ closer. This two-pair test supports the aggregate slope as a predictor within code; the construction and realized matching errors are given in Section J.9.

4.3 One anchor run predicts the return to more data

One training point supplies the amplitude missing from the compression readouts. At a fixed model size, let (D_0, L_0) be the loss after training on $D_0 = 12\text{M}$ tokens. Using the calibrated floor and exponent gives

$$\hat{L}(D) = \hat{E} + (L_0 - \hat{E})(D/D_0)^{-\hat{\alpha}_D}. \quad (7)$$

The anchor fixes the amplitude of the curve; the exponent determines how quickly its reducible loss falls. Across 882 larger-budget cells, with all calibration constants fitted without the predicted corpus, median relative error drops from 32% when carrying the anchor loss forward to 14% with this extrapolator; refitting the full curve gives 5.9%. This within-panel check supports predictions to roughly ten times the anchor’s data budget (Section J.10). It gives a concrete use for the raw-text readout: estimate the return to additional training data after paying for one low-data-budget run.

What the ordering establishes. The recovered ordering is an empirical ordering of the fitted data exponent, not of a latent dimension. Across 38 corpora the rank correlation is $\rho = 0.84$, and it ranges from 0.88 to 0.96 within each model size. The block-order intervention provides within-corpus evidence: it preserves the strings inside each block while changing longer-range order, and the compression slope and fitted exponent move together ($p = 0.00013$). The readout tracks where accessible sequential structure sits, not how much text there is.

The floor separates by data type more than it separates within one, which the frame also predicts: the level a curve settles at is set by the source’s irreducible content, and code, math, and web differ in that content by more than corpora inside any one of them do. Per-type intercepts take the floor panel from $R^2 = 0.483$ to 0.742 on a common slope of $+0.362$, while within web the association is $\rho = +0.64$ and within code it is -0.10 on five corpora spanning 27% of the pooled $\hat{h}$ range (Section J.3). The two coordinates are not two views of one axis.

Table 4: **Downstream accuracy of the four quadrants.** Mean accuracy (%) across four multiple-choice tasks for the four DMO-Doc quadrants. **Bold** marks the best and underline the worst quadrant at each model size.

Size	FORMAL	RICH	ROTE	COLLOQUIAL	Left–Right (pt)
150M	44.7	44.5	44.9	<u>43.8</u>	+0.7
300M	47.8	47.7	48.2	<u>47.4</u>	+0.5
600M	50.2	<u>49.4</u>	50.8	49.5	+1.1
2B	52.6	51.9	52.5	<u>51.6</u>	+0.8

Evaluation note. Results are 0-shot lm-eval averages over HellaSwag, ARC-Easy, PIQA, and SciQ, using 1,000 examples per task. LAMBADA is omitted because a formatting artifact produces 0–2% accuracy at these scales. Each cell is one run, and the 1.2B row is omitted for the reason given in [Section O.2](#); [Section 5.1](#) reports the held-out bits-per-byte results.

5 Extension: pre-training data selection

DMO-Doc uses compression level and sequential structure to select individual documents. Its two scores are the whole-document compression rate c_i^{doc} and Higuchi persistence $r_{\text{Hig}} := 2 - D_{\text{Hig}}$, computed in about 3 ms per document on one CPU core. These document scores have their own training tests: when aggregated over corpora, their correlations with $\hat{h}$ and $\hat{a}_{\text{LZ}}$ are 0.84 and 0.58, respectively ([Section O.1](#)). Splitting a pool at both medians gives four quadrants, carrying the corner names of [Figure 1](#) and named for what they contain: **RICH** (high compression rate, high persistence) holds diverse or rare text, **FORMAL** (low, high) coherent structured prose, **COLLOQUIAL** (high, low) templated or fragmented scrapes, and **ROTE** (low, low) repetitive text ([Section O.1](#)). What follows is the pretraining evidence: a median split of one pool across four model sizes, then a selection-strength sweep.

5.1 Pretraining across four model sizes

High-persistence data gives lower held-out Paloma loss at all four model sizes, and the lead holds rather than shrinking with scale. We split one DCLM ([Li et al., 2024](#)) pool at the medians of the two document scores and pretrain 150M–2B models on each quadrant at $D \approx 20P$ through 600M; at 2B the quadrant pool runs out first, so that size trains on the whole pool, 11.2–11.9B tokens, which leaves it at $D \approx 6P$. Averaging the two quadrants on each side of the persistence median, the high-persistence groups lead by 0.032 bits/byte at 150M and by 0.023 at 2B, with the two intermediate sizes inside that range ([Figure 4](#)). Each cell is a single run. A separate six-arm 150M grid at three seeds puts the single-run seed standard deviation at 0.0086 bits/byte and reproduces the persistence gap at 0.024.

That six-arm grid also says what the split buys against not selecting at all. **RICH** trains 0.015 bits per byte better than a random draw of its own pool ($t = 2.7$); a length-only filter blind to both document scores does not (+0.004); **FORMAL** matches the draw and **ROTE** is 0.028 worse ([Figure 6C](#)). Task accuracy follows the other split, favouring the more compressible quadrants ([Table 4](#)). A fifth size, 1.2B, is left out: its four quadrants drew from two different document pools in different proportions, and two of its cells were scored from checkpoints stopped mid-schedule. Per-size readings and the data checks are in [Sections O.2](#) and [T](#), which file the document-selection runs as Case 2.

5.2 Selection strength: keeping the top τ percent

Keeping the top-scoring tenth of a pool trains a better model than drawing the same number of documents from that pool at random, at all four sizes we ran. We rank a 350B-token FineWeb sample ([Penedo et al., 2024](#)) by the two scores the document selector records, a multi-window LZ rate slope and a PPM floor fit, keep the top τ percent, and pretrain models from 150M to 1.2B on a step budget fixed within each size. Each arm then draws its 200,000 training documents uniformly from its own top- τ pool, so $\tau = 100$ is a random draw of the whole pool at the same budget: the $\tau = 10$ and $\tau = 100$ arms share 0.1% of their documents,

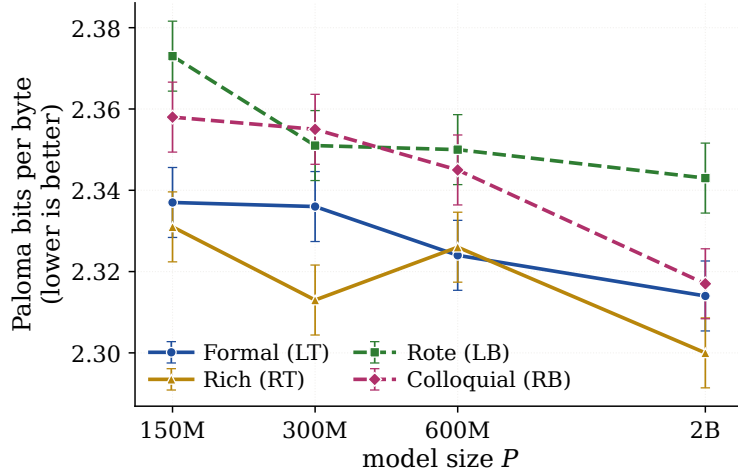


Figure 4: **The four quadrants keep their order from 150M to 2B: the two high-persistence pools train better models at every size, and the four curves descend in parallel.** Held-out mean Paloma bits per byte, lower is better, for DCLM split at the medians of the two document scores and pretrained at $D \approx 20P$ through 600M. At 2B the pool binds first: every quadrant trains on the whole of its own pool, 11.2–11.9B tokens, which leaves 2B at $D \approx 6P$. Solid lines are the high-persistence quadrants **FORMAL** and **RICH**, dashed the low-persistence **ROTE** and **COLLOQUIAL**. The gap between the two sides is 0.0315, 0.0285, 0.0225 and 0.0230 bits per byte at the four sizes, between 2.6 and 3.7 standard deviations each; a separate six-arm 150M grid at three seeds reproduces it at 0.0243 (4.9 sigma). Error bars are one single-run seed standard deviation, 0.0086 bits per byte, measured from that grid. Fitted slopes are -0.022 to -0.038 bits per byte per decade of P and no pair separates by more than 1.2 standard deviations, so the quadrant effect reads as an offset rather than a change in how each pool scales. Each cell of this figure is one run. A fifth size, 1.2B, was trained and is not plotted; [Section O.2](#) gives the reason. Model size is written P in the paper and N elsewhere.

and the mean scores of what each arm actually trained on move with τ in the predicted direction, by 0.65 standard deviations on the slope and 1.0 on the floor.

Mean 0-shot accuracy over eight tasks favours $\tau = 10$ over that random draw by 1.1, 0.6, 1.0 and 0.8 points at 150M, 300M, 600M and 1.2B ([Figure 6](#)). Accuracy falls as τ loosens at 150M and 600M, Spearman -1.0 , and less cleanly at 300M and 1.2B, -0.8 and -0.4 . The gain is concentrated: SciQ moves 4.2 points averaged over the four sizes and ARC-Easy 1.7, between them 0.74 of the 0.88-point mean. Counting each task once across sizes, six of the eight improve at 0-shot and all eight at 5-shot, a second reading of the same checkpoints, sign test $p = 0.008$.

Acknowledgement

We thank Boaz Barak and Elliot Paquette for helpful discussions. S.K. acknowledges the support from the National Science Foundation Grant under award IIS 2229881; H.Z. and S.K. acknowledge the Chan Zuckerberg Initiative Foundation for establishing the Kempner Institute for the Study of Natural and Artificial Intelligence.

References

- Wasi Uddin Ahmad, Aleksander Basant, Somshubra Majumdar, Sean Narenthiran, Aleksander Ficek, Siddhartha Jain, Vahid Noroozi, and Boris Ginsburg. OpenCodeInstruct: A large-scale instruction tuning dataset for code LLMs. *arXiv preprint arXiv:2504.04030*, 2025a.
- Wasi Uddin Ahmad, Sean Narenthiran, Somshubra Majumdar, Aleksander Ficek, Siddhartha Jain, Jocelyn Huang, Vahid Noroozi, and Boris Ginsburg. OpenCodeReasoning: Advancing data distillation for competitive coding. *arXiv preprint arXiv:2504.01943*, 2025b.

- Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1905.12784.
- Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. In *International Conference on Learning Representations*, 2024.
- Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences (PNAS)*; arXiv:2102.06701, 2024.
- Maissam Barkeshli, Alberto Alfarano, and Andrey Gromov. On the origin of neural scaling laws: from random graphs to natural language. *arXiv preprint*, 2026. arXiv:2601.10684.
- Luís Barreira and Benoît Saussol. Hausdorff dimension of measures via Poincaré recurrence. *Communications in Mathematical Physics*, 219(2):443–463, 2001.
- Luís Barreira, Yakov Pesin, and Jörg Schmeling. Dimension and product structure of hyperbolic measures. *Annals of Mathematics*, 149(3):755–783, 1999.
- Frederic H. Behr, Victoria Fossum, Michael Mitzenmacher, and David Xiao. Estimating and comparing entropy across written natural languages using PPM compression. Technical Report TR-12-02, Harvard University, Division of Engineering and Applied Sciences, 2003.
- Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.
- Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. Cosmopedia: How to create large-scale synthetic data for pre-training. In *Hugging Face Blog*, 2024.
- Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, et al. SmolLM2: When Smol goes big — data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.
- Michael D. Boshernitzan. Quantitative recurrence results. *Inventiones Mathematicae*, 113(1): 617–631, 1993.
- David Brandfonbrener, Nikhil Anand, Nikhil Vyas, Eran Malach, and Sham Kakade. Loss-to-loss prediction: Scaling laws for all datasets. *arXiv preprint arXiv:2411.12925*, 2024a.
- David Brandfonbrener, Hanlin Zhang, Andreas Kirsch, Jonathan Richard Schwarz, and Sham Kakade. Color-filter: Conditional loss reduction filtering for targeted language model pre-training. *Advances in Neural Information Processing Systems*, 37:97618–97649, 2024b.
- Alexander A Brudno. Entropy and the complexity of the trajectories of a dynamical system. *Transactions of the Moscow Mathematical Society*, 2:127–151, 1983.
- Francesco Cagnetta, Allan Raventós, Surya Ganguli, and Matthieu Wyart. Deriving neural scaling laws from the statistics of natural language. In *International Conference on Machine Learning*, 2026. URL <https://arxiv.org/abs/2602.07488>.
- Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Efficient approximation of deep ReLU networks for functions on low dimensional manifolds. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 8174–8184, 2019.
- Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of NAACL-HLT*, 2018.

- Łukasz Dębowski. On the vocabulary of grammar-based codes and the logical consistency of texts. *IEEE Transactions on Information Theory*, 57(7):4589–4599, 2011. arXiv:0810.3125.
- Łukasz Dębowski. From zipf’s law to neural scaling through heaps’ law and hilberg’s hypothesis. *arXiv preprint arXiv:2512.13491*, 2025.
- Grégoire Déléang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. Language modeling is compression. In *The Twelfth International Conference on Learning Representations*, 2024. arXiv:2309.10668.
- Ronald A. DeVore, Ralph Howard, and Charles A. Micchelli. Optimal nonlinear approximation. *Manuscripta Mathematica*, 63(4):469–478, 1989.
- Xin Du and Kumiko Tanaka-Ishii. Correlation dimension of natural language in a statistical manifold. *Physical Review Research*, 6(2):L022028, 2024.
- Werner Ebeling and Thorsten Pöschel. Entropy and long-range correlations in literary english. *Europhysics Letters*, 26(4):241–246, 1994.
- Ronen Eldan and Yuanzhi Li. TinyStories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*, 2023.
- Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. DOI: 10.1090/jams/852; arXiv:1310.0425.
- Stefano Galatolo. Dimension and hitting time in rapidly mixing systems. *Mathematical Research Letters*, 14(5):797–805, 2007.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- Aaron Gokaslan and Vanya Cohen. OpenWebText corpus. <http://Skylion007.github.io/OpenWebTextCorpus>, 2019.
- Xing Han, Ziyin Liu, Suchi Saria, and Paul Pu Liang. On the invariance and generality of neural scaling laws. *arXiv preprint arXiv:2605.07546*, 2026.
- Peter Henderson, Mark S Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel E Ho. Pile of Law: Learning responsible data filtering from the law and a 256GB open-source legal dataset. In *Advances in Neural Information Processing Systems*, 2022.
- Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.
- T. Higuchi. Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena*, 31(2):277–283, 1988.
- Wolfgang Hilberg. Der bekannte Grenzwert der redundanzfreien Information in Texten—eine Fehlinterpretation der Shannonschen Experimente? *Frequenz*, 44(9–10):243–248, 1990.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2203.15556.
- Siming Huang, Tianhao Cheng, Jason Klein Liu, Jiaran Hao, Liuyihan Song, Yang Xu, J Yang, JH Liu, Chenchen Zhang, Linzheng Chai, et al. OpenCoder: The open cookbook for top-tier code large language models. *arXiv preprint arXiv:2411.04905*, 2025.

- Yuzhen Huang, Jinghan Zhang, Zifei Shan, and Junxian He. Compression represents intelligence linearly. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.09937.
- Yiding Jiang, Allan Zhou, Zhili Feng, Sathika Malladi, and J. Zico Kolter. Adaptive data optimization: Dynamic sample selection with scaling laws. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/923285deb805c3e14e1aeebc9854d644-Abstract-Conference.html.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. The Stack: 3TB of permissively licensed source code. *Transactions on Machine Learning Research*, 2023.
- Andrei N. Kolmogorov and Vladimir M. Tihomirov. ε -entropy and ε -capacity of sets in functional spaces. *American Mathematical Society Translations, Series 2*, 17:277–364, 1961.
- Ioannis Kontoyiannis. Asymptotic recurrence and waiting times for stationary processes. *Journal of Theoretical Probability*, 11(3):795–811, 1998.
- Jakub Krajewski, Amitis Shidani, Dan Busbridge, Sam Wiseman, and Jason Ramapuram. Revisiting the scaling properties of downstream metrics in large language model training. *arXiv preprint arXiv:2512.08894*, 2025.
- Karl Küpfmüller. Die entropie der deutschen sprache. *Fernmeldetechnische Zeitschrift*, 7: 265–272, 1954.
- François Ledrappier and Lai-Sang Young. The metric entropy of diffeomorphisms. II. relations between entropy, exponents and dimension. *Annals of Mathematics*, 122(3): 540–574, 1985.
- Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, et al. DataComp-LM: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks*, 2024. arXiv:2406.11794.
- Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. StarCoder: May the source be with you! *Transactions on Machine Learning Research*, 2023.
- Henry W. Lin and Max Tegmark. Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7):299, 2017.
- Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. RegMix: Data mixture as regression for language model pre-training. *arXiv preprint*, 2024. arXiv:2407.01492.
- Guy Louchard and Wojciech Szpankowski. On the average redundancy rate of the Lempel–Ziv code. *IEEE Transactions on Information Theory*, 43(1):2–8, 1997.
- Anton Lozhkov, Loubna Ben Allal, Elie Bakouch, Leandro von Werra, and Thomas Wolf. FineMath: The finest collection of mathematical content. *Hugging Face Datasets*, 2025.
- Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. CodeXGLUE: A machine learning benchmark dataset for code understanding and generation. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks)*, 2021.

- Ian Magnusson, Tai Nguyen, Ben Bogin, David Heineman, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. DataDecide: How to predict best pretraining data with small experiments. *arXiv preprint*, 2025. arXiv:2504.11393.
- Alexander Maloney, Daniel A. Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.
- Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining LLMs at scale. *arXiv preprint arXiv:2309.04564*, 2023.
- Pertti Mattila. *Geometry of Sets and Measures in Euclidean Spaces*. Cambridge University Press, 1995.
- Prasanna Mayilvahanan, Thaddäus Wiedemer, Sayak Mallick, Matthias Bethge, and Wieland Brendel. Llm on the line: Data determines loss-to-loss scaling laws. *arXiv preprint arXiv:2502.12120*, 2025.
- Bettina Messmer, Mahdi Nikdan, Dan Alistarh, and Martin Jaggi. Enhancing language model pre-training with text quality filters. *arXiv preprint arXiv:2506.07733*, 2025.
- Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2303.13506.
- Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. AIMO-2 winning solution: Building state-of-the-art mathematical reasoning models with OpenMathReasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025.
- Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020.
- Yoonsoo Nam, Nayara Fonseca, Seok Hyeong Lee, Chris Mingard, and Ard A. Louis. An exactly solvable model for emergence and scaling laws in the multitask sparse parity problem. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL <https://openreview.net/forum?id=cuWsR25bbI>.
- Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 23, 2010.
- Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Bixi, et al. Sequence modeling and design from molecular to genome scale with Evo. In *Science*, 2024.
- Donald S. Ornstein and Benjamin Weiss. Entropy and data compression schemes. *IEEE Transactions on Information Theory*, 39(1):78–83, 1993.
- Rohan Pandey. gzip predicts data-dependent scaling laws. *arXiv preprint arXiv:2405.16684*, 2024.
- Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. OpenWebMath: An open dataset of high-quality mathematical web text. In *International Conference on Learning Representations*, 2024.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtessam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only. In *Advances in Neural Information Processing Systems*, 2023.

- Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, 2024.
- Yakov B. Pesin. Characteristic Lyapunov exponents and smooth ergodic theory. *Russian Mathematical Surveys*, 32(4):55–114, 1977.
- Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2104.08894.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- Vladimir A. Rokhlin. Lectures on the entropy theory of measure-preserving transformations. *Russian Mathematical Surveys*, 22(5):1–52, 1967.
- Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.
- Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.10938.
- Benoît Saussol, Serge Troubetzkoy, and Sandro Vaienti. Recurrence, dimensions, and Lyapunov exponents. *Journal of Statistical Physics*, 106(3–4):623–634, 2002.
- Serap A. Savari. Redundancy of the Lempel–Ziv incremental parsing rule. *IEEE Transactions on Information Theory*, 43(1):9–21, 1997.
- Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics*, 48(4):1875–1897, 2020.
- Uri Shaham, Alexander Cloninger, and Ronald R. Coifman. Provable approximation properties for deep neural networks. *Applied and Computational Harmonic Analysis*, 44(3):537–557, 2018. arXiv:1509.07385.
- Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *Journal of Machine Learning Research*, 23:1–34, 2022. arXiv:2004.10802.
- Jingzhe Shi, Qinwei Ma, Hongyi Liu, Hang Zhao, Jeng-Neng Hwang, and Lei Li. Intrinsic entropy of context length scaling in LLMs. *arXiv preprint arXiv:2502.01481*, 2025.
- Jonathan W. Siegel. Sharp lower bounds on the manifold widths of Sobolev and Besov spaces. *arXiv preprint arXiv:2402.04407*, 2024.
- Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627b token cleaned and deduplicated version of RedPajama. *Cerebras blog*, 2023.
- Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, and Cathy H Wu. UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.
- Ryosuke Takahira, Kumiko Tanaka-Ishii, and Łukasz Dębowski. Entropy rate estimates for natural language—a new extrapolation of compressed large-scale corpora. *Entropy*, 18(10):364, 2016.
- Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

- Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. OpenMathInstruct-2: Accelerating AI for math with massive open-source instruction data. In *International Conference on Learning Representations*, 2025.
- Alexandra Volkova, Mher Safaryan, Christoph H Lampert, and Dan Alistarh. Towards robust scaling laws for optimizers. *arXiv preprint arXiv:2602.07712*, 2026.
- Aaron D. Wyner and Jacob Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Transactions on Information Theory*, 35(6):1250–1258, 1989.
- Aaron D. Wyner and Jacob Ziv. The sliding-window Lempel-Ziv algorithm is asymptotically optimal. *Proceedings of the IEEE*, 82(6):872–877, 1994.
- Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a. arXiv:2305.10429.
- Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b. arXiv:2302.03169.
- Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017. arXiv:1610.01145.
- Jiasheng Ye, Peiju Liu, Tianxiang Sun, Yunhua Zhou, Jun Zhan, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. *arXiv preprint*, 2024. arXiv:2403.16952.
- Lai-Sang Young. Dimension, entropy and Lyapunov exponents. *Ergodic Theory and Dynamical Systems*, 2(1):109–124, 1982.
- Hanlin Zhang, Depen Morwani, Nikhil Vyas, Jingfeng Wu, Difan Zou, Udaya Ghai, Dean Foster, and Sham Kakade. How does critical batch size scale in pre-training? In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.21676.
- Hanlin Zhang, Jikai Jin, Vasilis Syrgkanis, and Sham M. Kakade. Prescriptive scaling reveals the evolution of language model capabilities. In *Forty-third International Conference on Machine Learning*, 2026. URL <https://openreview.net/forum?id=IkjsHRpuYY>.
- Yifan Zhang. StackMathQA: A curated collection of 2 million mathematical questions and answers sourced from Stack Exchange. *arXiv preprint arXiv:2411.04872*, 2024a.
- Yifan Zhang. Training and evaluating language models with template-based data generation. *arXiv preprint arXiv:2411.18104*, 2024b.
- Yifan Zhang, Yifan Luo, Yang Yuan, and Andrew Chi-Chih Yao. Autonomous data selection with language models for mathematical texts. *arXiv preprint arXiv:2402.07625*, 2024.
- Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.

Appendices

Table of Contents

A Setup: Compression, Source Geometry, and Prediction	18
B Theory: From Compression to Geometry and Model Scaling	21
C Available Conclusions and Required Information	26
D Controlled Checks of the Theory's Components	28
E Predictive Dimension Controls Direct-Head Approximation	29
F Local Dimension of the Source Measure	31
G From Symbolic Recurrence to Geometric Dimension	32
H Why Symbols Do Not Identify Geometric Dimension	34
I Compression measurement details	36
J Case-1 Evidence and Reproducibility	37
K Related Work	47
L Discussion	49
M Notation	50
N Deferred Proofs and Supporting Results	52
O Case-Study Details and Overlap Checks	58
P Data and Measurement Details	60
Q The Entropy-Matched Parity Family	63
R Calibration Certificate on a Source With a Known Exponent	64
S Estimator Sensitivity	65
T Additional document-selection results	66

A Setup: Compression, Source Geometry, and Prediction

DMO-Corpus measures a compression curve; the theory concerns entropy, source geometry, and prediction on a supplied state. These objects are related only under explicit assumptions. This section fixes the distinctions before stating the results.

A.1 Source and Predictive Dimension

The central distinction is between the latent source and the state used for prediction. d_{geo} is the Hausdorff dimension of the source support $\mathcal{M}_{\text{geo}}$. By contrast, d_{pred} is the dimension of the particular sufficient state and metric supplied to the predictor. The direct-head theorem uses d_{pred} ; the recurrence theorem concerns d_{geo} . The two dimensions need not be equal.

Quantity	Meaning and role
d_{geo}	Hausdorff dimension of the latent source support $\mathcal{M}_{\text{geo}}$; in the smooth case, $\mathcal{N}_{\text{cov}}(\varepsilon; \mathcal{M}_{\text{geo}}) = \Theta(\varepsilon^{-d_{\text{geo}}})$.
$d_{\mu}(x)$	Local mass-scaling exponent of the source measure. Under Assumption A.1 , it equals d_{geo} almost surely.
d_{pred}	Dimension of the supplied predictive-state law and metric; it appears in the one-sided direct-head approximation bound and is representation-relative.

Formally, the pointwise dimension is

$$d_{\mu}(x) = \lim_{r \downarrow 0} \frac{\log \mu(B(x, r))}{\log r},$$

when the limit exists.

Recurrence is an observable, not another dimension. Ideal LZ match lengths are dual to return times, and the Wyner–Ziv sliding-window rate has iterated expected limit h_X ([Wyner & Ziv, 1989](#); [Ornstein & Weiss, 1993](#)). Conformal dynamics, a generating code, and a shared clock then connect recurrence to source geometry through $d_{\text{geo}} = h_X / \lambda$ ([Proposition B.6](#)). An additional Hölder sufficient factor relates source geometry to the predictive state, whose direct-head rate obeys $\alpha_{\text{head}} \geq 2/d_{\text{pred}}$. The empirical finite-window slope $\hat{\alpha}_{\text{LZ}}$ is not any of these dimensions.

A.2 Compression Level and Learning Rate

The population *floor* is the irreducible per-symbol cross-entropy h_X of an infinite-context predictor under stationarity and ergodicity ([Theorem B.1](#)); a finite-context model can stop higher. Ideal universal coding identifies h_X asymptotically. The deployed $\hat{h}$ is instead a finite-window ranker. It inherits no consistency or coverage guarantee from that limit; the multi-document statistic U_h gives a finite-sample upper bound under its own sampling assumptions.

Units matter. Case 1 compression is bits per Unicode character (compressed UTF-8 bits divided by Python character count), while fitted neural floors are nats per token. The cross-unit floor comparison is a rank diagnostic, not a claim that the levels agree. Formal source-coding statements use bits and KL_2 ; logit calculations use natural logs. The Lyapunov exponent uses base-2 logs whenever entropy is divided by it. Converting bits to nats multiplies a loss by $\ln 2$ and leaves decay exponents unchanged.

The *exponent* depends on the scaling axis. Against parameter count, a smooth predictive-state problem admits the one-sided best-in-class head bound $\alpha_{\text{head}} \geq 2/d_{\text{pred}}$ ([Proposition B.4](#)). Against training data, long memory and estimator choice matter; on the tested panels, $\hat{\alpha}_{\text{LZ}}$ is an empirically calibrated ranker of α_D . The asymmetry follows because compression rates average surprisal, while their finite- N residuals reflect memory and estimator choice. Only a population recurrence observable, together with same-clock geometric assumptions, can

be related to d_{geo} ; the deployed $\hat{\alpha}_{\text{LZ}}$ estimates a different quantity ([Propositions B.6 and B.8](#) and [Theorem B.7](#)).

A.3 Assumptions behind Each Link

The assumptions fall into three groups. The LZ limit needs only a symbolic source; the direct-head rate needs predictive-state geometry and regularity; the recurrence bridge needs conformal source dynamics. A final Hölder condition connects the source and predictive states. Throughout, $(X_t)_{t \in \mathbb{Z}}$ is a two-sided stationary process on a finite alphabet Σ , so conditioning on the infinite past $X_{\leq t}$ is well defined; [Assumption A.3](#) adds ergodicity and fixes the entropy rate.

Source-side assumptions.

Assumption A.1 (Smooth exact-dimensional geometry). $\mathcal{M}_{\text{geo}} \subset \mathbb{R}^{d_{\text{geo,amb}}}$ is a compact d_{geo} -dimensional $C^{k_{\text{geo}}}$ Riemannian manifold ($k_{\text{geo}} \geq 2$), and $\mu \ll \mathcal{H}^{d_{\text{geo}}} |_{\mathcal{M}_{\text{geo}}}$ obeys

$$0 < c_1 \leq \frac{d\mu}{d\mathcal{H}^{d_{\text{geo}}}} \leq c_2 < \infty.$$

Thus small-ball masses and covering counts report the same d_{geo} . The arguments that do not use smoothness require only an exact-dimensional measure with local dimension d_{geo} ; the displayed form is the integer-dimensional specialization. Singular measures are handled only by the fractal extension of [Remark G.6](#). Write ρ_{geo} for the Riemannian distance on $\mathcal{M}_{\text{geo}}$.

Assumption A.2 (Geometric dynamics). The source geometry is a stationary ergodic system $(\mathcal{M}_{\text{geo}}, T, \mu)$ with T measure-preserving.

Assumption A.3 (Stationary ergodic symbolic source). $(X_t)_{t \in \mathbb{Z}}$ is a two-sided stationary ergodic process with law $\mathbb{P}_X$ on a finite alphabet Σ , with finite entropy rate $h_X = \lim_{m \rightarrow \infty} H(X_0 \mid X_{-m}^{-1})$. This assumption supports the ideal universal-LZ and symbolic return-time limits; it imposes no smooth geometry.

Assumption A.4 (Conformal expanding bridge [**Cond. D**]). In addition to [Assumptions A.1 and A.2](#), the symbols arise from a finite generating partition of finite entropy, and $(\mathcal{M}_{\text{geo}}, T, \mu)$ is a $C^{1+\beta_{\text{dyn}}}$ conformal expanding map for some $\beta_{\text{dyn}} > 0$ with absolutely continuous SRB/ACIP measure. The process (X_t) of [Assumption A.3](#) is the resulting symbol stream; a non-invertible T codes to a one-sided shift, whose unique two-sided stationary extension has the same entropy rate, so every statement proved under [Assumption A.3](#) applies verbatim to that coding. Thus $D_X T$ is a scalar multiple of an isometry, $|\det D_X T| = \|D_X T\|^{d_{\text{geo}}}$, $\|D_X T\| > 1$, and the single positive Lyapunov exponent is

$$\lambda = \int_{\mathcal{M}_{\text{geo}}} \log_2 \|D_X T\| d\mu > 0.$$

Their dimension consequence is stated in [Proposition B.6](#), with the full Rokhlin–Pesin argument in [Lemma G.4](#). The generating, absolute-continuity, and conformality clauses play distinct roles; the singular and non-conformal alternatives are treated in [Remarks G.6 and G.7](#).

Remark A.5 (Same clock and fixed blocking). Entropy and expansion must use the same time unit and logarithmic base. Under fixed k -blocking,

$$h_X^{(k)} = kh_X, \quad \lambda^{(k)} = k\lambda, \quad h_X^{(k)} / \lambda^{(k)} = h_X / \lambda,$$

so $2\gamma\lambda/h_X$ is invariant. Variable-length tokenization is not fixed blocking; both inputs must first be converted to a common clock. For UTF-8, [Lemma B.11](#) relates character-normalized population surprisal to the entropy rate of the character-aligned byte stream. It does not convert an arbitrary Lyapunov exponent or certify the finite-window DEFLATE statistic; a geometric bridge must justify the entropy and expansion clocks separately.

Remark A.6 (Ergodicity versus mixing). The entropy, return-time, and ideal universal-LZ limits require stationarity and ergodicity, not a mixing rate. Fast mixing enters only when the recurrence-dimension lower bound is upgraded to a two-sided limit ([Saussolet et al., 2002](#); [Galatolo, 2007](#)); [Assumption A.4](#) supplies it.

Predictive-side assumptions.

Assumption A.7 (Predictive-state law and direct-state input). Fix a past-measurable state $S_t = e(X_{\leq t})$ that is predictively sufficient:

$$\mathbb{P}(X_{t+1} \in \cdot \mid X_{\leq t}) = \mathbb{P}(X_{t+1} \in \cdot \mid S_t) \quad \text{a.s.}$$

Under stationarity let $\nu = \text{Law}(S_t)$ and $\mathcal{M}_{\text{pred}} = \text{supp } \nu$. For [Proposition B.4](#), $\mathcal{M}_{\text{pred}}$ is a compact embedded d_{pred} -dimensional $C^{k_{\text{pred}}}$ manifold, $k_{\text{pred}} \geq 2$, in $\mathbb{R}^{d_{\text{pred,amb}}}$, where $0 < d_{\text{pred}} < d_{\text{pred,amb}}$. Write ρ_{pred} for the induced Euclidean metric. A fixed affine change of coordinates places $\mathcal{M}_{\text{pred}}$ inside $[0, 1]^{d_{\text{pred,amb}}}$. Moreover,

$$\nu \ll \mathcal{H}^{d_{\text{pred}}} \big|_{\mathcal{M}_{\text{pred}}}, \quad 0 < c_1^{\text{pred}} \leq \frac{d\nu}{d\mathcal{H}^{d_{\text{pred}}}} \leq c_2^{\text{pred}} < \infty.$$

These are prediction-side conditions and imply no source geometry. The function class receives S_t directly. Thus α_{head} excludes the cost of learning e , optimization, and end-to-end Transformer realization. A zero-dimensional state is handled by finite or exact representation, not by the quotient $2/d_{\text{pred}}$.

The quantity d_{pred} belongs to this declared state representation and metric. It is unchanged by bi-Lipschitz reparameterization, but can change under non-bi-Lipschitz recoding or the addition of irrelevant past-measurable nuisance coordinates.

Assumption A.8 (Prediction regularity). On $\mathcal{M}_{\text{pred}}$, write $p(\cdot \mid s)$ for the conditional law of X_{t+1} given $S_t = s$, with coordinates $p_i(s) := \mathbb{P}(X_{t+1} = i \mid S_t = s)$. Assume $p_i(s) > 0$ for every $s \in \mathcal{M}_{\text{pred}}$ and $i \in \Sigma$, and define the canonical conditional-logit map by

$$\Lambda_i(s) = \ln p_i(s) - |\Sigma|^{-1} \sum_j \ln p_j(s).$$

Assume that Λ is L_{logit} -Lipschitz in sup norm. This regularity converts a geometric covering rate into cross-entropy achievability. On compact $\mathcal{M}_{\text{pred}}$ it also implies a finite bound $\|\Lambda(s)\|_{\infty} \leq B$ and hence $p_i(s) \geq e^{-2B}/|\Sigma|$.

Remark A.9 (Probability floor and the KL constant). Compactness and Lipschitz conditional logits imply a positive probability floor: $\|\Lambda(s)\|_{\infty} \leq B$ gives $p_i(s) \geq e^{-2B}/|\Sigma|$. The excess-risk proof uses the uniform softmax bound $(2 \ln 2)^{-1} \|\delta\|_{\infty}^2$, whose constant is independent of the value of that floor. Non-Lipschitz rare-token cliffs remain outside [Assumption A.8](#)'s scope.

Connecting source and predictive states.

Lemma A.10 (Hölder sufficient factor). Suppose $Z_t \in \mathcal{M}_{\text{geo}}$ and $S_t = \Phi(Z_t)$ almost surely, where S_t is past measurable and predictively sufficient as in [Assumption A.7](#), $\nu = \Phi_{\#}\mu$, $\mathcal{M}_{\text{pred}} = \Phi(\mathcal{M}_{\text{geo}})$, and

$$\rho_{\text{pred}}(\Phi(z), \Phi(z')) \leq K \rho_{\text{geo}}(z, z')^{\gamma}, \quad 0 < \gamma \leq 1.$$

Then

$$d_{\text{pred}} = \dim_{\text{H}} \Phi(\mathcal{M}_{\text{geo}}) \leq \frac{\dim_{\text{H}} \mathcal{M}_{\text{geo}}}{\gamma} = \frac{d_{\text{geo}}}{\gamma}.$$

Proof sketch. A γ -Hölder map sends a cover of diameter δ to one of diameter at most $K\delta^{\gamma}$; the Hausdorff-cover definition therefore gives $\dim_{\text{H}} \Phi(\mathcal{M}_{\text{geo}}) \leq \dim_{\text{H}}(\mathcal{M}_{\text{geo}})/\gamma$. See [Section N.1](#).

Remark A.11 (Geometric and predictive states). $\mathcal{M}_{\text{geo}}$ supports the dynamical system in [Assumptions A.2](#) and [A.4](#); $\mathcal{M}_{\text{pred}}$ supports the sufficient statistic used for prediction. Only [Lemma A.10](#) compares their dimensions. Simple examples show why this condition matters. The binary doubling map has nontrivial source geometry but an i.i.d. symbol stream, so its minimal predictive state is constant; feeding an exact deterministic state instead makes the next-symbol law a point mass and violates the finite-logit assumption. In the other direction,

a hidden-state filter can have a positive-dimensional Blackwell predictive law with smooth logits, but its dimension involves projection entropy and filter contraction, not the source Lyapunov exponent, and its support is generally Cantor rather than smooth.

The examples therefore instantiate the source and prediction results independently. The source-to-head bound in [Theorem C.1](#) additionally requires a positive-dimensional smooth sufficient state satisfying [Assumptions A.7](#) and [A.8](#), the conformal source bridge in [Assumption A.4](#), and the Hölder map above. No nontrivial source satisfies all of them at once, for the reason [Proposition C.2](#) proves.

A.4 Requirements and Notation

[Assumptions A.7](#) and [A.8](#) support the direct-state head bound; [Assumption A.3](#) supports the symbolic limit. The source-to-head composition additionally needs [Assumptions A.1](#), [A.2](#) and [A.4](#), same-clock λ , and [Lemma A.10](#). Finite- N refinements state their stronger assumptions locally.

We reserve P for parameter count, D for the training-token budget, N for text length, N_{tot} for total corpus length, w for the LZ window, m for context length, ℓ for a recurrence block, b for a certificate window in source symbols, r for a metric radius, and g for the geometric grid ratio. Empirical fits use P and D for the model and data axes and write α_D for the data-scaling exponent.

B Theory: From Compression to Geometry and Model Scaling

Compression provides two source-side quantities with different guarantees. An ideal universal-coding limit identifies the entropy rate h_X , whereas a multi-document construction gives a finite-sample upper certificate U_h . Under conformal dynamics, an independently supplied Lyapunov scale converts h_X into geometric dimension. Once a sufficient predictive state is supplied, a Hölder factor from source geometry can then turn that dimension into a direct-head approximation rate. This composition is conditional on a geometric scale that text alone leaves undetermined, uniformly over the stated class.

B.1 The Symbolic Side: An Ideal Limit and a Finite Certificate

We begin with the population entropy target and then give the multi-document procedure that covers it at finite sample size.

Theorem B.1 (Symbolic LZ entropy limit ([Wyner & Ziv, 1989](#); [Ornstein & Weiss, 1993](#); [Wyner & Ziv, 1994](#)); [[Theorem](#)]). *Under [Assumption A.3](#), let $C_{\text{WZ}}(N; w)$ be the per-symbol bit length of the sliding-window code analyzed by [Wyner & Ziv \(1994\)](#), with that paper’s parsing and pointer/length accounting. For each initial database, its suffix code is uniquely decodable. This ideal code is distinct from the fixed DEFLATE implementation used by the empirical instrument. Writing*

$$h_w := \limsup_{N \rightarrow \infty} \mathbb{E} C_{\text{WZ}}(N; w),$$

the ideal universal-LZ floor satisfies

$$\lim_{w \rightarrow \infty} h_w = h_X.$$

The order of limits is fixed: a diagonal $w = w(N)$ and any finite realized codelength sit outside it. If $h_X > 0$, the associated LZ match length L_N also obeys $L_N / \log_2 N \rightarrow 1/h_X$ in probability.

Proof sketch. Kraft’s inequality gives the lower bound $h_w \geq h_X$ for every fixed window, while the expected-rate theorem of [Wyner & Ziv \(1994\)](#) gives the converse after $w \rightarrow \infty$. The match-length conclusion is the Ornstein–Weiss/Wyner–Ziv recurrence theorem quoted in [Lemma G.2](#). The limit-order and initial-database details are proved in [Section N.1](#).

The corresponding source-side notion of memory is compressor-independent.

Proposition B.2 (Conditional-entropy tail). *For the two-sided stationary process of [Assumption A.3](#), let $h_m := H(X_0 | X_{-m}^{-1})$. Then $h_m \downarrow h_X$ and*

$$h_m - h_X = I(X_0; X_{-\infty}^{-m-1} | X_{-m}^{-1}).$$

Moreover, h_m is the minimum population log-loss among predictors using the last m symbols. Thus $h_m - h_X$ is exactly the Bayes penalty for truncating the past, distinct from a finite LZ residual or from $\hat{\alpha}_{LZ}$'s decay exponent.

Proof sketch. Martingale convergence sends the finite-past conditional law to the infinite-past law, and bounded continuity of entropy gives $h_m \downarrow h_X$. The conditional-mutual-information identity is then the entropy chain rule, while the Bayes claim follows from the usual cross-entropy–KL decomposition. See [Section N.1](#).

The ideal limit leaves finite-sample coverage to the certificate below, whose structure is simple: inflate the mean wrapped-compression rate to account for inserted separators, then add one error term for sampling window offsets and another for sampling documents. The statement gives the exact constants needed for coverage.

Proposition B.3 (Finite-sample one-sided floor certificate; [\[Theorem\]](#)). *Fix a finite source alphabet Σ , a window length b in source symbols, $M \geq 2$, and $\delta \in (0, 1)$. Concatenate J_{doc} independent blocks of fixed lengths $L_1, \dots, L_{J_{\text{doc}}}$ drawn from one common stationary source on Σ , whose entropy rate is h_X . Insert a fixed string $\sigma_{\text{sep}} \in \Sigma^{b_{\text{sep}}}$ at the positions determined by those known lengths; its positions, rather than its symbol values, identify it. Let the resulting total length be $N_{\text{tot}} \geq b$. Sample M offsets independently and uniformly from the legal window starts $\{0, \dots, N_{\text{tot}} - b\}$. Suppose each window contains at most $s_{\text{max}} < b$ inserted symbols. Encode each source symbol injectively with a prefix-free byte code of maximum length L_{max} , and apply the wrapped prefix-free DEFLATE construction of [Section N.1.1](#). Let $Z_1, \dots, Z_M$ be the resulting rates in bits per source symbol, with mean $\bar{Z}$ and sample variance*

$$S_Z^2 := \frac{1}{M-1} \sum_{i=1}^M (Z_i - \bar{Z})^2.$$

With $q_b = \lceil \log_2(8L_{\text{max}}b + 1) \rceil$, define

$$\begin{aligned} \kappa &= \frac{b}{b - s_{\text{max}}}, & B_b &= 8L_{\text{max}} + \frac{q_b + 1}{b}, & c_k &= \frac{B_b(L_k + b - 1)}{N_{\text{tot}} - b + 1}, \\ \varepsilon_M &= \sqrt{\frac{2S_Z^2 \ln(4/\delta)}{M}} + \frac{7B_b \ln(4/\delta)}{3(M-1)}, & \varepsilon_{\text{doc}} &= \sqrt{\frac{1}{2} \ln(2/\delta) \sum_{k \leq J_{\text{doc}}} c_k^2}. \end{aligned}$$

Then, jointly over document contents and sampled offsets,

$$\mathbb{P}[h_X \leq U_h(b, M) := \kappa(\bar{Z} + \varepsilon_M + \varepsilon_{\text{doc}})] \geq 1 - \delta, \quad (8)$$

with h_X and $U_h(b, M)$ in the same source-symbol unit. The byte identity map gives a bits-per-byte certificate ($L_{\text{max}} = 1$); UTF-8 gives a bits-per-character certificate ($L_{\text{max}} \leq 4$). The result is finite sample at fixed b and fixed wrapped compressor and assumes no power-law residual model.

Proof sketch. Kraft's inequality lower-bounds the expected wrapped length by the entropy in each sampled window. Empirical Bernstein controls offset sampling within a realized corpus, McDiarmid controls replacement of its independent documents, and a union bound yields (8). The exact wrapper and proof are in [Section N.1.1](#).

In practice, one sufficiently long document can support the empirical scores $(\hat{h}, \hat{\alpha}_{LZ})$. The certificate U_h instead requires independent documents and randomized long windows; only U_h has finite-sample coverage.

B.2 From Predictive Geometry to Model-Size Scaling

We next assume that a predictive state is supplied. The parameter count below covers only the prediction head, not the cost of learning or computing the state from token history.

Proposition B.4 (Cross-entropy achievability exponent). *Let [Assumptions A.7](#) and [A.8](#) hold on a predictive-state manifold $\mathcal{M}_{\text{pred}}$ of dimension $d_{\text{pred}} > 0$ with state law ν . For the P -parameter*

deep-ReLU logit class of [Lemma E.3](#), define

$$\begin{aligned}\Delta\mathcal{L}(P) &:= \inf_{f \in \mathcal{F}_P} \mathbb{E}_{s \sim \nu} \text{KL}_2(p(\cdot | s) \parallel \text{softmax}(f(s))), \\ \alpha_{\text{head}} &:= \sup\{a : \Delta\mathcal{L}(P) = O(P^{-a})\}.\end{aligned}\tag{9}$$

For every $\eta > 0$ there is a sequence $f_P \in \mathcal{F}_P$ satisfying

$$\mathbb{E}_{s \sim \nu} \text{KL}_2(p(\cdot | s) \parallel \text{softmax}(f_P(s))) = O_\eta(P^{-2/(d_{\text{pred}} + \eta)}).$$

Consequently $\Delta\mathcal{L}(P)$ obeys the same bound and $\alpha_{\text{head}} \geq 2/d_{\text{pred}}$. This is a one-sided best-in-class result about what the function class can reach, independent of the optimizer used to train it.

Proof sketch. The manifold approximation theorem gives uniform logit error $O_\eta(P^{-1/(d_{\text{pred}} + \eta)})$. Hoeffding's lemma turns that error into a uniform quadratic softmax-KL bound. Because this holds for every $\eta > 0$, taking $\eta \downarrow 0$ in the definition of α_{head} gives $\alpha_{\text{head}} \geq 2/d_{\text{pred}}$. See [Section N.1](#) for the endpoint and constant bookkeeping.

Lemma B.5 (Predictive states carry the entropy floor). *For the stationary process of [Assumption A.3](#), if S_t is a measurable function of $X_{\leq t}$ and is predictively sufficient, then*

$$\mathbb{E}H(p(\cdot | S_t)) = H(X_{t+1} | X_{\leq t}) = h_X.$$

An exact latent state of a deterministic system need not have this property; the theorem therefore concerns $\mathcal{M}_{\text{pred}}$, not an arbitrary geometric state space.

Proof sketch. Predictive sufficiency identifies conditioning on S_t with conditioning on the complete past; stationarity then identifies that conditional entropy with h_X . The two-line formal argument is recorded in [Section N.1](#).

Compression reaches source geometry only after one geometric scale is supplied.

Proposition B.6 (Conformal entropy–dimension bridge [[Cond. D](#)]). *Under [Assumptions A.1](#) to [A.4](#), with a conformal expanding map, an absolutely continuous invariant measure, and a generating coding,*

$$d_{\text{geo}} := \dim_{\text{H}}(\mathcal{M}_{\text{geo}}) = \frac{h_X}{\lambda}, \quad \lambda := \int_{\mathcal{M}_{\text{geo}}} \log_2 \|D_x T\| d\mu(x) > 0.$$

Entropy and λ must use the same logarithmic base and the same iterate clock. The scalar identity is specific to conformality; non-conformal systems require the Lyapunov vector formula of [Remark G.7](#).

Proof. The coding partition of [Assumption A.4](#) is generating, so the symbolic entropy rate is the measure-theoretic entropy of the map, $h_X = h_\mu(T)$. Rokhlin–Pesin for the absolutely continuous invariant measure gives $h_\mu(T) = \int_{\mathcal{M}_{\text{geo}}} \log_2 |\det D_x T| d\mu(x)$, and conformality replaces the Jacobian by $\|D_x T\|^{d_{\text{geo}}}$, so $h_\mu(T) = d_{\text{geo}} \lambda$. Since $\lambda > 0$, dividing gives $d_{\text{geo}} = h_X / \lambda$. [Lemma G.4](#) records the same calculation and extends it to the pointwise dimension. $\square$

Fixed blocking preserves this ratio: grouping k source steps multiplies both entropy and the Lyapunov exponent by k . This same-clock invariance is formalized in [Remark A.5](#).

B.3 The Same Symbolic Law Admits Different Source Dimensions

Some non-symbolic geometric side information is necessary over this class; our conformal bridge uses a same-clock Lyapunov exponent as one sufficient choice, though other geometric information may also suffice.

Theorem B.7 (Symbolic obstruction to recovering dimension). *Geometric dimension is not a function of the symbolic process law. There exist two systems satisfying [Assumptions A.1](#), [A.2](#) and [A.4](#) whose symbolic codings are identical in law, and hence agree on every finite or infinite stream statistic, but whose geometric dimensions differ. Consequently, no symbolic-only estimator can be consistent for d_{geo} over this class. The statement includes every compression curve.*

Proof sketch. A q^k -fold map on the circle and a coordinatewise q -fold map on the k -torus both emit the i.i.d. uniform process on q^k symbols under their natural generating partitions. Their dimensions are nevertheless 1 and k , so a common-law estimator cannot converge to both. The construction and consistency argument are in [Section N.1](#). Every compression curve is a measurable function of the common symbolic law and any auxiliary compressor seed, so the two witnesses induce the same law on it too.

Known finite- N redundancy results sharpen this boundary but do not replace it.

Proposition B.8 (Finite- N LZ redundancy: a hierarchy of claims). *For the unifilar and Markov source classes covered by [Savari \(1997\)](#), the corresponding idealized LZ78/LZW rates satisfy*

$$\mathbb{E}C_{\text{LZ78}}(N) - h_X = O(1/\log_2 N) \quad (\text{Savari, 1997}). \quad (10)$$

This is an upper envelope: it leaves the decay order and the leading residual open. For the nondegenerate binary memoryless sources and the precise LZ78 phrase-code convention covered by [Louchard & Szpankowski \(1997\)](#), the redundancy has the exact logarithmic leading form

$$\frac{A_{\text{LZ}} + \delta(N)}{\log_2 N} + O\left(\frac{\log_2 \log_2 N}{(\log_2 N)^2}\right), \quad (11)$$

with a non-vanishing leading coefficient and bounded oscillation. The memoryless case therefore rules out a universal $N^{-1/d}$ law as the leading LZ78 asymptotic. Both theorems are about LZ78/LZW; coverage for the deployed fixed-window statistic comes from the separate U_h of [Proposition B.3](#), under its own sampling assumptions.

Proof sketch. The first claim is the upper bound of [Savari \(1997\)](#). The binary-memoryless expansion of [Louchard & Szpankowski \(1997\)](#) supplies the nonzero logarithmic leading term that rules out a universal polynomial LZ78 law. The precise scope comparison is in [Section N.1](#).

Remark B.9 (Two axes and three exponents). The model and data axes are different. α_{head} is the one-sided, best-in-class model-axis exponent of [Proposition B.4](#); α_D is a trained model's fitted data-scaling exponent, which can depend on memory and spectral structure; and $\hat{\alpha}_{\text{LZ}}$ is the estimator-dependent finite-scale compression slope used to rank α_D empirically. The first composes with source geometry only under the factor assumptions of [Theorem C.1](#), while the $\hat{\alpha}_{\text{LZ}}$ -to- α_D link is only a calibration. These are the only established pairings, and [Theorem B.7](#) rules out interpreting $\hat{\alpha}_{\text{LZ}}$ universally as geometric $1/d_{\text{geo}}$.

B.4 Entropy Does Not Determine the Data Exponent

[Theorem B.7](#) concerns geometric dimension, which is not what the corpus experiments predict. The corresponding statement about the fitted data exponent needs a different witness, because α_D is a functional of the process law: two processes with the same law have the same exponent. What is needed is a pair with the same *entropy function* and different laws.

Proposition B.10 (An entropy-matched family). *Fix $k \geq 1$ and $a \in \{0, 1\}^k$ with $a_1 = 1$. Let $(U_t)_{t \in \mathbb{Z}}$ be i.i.d. uniform on $\{0, 1\}$ and set*

$$X_t^{(a)} := (U_t, \langle a, (U_{t-k}, \dots, U_{t-1}) \rangle) \in \{0, 1\}^2,$$

with $\langle \cdot, \cdot \rangle$ the $\text{GF}(2)$ inner product. Every $X^{(a)}$ is stationary and ergodic, and for every $N \geq 1$

$$H(X_1^{(a)}, \dots, X_N^{(a)}) = N + \min(N, k). \quad (12)$$

The right-hand side does not depend on a . Consequently all members share the entropy rate $h_X = 1$, the conditional-entropy tail $h_m = 2$ for $m < k$ and $h_m = 1$ for $m \geq k$, the excess entropy k , and hence the value of every functional of the block-entropy function, including every Hilbert-type exponent.

Proof. Write $u_j := U_{j-k}$ for $j = 1, \dots, k$: the prefix bits that precede the observation window. They are uniform on $\{0, 1\}^k$ and independent of U_1^N . Set $r := \min(N, k)$.

Step 1: the first coordinates contribute N bits. They are $U_1, \dots, U_N$, so $H(U_1^N) = N$.

Step 2: only the first r second coordinates stay random once U_1^N is known. The window read at time t is $(U_{t-k}, \dots, U_{t-1})$, whose i -th entry $U_{t-k+i-1}$ is a prefix bit exactly when $t - k + i - 1 \leq 0$, that is when $i \leq k + 1 - t$, and then equals u_{t+i-1} . For $t \geq k + 1$ no entry is a prefix bit, so the second coordinate is a function of U_1^N . For $1 \leq t \leq r$,

$$X_{t,2}^{(a)} = \underbrace{\bigoplus_{i \leq k+1-t} a_i u_{t+i-1}}_{\langle v_t, u \rangle} \oplus \underbrace{\bigoplus_{i > k+1-t} a_i U_{t-k+i-1}}_{\text{measurable w.r.t. } U_1^N}. \quad (13)$$

Step 3: the rows $v_1, \dots, v_r$ are independent over $\text{GF}(2)$. The row v_t is supported on coordinates $t, \dots, k$, and its entry at t is $a_1 = 1$. Stacking $t = 1, \dots, r$ gives a matrix in row echelon form with unit pivots, of rank r whatever the remaining entries of a .

Step 4: conclude. Conditionally on U_1^N , the vector $(X_{t,2}^{(a)})_{t \leq r}$ in (13) is a rank- r linear image of the uniform u shifted by a constant, hence uniform on a coset of an r -dimensional subspace. Its conditional entropy is r bits and does not depend on the conditioning value, so adding Step 1 gives $H(X_1^{(a)}, \dots, X_N^{(a)}) = N + r$, which is (12). $\square$

What separates the witnesses needs no reference to a learner. At $a = e_1$ the second coordinate is the window's first bit, readable from lag k alone; at $a = 1^k$ it is the parity of all k , and nothing short of all k positions helps at all. The two ends of the family differ in whether the next symbol admits partial credit, not in how much information it carries. Sparse parity is already the standard hard target in this literature, and power-law scaling has been derived for it (Nam et al., 2024) and observed in data without explicit power-law structure (Barkeshli et al., 2026). What (12) adds is exact entropy matching, which turns the observation into a statement about what a training-free predictor cannot do.

The deployed readout cannot separate them either. On five arms at $k = 16$ with $|a| \in \{1, 2, 4, 8, 16\}$, each rendered so that one process step is one token, $\hat{\alpha}_{\text{LZ}}$ spans 0.1979 to 0.1984: a cross-arm spread of 0.0005 against a start-index noise of 0.0009, so 0.6 of its own noise. The three-point floor agrees to 0.001. The frozen map therefore predicts the same exponent, 0.188, for every arm. Whether the trained exponents agree is a measurement, recorded with the family's construction in Section Q.

B.5 The Character Clock of the Deployed Instrument

The symbolic results are normalized per source symbol. Case 1 instead compresses UTF-8 bytes and divides by a Unicode-character count. The following lemma places that quotient on a declared clock, not on the byte-level certificate.

Lemma B.11 (Character clock of a UTF-8 encoded source). *Let $(X_t)_{t \in \mathbb{Z}}$ satisfy Assumption A.3 on a finite character alphabet, with h_X in bits per character. Let ϕ be the prefix-free UTF-8 map, write $\ell_{\text{utf}}(a) = |\phi(a)| \in \{1, 2, 3, 4\}$, $\bar{\ell}_{\text{utf}} = \mathbb{E} \ell_{\text{utf}}(X_0)$, and let $K_N = \sum_{t \leq N} \ell_{\text{utf}}(X_t)$ be the byte length of $\phi(X_1^N)$. Then:*

- (i) $K_N / N \rightarrow \bar{\ell}_{\text{utf}}$ almost surely and in L^1 .
- (ii) For every lossless binary encoder L of byte strings whose range is prefix-free,

$$\frac{\mathbb{E} |L(\phi(X_1^N))|}{N} \geq \frac{H(X_1^N)}{N} \geq h_X \quad (N \geq 1).$$

- (iii) For prefixes anchored at character boundaries,

$$-\frac{1}{K_N} \log_2 \mathbb{P}(X_1^N) \rightarrow \frac{h_X}{\bar{\ell}_{\text{utf}}} \quad \text{almost surely.}$$

Proof sketch. Birkhoff controls K_N/N , Kraft gives the expected-length inequality, and Shannon–McMillan–Breiman combined with $N/K_N \rightarrow 1/\ell_{\text{utf}}$ gives the anchored byte rate. See [Section N.1](#).

Implication for Case 1. Compressed bits divided by character count are on the character clock of h_X . The inequality above holds in expectation, so coverage for a realized fixed-window DEFLATE curve comes from the byte-level certificate instead, which assumes a stationary byte source and returns bits per byte. Part (iii) converts character-aligned population surprisal, and each guarantee stays in the normalization it was stated in.

C Available Conclusions and Required Information

The results require different kinds of information. Text alone gives empirical corpus scores; a multi-document protocol gives a one-sided entropy certificate. Geometry additionally requires a same-clock dynamical scale; a source-to-head bound further requires a sufficient predictive state and a Hölder factor to it. The table below is a lookup: bring the information in a row’s middle column and that row’s statement is available, whatever the other rows have.

Table 5: **Available conclusions.** Each row lists its required information and the resulting statement. The last two rows are conditional implications, and [Proposition C.2](#) proves that an expanding source never supplies their Hölder factor, so an instance has to come from elsewhere.

Conclusion	Required information	Statement
Empirical corpus scores	Candidate corpus	$(\hat{h}, \hat{a}_{\text{LZ}})$: empirical point estimates, used as rankers
Certified entropy ceiling	Declared source alphabet and clock, target δ_h , and the multi-document protocol of Proposition B.3	$h_X \leq U_h$ with probability at least $1 - \delta_h$
Population geometry	Conformal-source hypotheses; exact h_X and external $\lambda > 0$ on the same declared source clock/log base	$d_{\text{geo}} = h_X / \lambda$
Finite-sample geometry	Conformal-source hypotheses; one-sided $U_h, \lambda_L > 0$ on the same declared source clock/log base, with their marginal guarantees	$d_{\text{geo}} \leq U_h / \lambda_L$ with probability at least $1 - \delta_h - \delta_\lambda$
Direct-head rate	Supplied sufficient state satisfying Assumptions A.7 and A.8	$\alpha_{\text{head}} \geq 2/d_{\text{pred}}$
Population source-to-head	Population-geometry information and a fixed factor (Φ, γ) satisfying Lemma A.10 ; Assumptions A.7 and A.8 additionally for the exponent bound	$d_{\text{pred}} \leq h_X / (\gamma\lambda)$; under those additional assumptions, also $\alpha_{\text{head}} \geq 2/d_{\text{pred}} \geq 2\gamma\lambda/h_X$
Finite-sample source-to-head	Finite-sample geometry information and a fixed factor (Φ, γ) satisfying Lemma A.10 ; Assumptions A.7 and A.8 additionally for the exponent bound	At confidence $1 - \delta_h - \delta_\lambda$: $d_{\text{pred}} \leq U_h / (\gamma\lambda_L)$; under those additional assumptions, also $\alpha_{\text{head}} \geq 2\gamma\lambda_L/U_h$

The final two rows of [Table 5](#) combine a geometric comparison with a neural approximation result. A γ -Hölder map can increase Hausdorff dimension by at most a factor $1/\gamma$; if its image is a sufficient predictive state, a direct prediction head inherits the corresponding dimension-dependent rate.

Theorem C.1 (Conditional source-to-head composition [[Cond. D](#)]). *Let $(\mathcal{M}_{\text{geo}}, T, \mu)$ satisfy the ideal LZ limit and the conformal bridge of [Theorem B.1](#) and [Proposition B.6](#), so that*

$$d_{\text{geo}} := \dim_{\text{H}}(\mathcal{M}_{\text{geo}}) = \frac{h_X}{\lambda},$$

where $\lambda > 0$ is supplied independently on the same clock and in the same logarithmic base as h_X . Let [Assumptions A.7](#) and [A.8](#) hold for a fixed sufficient state S_t with law ν on a $d_{\text{pred}} > 0$ dimensional manifold $\mathcal{M}_{\text{pred}}$. Finally, suppose there is a γ -Hölder sufficient factor $\Phi : \mathcal{M}_{\text{geo}} \rightarrow \mathcal{M}_{\text{pred}}$, $0 < \gamma \leq 1$, satisfying [Lemma A.10](#). Then

$$\alpha_{\text{head}} \geq \frac{2}{d_{\text{pred}}} \geq \frac{2\gamma}{d_{\text{geo}}} = \frac{2\gamma\lambda}{h_X}.$$

Scope. The bound is one-sided and best-in-class, for a head that directly receives S_t : it says what the function class can reach, not what a given optimizer or history encoder reaches. The empirical $\hat{\alpha}_{\text{LZ}} \rightarrow \alpha_D$ calibration rests on trained grids, not on this chain.

Proof sketch. Three links yield the chain: the conformal bridge gives $d_{\text{geo}} = h_X / \lambda$, the Hölder factor gives $d_{\text{pred}} \leq d_{\text{geo}} / \gamma$, and the direct-state theorem gives $\alpha_{\text{head}} \geq 2/d_{\text{pred}}$. Predictive sufficiency puts the head loss above the same Bayes floor h_X . The full argument is in [the deferred oracle proofs](#).

Expansion and prediction pull the state in opposite directions, and that is a theorem rather than a gap in our search.

Proposition C.2 (Expanding sources cannot supply the factor). *Let $(\mathcal{M}_{\text{geo}}, T, \mu)$ satisfy [Assumption A.4](#), so that T is a non-invertible conformal expanding map with a finite generating partition, and let (X_t) be the resulting symbol stream. Write Z_t for the source point at time t . Then Z_t is measurable with respect to $\sigma(X_{t+1}, X_{t+2}, \dots)$ up to a μ -null set. Consequently a state S_t that is past measurable as [Assumption A.7](#) requires and also satisfies $S_t = \Phi(Z_t)$ almost surely as [Lemma A.10](#) requires is measurable with respect to $\sigma(X_{\leq t}) \cap \sigma(X_{> t})$. Whenever the past and the future are independent, as they are for the doubling map with Lebesgue measure, that σ -algebra is trivial, S_t is almost surely constant, and $d_{\text{pred}} = 0$. No expanding source of this kind therefore carries a Hölder factor into a predictive state of positive dimension, which is what the last two rows of [Table 5](#) would need.*

Proof. The source point lies in the future. A generating partition determines Z_t from its forward itinerary $(X_{t+1}, X_{t+2}, \dots)$ up to a μ -null set; for the doubling map the coding is the binary expansion

$$Z_t = \sum_{i \geq 1} X_{t+i} 2^{-i}.$$

So $Z_t \in \sigma(X_{> t}) \bmod \mu$, and so is $\Phi(Z_t)$ for any measurable Φ .

A past-measurable state is then trivial. If also $S_t \in \sigma(X_{\leq t})$, then S_t is measurable with respect to $\sigma(X_{\leq t}) \cap \sigma(X_{> t})$. For independent σ -algebras $\mathcal{F}, \mathcal{G}$, every $A \in \mathcal{F} \cap \mathcal{G}$ satisfies $\mathbb{P}(A) = \mathbb{P}(A)^2$ and so has probability 0 or 1; the intersection is trivial and S_t is almost surely constant. A constant state makes $\mathcal{M}_{\text{pred}}$ a single point, which $d_{\text{pred}} > 0$ in [Assumption A.7](#) excludes. $\square$

Expansion is what makes $d_{\text{geo}} = h_X / \lambda$ available and it is also what puts the source point in the future rather than the past. Prediction needs the opposite. The two halves of [Theorem C.1](#) therefore cannot be joined by a source of the kind [Assumption A.4](#) describes, and the way to join them is to replace the expanding source by an invertible one whose past determines its state, at the price of restating the entropy–dimension bridge for the invertible extension. [Remark A.11](#) records the other direction, where a filter supplies a positive-dimensional predictive law on a generally Cantor support.

Corollary C.3 (Finite-sample propagation). *Assume the conformal-source hypotheses of [Proposition B.6](#), and suppose two procedures have the marginal guarantees*

$$\mathbb{P}(h_X \leq U_h) \geq 1 - \delta_h, \quad \mathbb{P}(\lambda \geq \lambda_L) \geq 1 - \delta_\lambda,$$

where $U_h, \lambda_L > 0$ and $\delta_h + \delta_\lambda < 1$. Then, with probability at least $1 - \delta_h - \delta_\lambda$,

$$d_{\text{geo}} \leq \frac{U_h}{\lambda_L}.$$

If a fixed factor (Φ, γ) satisfying [Lemma A.10](#) is also supplied, then on the same event

$$d_{\text{pred}} \leq \frac{U_h}{\gamma \lambda_L},$$

and if the direct-head assumptions [Assumptions A.7](#) and [A.8](#) also hold, then

$$\alpha_{\text{head}} \geq \frac{2\gamma \lambda_L}{U_h}.$$

No independence is required. For known exact λ , set $\lambda_L = \lambda$ and $\delta_\lambda = 0$. Here γ is fixed; a random lower bound for it would need its own failure event. The entropy certificate and Lyapunov bound must use the same alphabet clock ([Remark A.5](#)).

Proof sketch. Intersect the two marginal events, use the union bound, and substitute the one-sided bounds into whichever of the bridge, factor, and head inequalities have their stated inputs. See [the deferred oracle proofs](#).

The two readouts and the certificate are separate purchases. The instrument of [Section 2](#) returns $(\hat{h}, \hat{\alpha}_{LZ})$ from one pass over the text and ranks corpora with them; when a number with coverage is wanted instead, the multi-document protocol of [Proposition B.3](#) buys $h_X \leq U_h$, and the geometric rows of [Table 5](#) buy the rest once their dynamical inputs are in hand.

D Controlled Checks of the Theory’s Components

On problems with known answers the components behave: switching long-range structure on in a source whose conditional-entropy tail exponent is known by construction moves $\hat{\alpha}_{LZ}$ by six to thirty times its noise floor, the finite-grid residual estimator tracks $1/d$ on four manifold families, 1-NN covering rates land within 1% of $2/d$, and independently measured entropy and expansion return the known dimension on a conformal torus. The nearest-neighbour experiments run a different learner and loss from [Proposition B.4](#): they test the covering-rate mechanism, not end-to-end Transformer scaling. Supporting figures and measurement protocols for the corpus readouts are in [Section P](#); the Case-1 training record is in [Section J](#).

D.1 Controlled Readout Checks with Known Targets

On four controlled manifold families the finite-grid residual estimator follows $\hat{\alpha}_s^{(3)} \approx 1/d$, and the floor-side estimates line up with the known entropy rates. That agreement is an empirical calibration on this panel; coverage is what the U_h procedure of [Proposition B.3](#) supplies.

D.2 The Readout on a Source With a Known Exponent

Switching long-range structure on, with the string held fixed, moves $\hat{\alpha}_{LZ}$ by six to thirty times its own noise floor. The source is the Santa Fe process of [Dębowski \(2011\)](#), the toy source [Dębowski \(2025\)](#) uses to exhibit Hilberg scaling: draw K_t from a Zipf law $\mathbb{P}(K = k) \propto k^{-1/\beta}$ over 10^6 facts and emit the pair (K_t, Z_{K_t}) from a fixed knowledge base of bits. Conditioning on that base makes the process i.i.d., so the entropy rate is exactly $H(K)$, while $H(X_{1:N}) - Nh_X \sim N^\beta$. The conditional-entropy tail exponent is $\gamma_{\text{CE}} = 1 - \beta$, known by construction.

The comparison is paired. We keep the drawn fact sequence, the string template, every character position and every field width, and change one thing: whether a repeated fact carries a repeated answer. At $\gamma_{\text{CE}} = 0.70, 0.50, 0.30, 0.10$ the readout rises by 0.060, 0.014, 0.014 and 0.006 when reuse is switched on, against a window-resampling standard deviation of 0.0006–0.0022 measured by rerunning the same readout at five start-index seeds. Four of four, and the effect grows with the known exponent.

Long-range structure supplies 4% to 19% of the value, and that number sets how the readout is read. Without reuse the same sources still return $\hat{\alpha}_{\text{LZ}}$ between 0.129 and 0.250, from the Zipf marginal and the rendering alone. An absolute $\hat{\alpha}_{\text{LZ}}$ therefore mixes memory with surface form, and holding the surface form fixed is what separates them, which is what the block-order intervention of [Section 4.2](#) does.

One reading of this family stays ambiguous. In the unmodified Santa Fe sources $\hat{\alpha}_{\text{LZ}}$ orders γ_{CE} perfectly across all seven, but $h_X = H(K)$ and β are both set by the same Zipf exponent there, so that ordering alone cannot say which of the two produced it. The paired reuse switch above is what breaks the tie. [Section R](#) gives the full sweep.

D.3 A Covering-Rate Analogue and a Component-Wise Conformal Check

Noiseless 1-NN regression on $\text{Unif}([0, 1]^d)$ recovers the expected covering rate to within 1%: $\hat{\alpha}_{\text{1NN}}d = 2.013 \pm 0.018$ for $d = 1, \dots, 8$. The constant-scale conformal $\times 2$ torus lets the two sides be measured separately. LZ estimates h_X , the known $\lambda = \log_2 2 = 1$ turns it into $\hat{d} = \hat{h}_X / \lambda$, and pairing that source-side estimate with the independent nearest-neighbour curve gives $\hat{\alpha}_{\text{1NN}}\hat{d} = 2.03 \pm 0.02$. Away from sharp spectra the scalar summary fails: three shapes at matched $d_{\text{PR}} = 4$ spread $\hat{\alpha}_{\text{1NN}}$ by 47%, ruling out that scalar as a sufficient description of the tested learner.

Roadmap of this appendix. The four pillars are the head-approximation, local-dimension, conformal entropy–dimension, and non-identification results. Their principal statements are [Propositions B.4](#) and [B.6](#), [Lemma F.1](#), and [Theorem B.7](#). The finite- N coding result in [Proposition B.8](#) supplies separate estimator scope. This appendix supplies the supporting lemmas and scope, with longer proofs in [Section N](#).

E Predictive Dimension Controls Direct-Head Approximation

[Assumptions A.7](#) and [A.8](#) provide the predictive manifold and its Lipschitz logit map. [Assumption A.3](#) is needed only to identify the symbolic cross-entropy reference loss.

Why should predictive dimension affect a model-size exponent? Covering $\mathcal{M}_{\text{pred}}$ at resolution ε costs $\Theta(\varepsilon^{-d_{\text{pred}}})$ balls ([Lemma E.2](#)); a ReLU network with P parameters can approximate at every rate $O_\eta(P^{-1/(d_{\text{pred}}+\eta)})$, $\eta > 0$ ([Lemma E.3](#)); and softmax cross-entropy admits a uniform quadratic upper bound in the logit error. Together these facts give the one-sided, best-in-class bound $\alpha_{\text{head}} \geq 2/d_{\text{pred}}$. They do not determine the exponent of a particular trained model ([Remarks E.4](#) and [E.5](#)).

Remark E.1 (Specialization to Lipschitz ($\beta_{\text{app}} = 1$)). For the remainder of this paper we specialize to the Lipschitz case $\beta_{\text{app}} = 1$, as stated in [Assumption A.8](#). If each target coordinate instead admits an ambient extension in the Hölder class of order β_{app} (with the usual derivative interpretation when $\beta_{\text{app}} > 1$), the same argument gives the sup-norm rate $O_\eta(P^{-\beta_{\text{app}}/(d_{\text{pred}}+\eta)})$ for every $\eta > 0$ and $\alpha_{\text{head}} \geq 2\beta_{\text{app}}/d_{\text{pred}}$ in cross-entropy. That extension hypothesis is automatic for $\beta_{\text{app}} = 1$ by McShane’s theorem, but is additional structure for $\beta_{\text{app}} > 1$; the Lipschitz restriction keeps the paper’s assumptions uniform.

Lemma E.2 (Kolmogorov-Tihomirov Metric Entropy). *Let $\mathcal{M}$ be a compact d -dimensional C^1 embedded Riemannian manifold. Its minimum ε -cover size satisfies*

$$\mathcal{N}_{\text{cov}}(\varepsilon; \mathcal{M}) = \Theta(\varepsilon^{-d}) \quad \text{as } \varepsilon \rightarrow 0.$$

Proof. Compactness and a finite smooth atlas give constants $r_0, c_-, c_+ > 0$ such that, for every $x \in \mathcal{M}$ and $0 < r < r_0$,

$$c_-r^d \leq \mathcal{H}^d(B(x, r) \cap \mathcal{M}) \leq c_+r^d. \quad (14)$$

The same bounds hold whether the balls are taken in the intrinsic metric or in the ambient metric, after changing the constants, because the two metrics are uniformly locally bi-Lipschitz on a compact embedded manifold.

Let $\mathcal{N}_{\text{cov}}(\varepsilon; \mathcal{M})$ denote the least number of radius- ε balls covering $\mathcal{M}$. If J such balls cover $\mathcal{M}$, discard any that miss $\mathcal{M}$ and pick a point $x_j \in \mathcal{M}$ inside each of the balls that remain. A remaining ball has radius ε and contains x_j , so its intersection with $\mathcal{M}$ lies in $B(x_j, 2\varepsilon) \cap \mathcal{M}$. The upper bound in (14) and subadditivity of $\mathcal{H}^d$ give

$$\mathcal{H}^d(\mathcal{M}) \leq Jc_+(2\varepsilon)^d,$$

and hence $\mathcal{N}_{\text{cov}}(\varepsilon; \mathcal{M}) \geq \mathcal{H}^d(\mathcal{M}) / (2^d c_+ \varepsilon^d)$. For the reverse inequality, choose a maximal ε -separated set $\{x_1, \dots, x_J\} \subset \mathcal{M}$. Maximality makes the radius- ε balls about these points a cover, whereas the radius- $\varepsilon/3$ balls are pairwise disjoint. The lower bound in (14) therefore yields

$$Jc_-(\varepsilon/3)^d \leq \mathcal{H}^d(\mathcal{M}), \quad \mathcal{N}_{\text{cov}}(\varepsilon; \mathcal{M}) \leq J \leq \frac{3^d \mathcal{H}^d(\mathcal{M})}{c_- \varepsilon^d}.$$

Both estimates hold for all sufficiently small ε , proving the asserted $\Theta(\varepsilon^{-d})$ rate; see also [Kolmogorov & Tihomirov \(1961\)](#). $\square$

Lemma E.3 (ReLU approximation rate on manifolds). *For a positive integer P , let $\mathcal{N}_P$ denote the scalar-output deep ReLU networks with at most P trainable parameters. For $P \geq |\Sigma|$, let*

$$\mathcal{F}_P := \left\{ (f_i)_{i \in \Sigma} : f_i \in \mathcal{N}_{P_i} \text{ for integers } P_i \geq 1, \sum_{i \in \Sigma} P_i \leq P \right\};$$

thus P is the total, rather than per-coordinate, budget. Under [Assumptions A.7](#) and [A.8](#):

- (a) (Upper bound) *For every $\eta > 0$, there is an $f_P \in \mathcal{F}_P$ such that*

$$\sup_{s \in \mathcal{M}_{\text{pred}}} \|f_P(s) - \Lambda(s)\|_\infty = O_\eta \left(P^{-1/(d_{\text{pred}} + \eta)} \right)$$

in the Lipschitz case. More generally, if each target coordinate extends to an ambient Hölder class of smoothness order β_{app} , the exponent 1 is replaced by β_{app} .

- (b) (Sup-norm parameter-count lower bounds, [Yarotsky 2017](#)) *For an integer smoothness order $\beta_{\text{app}} \geq 1$ and the unit ball of $W^{\beta_{\text{app}}, \infty}([0, 1]^{d_{\text{pred}}})$ in $\|\cdot\|_\infty$: (b1, unconditional, Thm 4a) any fixed ReLU architecture achieving error ε uniformly over this ball has $P = \Omega(\varepsilon^{-d_{\text{pred}}/(2\beta_{\text{app}})})$; (b2, bounded depth, Thm 4b) for fixed $C_{\text{depth}} > 0$ and $p_{\text{dep}} \geq 0$, any fixed ReLU architecture of depth $L_{\text{net}} \leq C_{\text{depth}} \log^{p_{\text{dep}}}(1/\varepsilon)$ that achieves error ε uniformly over this ball has $P = \Omega(\varepsilon^{-d_{\text{pred}}/\beta_{\text{app}}} \log^{-2p_{\text{dep}}-1}(1/\varepsilon))$. These are class-uniform (minimax) bounds over the Sobolev ball, not per-target bounds.*

Proof sketch. Part (a) is cover-and-glue on local charts ([Nakada & Imaizumi, 2020](#)); part (b) is VC-dimension counting ([Yarotsky, 2017](#)), whose quadratic parameter-count growth is where the factor-two gap of [Remark E.4](#) originates. The full argument is in [Section N](#).

Together, the lemmas yield the main-text result [Proposition B.4](#): the sup-norm approximation rate $O_\eta(P^{-1/(d_{\text{pred}} + \eta)})$ becomes $\alpha_{\text{head}} \geq 2/d_{\text{pred}}$ in cross-entropy. The proof applies [Lemma E.3\(a\)](#) and the softmax's quadratic bound without introducing a probability-floor-dependent constant ([Remark A.9](#)); see [Section N.1](#). The following remarks separate this achievability result from a two-sided law and from finite-resolution spectral heuristics.

Remark E.4 (Approximation-Error Lower Bound and the Loss Exponent). The Yarotsky/DeVore approximation-error bounds (sup-norm, parameter count) do not transfer linearly to the loss exponent.

- (i) *Approximation-error metric.* [Lemma E.3\(b1\)](#) ([Yarotsky 2017](#), Thm 4a) gives $P = \Omega(\varepsilon^{-d_{\text{pred}}/2})$ for Lipschitz targets, so the minimax sup-norm error has decay order at most $2/d_{\text{pred}}$, while part (a) certifies every decay order below $1/d_{\text{pred}}$. The resulting interval of possible endpoint orders is therefore $[1/d_{\text{pred}}, 2/d_{\text{pred}}]$. The interval closes to the endpoint $1/d_{\text{pred}}$ under bounded (poly-log) depth (Thm 4b) or continuous weight selection ([DeVore](#)

et al., 1989; Siegel, 2024), neither of which is automatic for arbitrarily-trained unbounded-depth nets (Yarotsky, 2017).

(ii) *Transfer to the loss exponent.* The softmax makes excess cross-entropy quadratic in logit error, which is what turns an approximation rate into a loss rate:

$$\text{KL}_2 \leq \frac{\varepsilon^2}{2 \ln 2} \implies \alpha_{\text{head}} \geq \frac{2}{d_{\text{pred}}} \quad (\text{Proposition B.4}).$$

This *strengthens* achievability from $1/d_{\text{pred}}$ to the proven $2/d_{\text{pred}}$ but does *not* pin it from above: a trained network can decay faster (favourable loss geometry) or slower (optimization slack), consistent with the $\alpha_{\text{head}} \propto 1/d_{\text{pred}}$ scaling of the empirical literature. Sharma–Kaplan’s predicted loss exponent $4/d_{\text{pred}}$ (Sharma & Kaplan, 2022) is itself a *theory* prediction (smooth-target manifold regression, an order-2 approximation rate squared by the quadratic loss), so the factor-2 gap from our Lipschitz-target $2/d_{\text{pred}}$ floor is explained by target smoothness ($\beta_{\text{app}}=2$ vs. $\beta_{\text{app}}=1$, Remark E.1). The quadratic loss supplies the common multiplier (Bahri et al., 2024).

(iii) *Scope.* The result is best-in-class: it concerns the best P -parameter model, not any particular trained run. The fitted data-scaling exponent is the memory-governed quantity of Remark B.9.

(iv) *Endpoint.* For each $\eta > 0$ the construction has order $P^{-2/(d_{\text{pred}}+\eta)}$, with a constant depending on η , the logit Lipschitz constant, and $|\Sigma|$. This does not give a single bound at the endpoint $P^{-2/d_{\text{pred}}}$, nor does it determine the loss magnitude.

Remark E.5 (Scope of the proportionality: a sharp-spectrum hypothesis beyond the exact-manifold model). The predictive reading that d_{pred} alone fixes a trained-run exponent through $\alpha_{\text{head}} \propto 1/d_{\text{pred}}$ is not a corollary (Remark E.4). On the exact compact $C^{k_{\text{pred}}}$ manifold of Assumption A.7, upper Minkowski, Hausdorff, and intrinsic dimension coincide at d_{pred} (proof of Lemma E.3); the theorem needs no covariance spectrum.

A different, finite-resolution question is whether a spectral summary of an approximately low-dimensional cloud predicts the rate of a specified learner. The following is a heuristic, not part of Proposition B.4. Order the local covariance or kernel eigenvalues as $\zeta_1 \geq \zeta_2 \geq \dots$. A sharp gap separating a small set of resolved directions from a negligible tail is one sufficient scenario in which an effective direction count might summarize the spectrum. Turning that scenario into an inverse-dimension learning law would still require a separate theorem specifying the learner, sampling regime, resolution, and loss. With a slowly decaying spectrum, the whole eigenvalue profile may matter instead.

The fixed-participation-ratio experiment in Section D illustrates only this last possibility: within the tested construction, two spectral shapes matched on d_{PR} produce a 47% spread in $\hat{\alpha}_{1\text{NN}}$. It shows that this scalar is insufficient there. The arms are full rank in $\mathbb{R}^{24}$, so their exact manifold dimension and their finite-resolution participation ratio are also different objects. The analogy with the non-conformal Lyapunov-vector obstruction is therefore structural rather than a theorem transferring one setting to the other.

F Local Dimension of the Source Measure

The conformal bridge uses the local mass-scaling exponent of the source measure. Under Assumption A.1, this exponent agrees with the manifold dimension.

Lemma F.1 (Local Dimension Constancy). *Under Assumption A.1, the local dimension*

$$d_{\mu}(x) := \lim_{r \rightarrow 0} \frac{\log \mu(B(x, r))}{\log r}$$

exists and equals d_{geo} for μ -almost every $x \in \mathcal{M}_{\text{geo}}$.

Proof. Let $f = d\mu/d(\mathcal{H}^{d_{\text{geo}}}|\mathcal{M}_{\text{geo}})$. By [Assumption A.1](#), $c_1 \leq f \leq c_2$ almost everywhere. Combining this with the uniform small-ball estimate [\(14\)](#), for every $x \in \mathcal{M}_{\text{geo}}$ and every sufficiently small r we have

$$c_1 c_- r^{d_{\text{geo}}} \leq \mu(B(x, r)) \leq c_2 c_+ r^{d_{\text{geo}}}.$$

Consequently $\log \mu(B(x, r)) = d_{\text{geo}} \log r + O(1)$, with the $O(1)$ uniform in x . Dividing by $\log r$ and letting $r \downarrow 0$ proves $d_\mu(x) = d_{\text{geo}}$ (in fact, at every $x \in \mathcal{M}_{\text{geo}}$). The same local-volume estimate is discussed in [Mattila \(1995\)](#). $\square$

G From Symbolic Recurrence to Geometric Dimension

[Assumption A.3](#) supplies the symbolic entropy limit; [Assumptions A.1, A.2](#) and [A.4](#) supply the conformal geometric bridge. Results marked **[Cond. D]** require the smooth conformal-expanding structure of [Assumption A.4](#).

Three similar-looking recurrence quantities play different roles. The block return time R_ℓ and LZ match length L_N recover the symbolic entropy rate, whereas the metric return time τ_r recovers local geometric dimension. Conformality and the Rokhlin–Pesin identity, not a formal comparison of these estimators, put the symbolic and geometric quantities on one clock and give $d_{\text{geo}} = h_X/\lambda$.

Concretely, symbolic block returns grow like $2^{h_X \ell}$, while metric returns to radius r grow like r^{-d_μ} . A direct cylinder-to-ball comparison would require additional Markov-partition and distortion hypotheses and is not used here. The bridge uses the population limit $\lim_{w \rightarrow \infty} h_w$; finite- N compression rates are treated separately in [Remark G.5](#).

Lemma G.1 (Ornstein–Weiss Theorem ([Ornstein & Weiss, 1993](#))). *Under [Assumption A.3](#), the return-time complexity satisfies*

$$\frac{\log_2 R_\ell(x)}{\ell} \xrightarrow{\text{a.s.}} h_X,$$

where $R_\ell(x)$ is the first return time of the ℓ -block starting at x , and h_X is the entropy rate of [Assumption A.3](#). Moreover, if $M_N(x)$ is the LZ78 phrase count, then $M_N(x) \log_2 M_N(x) / N \rightarrow h_X$ a.s. ([Ziv & Lempel, 1978](#)).

Proof. Apply the Ornstein–Weiss recurrence theorem to the finite-alphabet stationary ergodic shift specified by [Assumption A.3](#). Its cylinder return time is precisely the random variable R_ℓ in the statement, so the theorem gives $\ell^{-1} \log_2 R_\ell \rightarrow h_\mu(\sigma) = h_X$ almost surely, including when $h_X = 0$. The parsing assertion is the coding theorem of [Ziv & Lempel \(1978\)](#) applied to the same shift. It is a separate consequence of stationarity and ergodicity, rather than a use of Kac’s mean-return identity as a pathwise approximation. $\square$

The next lemma records the equivalent match-length formulation of symbolic recurrence; the expected-rate limit and its order of limits are in [Theorem B.1](#).

Lemma G.2 (Match-length/recurrence-time duality ([Wyner & Ziv, 1989](#))). *Under [Assumption A.3](#), the LZ77 match length $L_N(x)$, the length of the longest prefix of $(x_{N+1}, x_{N+2}, \dots)$ occurring as a substring of x_1^N , satisfies*

$$\frac{\log_2 N}{L_N(x)} \xrightarrow{\mathbb{P}} h_X, \quad \frac{1}{\ell} \log_2 R_\ell(x) \xrightarrow{\text{a.s.}} h_X,$$

where R_ℓ is the waiting time until the first ℓ -block recurs. Thus match length is the inverse statistic associated with symbolic recurrence time. If $h_X > 0$, the first convergence is equivalently $L_N / \log_2 N \xrightarrow{\mathbb{P}} 1/h_X$. For $h_X = 0$, only the reciprocal formulation is asserted; division by the entropy rate is not defined.

Proof. The waiting-time limit is the Ornstein–Weiss theorem for cylinder sets, in the formulation of [Ornstein & Weiss \(1993\)](#); [Kontoyiannis \(1998\)](#). The match-length theorem of [Wyner &](#)

Ziv (1989) gives $\log_2 N / L_N \rightarrow h_X$ in probability for every stationary ergodic finite-alphabet source. When $h_X > 0$, the map $u \mapsto 1/u$ is continuous at h_X , so the continuous-mapping theorem yields $L_N / \log_2 N \rightarrow 1/h_X$ in probability. No reciprocal is taken in the zero-entropy case. $\square$

Lemma G.3 (Recurrence-Dimension Identity (Barreira & Saussol, 2001) [Cond. D]). *Under Assumption A.4, the metric return time $\tau_r(x) := \inf\{t \geq 1 : \rho_{\text{geo}}(T^t x, x) < r\}$ satisfies, for μ -a.e. x ,*

$$\lim_{r \downarrow 0} \frac{\log_2 \tau_r(x)}{-\log_2 r} = d_\mu(x) = d_{\text{geo}},$$

a genuine limit (lower and upper recurrence rates coincide).

Proof. For an ergodic equilibrium measure of a Hölder potential on a $C^{1+\beta_{\text{dyn}}}$ conformal expanding system, the quantitative recurrence theorem of Barreira & Saussol (2001) identifies the lower and upper recurrence exponents with the corresponding lower and upper pointwise dimensions:

$$\liminf_{r \downarrow 0} \frac{\log_2 \tau_r(x)}{-\log_2 r} = \underline{d}_\mu(x), \quad \limsup_{r \downarrow 0} \frac{\log_2 \tau_r(x)}{-\log_2 r} = \bar{d}_\mu(x)$$

for μ -almost every x . Under Assumption A.4, the ACIP is the equilibrium (Gibbs) measure for the Hölder geometric potential $-\ln |\det D_x T|$; Hölder regularity follows from $T \in C^{1+\beta_{\text{dyn}}}$. Thus the recurrence theorem applies. If one uses a formulation stated for a topologically mixing expanding map, apply it to the mixing components of a finite cyclic decomposition and to the corresponding fixed iterate $T^{n_{\text{cyc}}}$; multiplying return times by the fixed factor n_{cyc} does not change their logarithmic exponent. By Lemma F.1, $\underline{d}_\mu(x) = \bar{d}_\mu(x) = d_{\text{geo}}$ almost everywhere, so the displayed liminf and limsup coincide and equal d_{geo} . $\square$

Lemma G.4 (Pesin Entropy Equality, Conformal ACIP Case (Rokhlin, 1967; Pesin, 1977; Young, 1982; Ledrappier & Young, 1985) [Cond. D]). *Under Assumption A.4 (conformal expanding map with absolutely continuous invariant measure $\mu \ll \mathcal{H}^{d_{\text{geo}}} |_{\mathcal{M}_{\text{geo}}}$), the pointwise dimension is constant and equals the entropy/exponent ratio:*

$$d_\mu(x) = \frac{h_X}{\lambda} = d_{\text{geo}} \quad (\mu\text{-a.e.}),$$

where h_X is the symbolic entropy rate, equal under the generating code to the measure-theoretic entropy $h_\mu(T)$; λ is the conformal Lyapunov exponent of Assumption A.4.

Proof. The Rokhlin–Pesin entropy formula for the absolutely continuous invariant measure in Assumption A.4 gives

$$h_\mu(T) = \int_{\mathcal{M}_{\text{geo}}} \log_2 |\det D_x T| d\mu(x).$$

Conformality implies $|\det D_x T| = \|D_x T\|^{d_{\text{geo}}}$, and hence

$$h_\mu(T) = d_{\text{geo}} \int_{\mathcal{M}_{\text{geo}}} \log_2 \|D_x T\| d\mu(x) = d_{\text{geo}} \lambda.$$

The partition in Assumption A.4 is generating, so its symbolic entropy rate h_X equals $h_\mu(T)$. Since $\lambda > 0$, division gives $h_X / \lambda = d_{\text{geo}}$; combining this identity with Lemma F.1 yields $d_\mu(x) = h_X / \lambda = d_{\text{geo}}$ for μ -almost every x . $\square$

Proposition B.6 ($d_{\text{geo}} = h_X / \lambda$, conformal case [Cond. D]) is proved in the body (Section B) from this lemma, which also supplies the pointwise statement. The remarks that follow delimit its scope: the finite-window bias, the fractal Gibbs extension, and the non-conformal Lyapunov vector.

Remark G.5 (What LZ Measures: The Entropy Side of the Bridge). [Proposition B.6](#) uses the population identity $h_X = \lim_{w \rightarrow \infty} h_w$ together with the supplied scale λ ; it does not use a finite- N residual exponent. The ideal code provides h_X , and the geometric dynamics provide its conversion to dimension. The deployed DEFLATE curve at $w_0 = 32,768$ is a different statistic: it is not identified with h_{w_0} and has no coverage statement here. Finite-sample coverage belongs instead to the wrapped statistic U_h under the independent-document and random-offset protocol of [Proposition B.3](#). The non-identification result in [Section H](#) shows why a symbolic statistic alone cannot supply the missing geometric scale uniformly over its stated class.

Remark G.6 (General Gibbs / information-dimension extension). The absolute-continuity hypothesis of [Assumption A.4](#) is what pins d_{geo} to the *integer* manifold dimension. If we drop it and instead let μ be a general ergodic equilibrium (Gibbs) state of a Hölder potential on a $C^{1+\beta_{\text{dyn}}}$ conformal expanding repeller, then Ruelle's inequality gives $h_X \leq d_{\text{geo}}\lambda$. Under the usual rigidity hypotheses, equality is the ACIP/SRB case; non-ACIP Gibbs states are singular with respect to $\mathcal{H}^{d_{\text{geo}}}|\mathcal{M}_{\text{geo}}$ and satisfy the strict inequality $h_X < d_{\text{geo}}\lambda$. The *same* ratio identity survives, but now reads the *measure / information dimension* rather than the manifold dimension:

$$\dim_{\text{H}}(\mu) = d_{\mu}(x) = \frac{h_X}{\lambda} \quad (\mu\text{-a.e.}),$$

by the Ledrappier–Young formula in the conformal case ([Ledrappier & Young, 1985](#); [Young, 1982](#)). Here $\dim_{\text{H}}(\mu) = h_X/\lambda \leq d_{\text{geo}}$, with equality iff μ is the ACIP of [Assumption A.4](#). Within this conformal Gibbs class the population Rényi dimensions satisfy $D_2 \leq D_1$, and the ratio reads the information dimension D_1 , not an embedding dimension. A finite-sample empirical estimate $\hat{D}_2$, however, is not a certified lower bound on D_1 without an error analysis. If a conformal Gibbs model were justified for language belief dynamics, the relevant population object would likewise be D_1 ; that model identification is not made here. [Proposition B.6](#) remains the absolutely continuous special case.

Remark G.7 (Non-conformal expanding case: the Ledrappier–Young vector formula). The identity $\dim_{\text{H}} \mu = h_X/\lambda$ needs conformality: a single exponent λ expanding every direction equally. For a $C^{1+\beta_{\text{dyn}}}$ non-conformal expanding system the distinct exponents $\lambda^{(1)} > \lambda^{(2)} > \dots$ are anisotropic and must *not* be collapsed into one λ . Under the hypotheses of the Ledrappier–Young dimension formula, the dimension is instead a *sum of partial dimensions* over the expanding Oseledets subspaces,

$$\dim_{\text{H}} \mu = \sum_i \vartheta_i, \quad \vartheta_i = \frac{H_i - H_{i-1}}{\lambda^{(i)}},$$

with $\lambda^{(i)}$ the distinct (unstable) exponents and $H_i - H_{i-1} \geq 0$ the entropy increments subordinate to the i -th invariant foliation (so $\sum_i (H_i - H_{i-1}) = h_X$, $0 \leq \vartheta_i \leq \dim E^{(i)}$). Conformality ([Assumption A.4](#)) is the degenerate case with all $\lambda^{(i)} \equiv \lambda$, collapsing to h_X/λ ; see [Ledrappier & Young \(1985\)](#); [Barreira et al. \(1999\)](#) for the hyperbolic product/dimension structure. For an invertible hyperbolic diffeomorphism, the displayed sum describes the conditional dimension on unstable leaves; the total pointwise dimension also contains the stable conditional contribution, obtained from the inverse dynamics. It is therefore not equal to the unstable sum without an additional hypothesis. Hence the LZ-based oracle $d_{\text{geo}} = h_X/\lambda$ of [Proposition B.6](#) is a genuinely *conformal* reading of h_X ; non-conformal geometry carries strictly more information than this ratio.

H Why Symbols Do Not Identify Geometric Dimension

[Theorem B.7](#) gives the exact equal-law non-identification result in the body. This section quantifies its irreducible error and its degradation when two symbolic laws nearly collide. The finite- N LZ hierarchy in [Proposition B.8](#) is a separate coding result; neither statement turns a symbolic residual into a geometric observable.

Theorem H.1 (Quantitative obstruction: an irreducible minimax error). *Fix integers $q, k \geq 2$ and take the corresponding witnesses W_1, W_2 constructed in the proof of [Theorem B.7](#), with*

dimensions 1 and k , respectively. For every $N \geq 1$ and every measurable estimator $\hat{d}_N : \{0, \dots, q^k - 1\}^N \rightarrow \mathbb{R}$ that reads the token stream alone,

$$\max \left\{ \mathbb{E}_{W_1} |\hat{d}_N - 1|, \mathbb{E}_{W_2} |\hat{d}_N - k| \right\} \geq \frac{k-1}{2},$$

with equality for the constant estimator $\hat{d}_N \equiv \frac{1+k}{2}$. Taking $q = 2$, the family of 2^k -letter witnesses has irreducible error $(k-1)/2 \rightarrow \infty$: no token-stream estimator is uniformly consistent for d_{geo} over the conformal-expanding class.

Proof. By [Theorem B.7](#) the stream $X_{1:N}$ has one common law under W_1 and W_2 (i.i.d. uniform on q^k letters), so $\hat{d}_N(X_{1:N})$ has a single law $\mathbf{Q}_N$ on $\mathbb{R}$ and

$$\mathbb{E}_{W_i} |\hat{d}_N - d_i| = \int |t - d_i| d\mathbf{Q}_N(t), \quad (15)$$

with the two dimensions $d_1 = 1, d_2 = k$.

A maximum dominates an average, and $|t-1| + |t-k| \geq k-1$ for all t (triangle inequality, $k \geq 2$), so (15) gives

$$\max_{i \in \{1,2\}} \mathbb{E}_{W_i} |\hat{d}_N - d_i| \geq \frac{1}{2} \int (|t-1| + |t-k|) d\mathbf{Q}_N(t) \geq (k-1)/2, \quad (16)$$

At $t = (1+k)/2$ both terms equal $(k-1)/2$, so the constant estimator named in the statement attains the bound. The estimate is uniform in N because the two observation laws coincide; this is precisely the degenerate Le Cam two-point experiment. $\square$

The obstruction does not hinge on the exact symbolic collision: the bound degrades continuously in the separation between the two observation laws, which converts the impossibility statement into a sample-size requirement for any pair of sources that merely *nearly* collide.

Proposition H.2 (Robust obstruction and a sample-size floor). *Let W_1, W_2 satisfy [Assumptions A.1, A.2](#) and [A.4](#) with geometric dimensions d_1, d_2 and gap $\Delta = |d_1 - d_2| > 0$. Write $\mathbb{P}_i^{(N)}$ for the law of $X_{1:N}$ under W_i and $V_N = \text{TV}(\mathbb{P}_1^{(N)}, \mathbb{P}_2^{(N)})$. For every $N \geq 1$ and every (possibly randomized) estimator $\hat{d}_N$ that reads the token stream alone,*

$$\max_{i \in \{1,2\}} \mathbb{E}_{W_i} |\hat{d}_N - d_i| \geq \frac{\Delta}{2} (1 - V_N). \quad (17)$$

If moreover the symbolic laws have natural-log divergence rate at most δ , in the sense that $\text{KL}_{\text{nat}}(\mathbb{P}_1^{(N)} \parallel \mathbb{P}_2^{(N)}) \leq N\delta$ for all N , then

$$\max_{i \in \{1,2\}} \mathbb{E}_{W_i} |\hat{d}_N - d_i| \geq \frac{\Delta}{2} \left(1 - \sqrt{N\delta/2}\right), \quad (18)$$

so the risk remains at least $\Delta/4$ for every $N \leq 1/(2\delta)$. Therefore, any procedure whose minimax expected absolute error on this pair is strictly below $\Delta/4$ must use

$$N > \frac{1}{2\delta} \quad (19)$$

symbols. This necessary condition is independent of the estimator. Setting $\delta = 0$ returns [Theorem H.1](#).

Proof. Let $f_i^{(N)}$ be the counting density of $\mathbb{P}_i^{(N)}$ on Σ^N , and let $\xi_N := \mathbb{P}_1^{(N)} \wedge \mathbb{P}_2^{(N)}$ be the measure with counting density $\min(f_1^{(N)}, f_2^{(N)})$. Then $\xi_N(\Sigma^N) = 1 - V_N$ and $\mathbb{P}_i^{(N)} \geq \xi_N$ for $i = 1, 2$. Because $|t - d_1| + |t - d_2| \geq \Delta$ for every $t \in \mathbb{R}$,

$$\mathbb{E}_{W_1} |\hat{d}_N - d_1| + \mathbb{E}_{W_2} |\hat{d}_N - d_2| \geq \int (|\hat{d}_N - d_1| + |\hat{d}_N - d_2|) d\xi_N \geq \Delta (1 - V_N),$$

and a maximum dominates half the sum, giving (17); for randomized $\hat{d}_N$ apply the display conditionally on the randomization seed and average. Pinsker's inequality gives $V_N \leq \sqrt{\text{KL}_{\text{nat}}(\mathbb{P}_1^{(N)} \parallel \mathbb{P}_2^{(N)})/2} \leq \sqrt{N\delta/2}$, which is (18); and $\sqrt{N\delta/2} \leq \frac{1}{2}$ exactly when $N \leq 1/(2\delta)$. $\square$

Two readings matter. First, the witnesses of Theorem H.1 sit at $\delta = 0$, where (19) is vacuous and the floor $\Delta/2$ holds at every N : that is the impossibility. Second, (18) is informative only while $N\delta < 2$, so for sources that are genuinely distinguishable it is a small-sample statement rather than an impossibility, and it says what the finite-corpus cost of a near-collision is. Neither reading licenses a dimension estimate from symbols; together they show the boundary is a neighborhood rather than a single point.

Remark H.3 (Status of the three-point estimators). On an admissible geometric three-point grid, the explanatory estimators recover the intercept and residual exponent exactly when the measured values follow a nonconstant pure-power template. Section P.4 gives their definitions, domain conditions, and full algebra.

The coding results do not validate this template for the deployed inputs C_i : Proposition B.8 is code- and source-class-specific. Thus $\hat{a}_s^{(3)}$ remains a descriptive finite-grid fit. The general separation among the empirical rankers, the certified U_h , and the conformal population identity is stated in Remark G.5.

Which side of h_X does a three-point floor land on? Write $a_i = C(N_i) - h_X$ and $\Delta_a = a_1 + a_3 - 2a_2$. If $\Delta_a = 0$, the extrapolant is undefined. If $\Delta_a \neq 0$, direct subtraction gives the exact identity

$$\hat{H}_0^{(3)} - h_X = \frac{a_1 a_3 - a_2^2}{\Delta_a},$$

and therefore

$$\text{sgn}(\hat{H}_0^{(3)} - h_X) = \text{sgn}(a_1 a_3 - a_2^2) \text{sgn}(\Delta_a).$$

If $a_1, a_2, a_3 > 0$ and $\Delta_a > 0$, log-convexity on the grid ($a_2^2 \leq a_1 a_3$) implies $\hat{H}_0^{(3)} \geq h_X$; when $\Delta_a < 0$ the implication is reversed, and for residuals of mixed sign log-convexity is not defined. No theorem here supplies those hypotheses for the deployed curve. Which side of h_X the three-point floor lands on therefore remains open. In a PPM screening sweep (Section S.1) the raw extrapolant for `opc_synth` is negative and the clamped $\hat{h}$ is exactly 0. The clamp never fired on a reported number: no floor in any deployed panel sits at or below 0, the smallest being 0.169.

I Compression measurement details

Remark I.1 (Total-rate slopes and entropy tails). For a stationary finite-alphabet source on one fixed alphabet and clock, let $H_n^{\text{cond}} = H(X_{n+1} \mid X_{1:n})$. The chain rule gives

$$\frac{H(X_{1:N})}{N} - h_X = \frac{1}{N} \sum_{n=0}^{N-1} (H_n^{\text{cond}} - h_X).$$

Thus a tail $H_n^{\text{cond}} - h_X \asymp n^{-\gamma_{\text{CE}}}$ with $0 < \gamma_{\text{CE}} < 1$ transfers that exponent to the ideal *residual* block-entropy rate. If an idealized total rate obeys $C(N) = H + AN^{-\gamma_{\text{CE}}}$, its positive local log-log slope is

$$-\frac{d \log C(N)}{d \log N} = \gamma_{\text{CE}} \frac{AN^{-\gamma_{\text{CE}}}}{H + AN^{-\gamma_{\text{CE}}}}.$$

For $H, A > 0$, the floor attenuates the slope. Attenuation alone therefore cannot explain a compression fit larger than the entropy-tail exponent. The deployed eight-point estimate also depends on the source, sample, compressor, finite window, grid, alphabet, and clock. Outside this idealization, the calculation fixes neither the sign nor the magnitude of its gap from γ_{CE} . In particular, the raw-rate slope $\hat{a}_{\text{LZ}}$, the template exponent α_s in (1), and the conditional-entropy-tail exponent are distinct measurements.

Compressor comparisons. The four LZ-slope variants agree with one another at $\rho \geq 0.93$ (≥ 0.81 after controlling for $\hat{h}$), and each ranks α_D at $\rho \geq 0.63$ on all three corpus sets. The compressor and grid were selected on the Case 1 calibration panel; [Section S.1](#) and [Table 9](#) reports the full sweep. The prospective family test evaluates the selected estimator without refitting.

The exploratory level-rate plane. [Figure 1](#) plots a three-point analogue of the level-rate pair to display how text types occupy different regions of the plane. The deployed Case 1 estimator is the higher-sample, eight-point map in [Table 1](#).

J Case-1 Evidence and Reproducibility

This section answers four questions about Case 1: how the neural targets were measured, how the calibration changes across corpus sets and model scales, whether the exponent map transfers to new parent families, and whether changing sequence order moves both the readout and the fitted exponent. Data sources and preprocessing are given in [Section P](#), and compressor comparisons in [Section S.1](#).

Prediction metrics. MAE is the mean absolute difference between a prediction and the fitted exponent. The usual squared-error prediction coefficient is

$$Q^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \hat{y}_i^{\text{ref}})^2}.$$

Its value depends on the reference predictor and on how predictions are pooled across folds, so we fix both: the reference $\hat{y}_i^{\text{ref}}$ is the panel mean $\bar{\alpha}_D = 0.1828$, and held-out predictions are pooled over folds before scoring. Every leave-one-parent-family-out number below is recomputed under that definition from the frozen 38-row calibration artifact by `pre-train/scripts/case1_lofo_recompute.py`, which refits the predictor with one parent family removed, predicts that family’s rows, pools, and scores. The reconstruction returns the calibration line $1.1821\hat{\alpha}_{\text{LZ}} - 0.0461$, in-sample $R^2 = 0.635$ and $\rho = 0.845$, and the five frozen domain means to six digits, so the artifact and the reported calibration agree. One number does not reproduce exactly: the domain-mean baseline’s Q^2 is 0.317 here against 0.332 as first recorded. The panel’s other domain has a single member, so the fold that removes it has no in-fold domain mean, and the two runs fall back differently.

J.1 Full Compression Measurement Protocol

Training target. Let P be the parameter count and D the training-token budget. The training grids fit the additive scaling model

$$L(P, D) = E + A_P P^{-\alpha_P} + A_D D^{-\alpha_D}. \quad (20)$$

Our target is the fitted α_D , which depends on the model family, tokenizer, training schedule, and range of budgets. At fixed P , the model-size term is constant; the sensitivity analysis instead fits $L(D) = E_P + A'_P D^{-\alpha_{D,P}}$ separately at each size with a free floor. The primary grids use GPT-2-style models from 16M to 1B parameters, sequence length 512, and token budgets $D \in \{12, 40, 130, 250, 700\}$ million. Longer runs at 2–20B tokens support the separate floor analysis. [Section J.2](#) gives the training schedules. The two exponent analyses are reported separately in [Sections 4](#) and [J.4](#). A larger exponent implies faster reduction of excess loss when the floor and amplitude are held fixed.

Compression measurement. We concatenate the corpus documents and sample character windows with replacement, uniformly by starting position. For a window $x_{i:i+N}$, its compressed rate is

$$c_N(x, i) = \frac{8 \lfloor \text{Deflate}_{9,-15}(\text{UTF8}(x_{i:i+N})) \rfloor}{N} \quad \text{bits/character.}$$

Level 9 and `wbits = -15` specify headerless DEFLATE with a 32,768-byte LZ77 window. Write C_N for the mean rate at length N . We define the compression slope as

$$\hat{a}_{LZ} = |\text{OLS slope}(\log C_N \text{ on } \log N)|, \quad N \in \{64, 128, \dots, 8192\}. \quad (21)$$

The fit uses the total rates. Longer 16k and 32k windows are omitted from this slope because of compressor saturation. The fixed corpus sample contains 10,000 documents; the numbers of windows at the eight lengths are 300, 150, 100, 75, 60, 50, 42, 37.

Calibration. The prediction rule is

$$\hat{a}_D = 1.182\hat{a}_{LZ} - 0.046. \quad (22)$$

The compressor, length grid, and linear map were selected on the calibration panel of 38 corpora. The compression slopes on that panel range from 0.119 to 0.454, with median 0.180. Applying the frozen rule to a new corpus requires the compression measurement above. [Section S.1](#) records the estimator comparisons; [Section J.7](#) records the subsequent transfer test.

Companion floor estimate. A separate sample of 100,000 documents gives a three-point extrapolation from rates at $N = 128, 2048, 32768$, with 2000, 1000, 500 windows, respectively:

$$\hat{h} = \max\left\{0, \frac{C_{128}C_{32768} - C_{2048}^2}{C_{128} + C_{32768} - 2C_{2048}}\right\}. \quad (23)$$

This expression follows from the finite-range template

$$C(N) = H_0 + A_H N^{-\alpha_s}. \quad (24)$$

The rate slope $\hat{a}_{LZ}$ is computed independently of this extrapolation, so the template exponent α_s and the total-rate slope measure different features of the curve. Exact preprocessing and the near-zero-denominator fallback are specified in [Section P](#). The two empirical outputs are $Q_{\text{DMO}}(X; \mathcal{I}) = (\hat{h}(X; \mathcal{I}), \hat{a}_{LZ}(X; \mathcal{I}))$, where $\mathcal{I}$ denotes the sampled start indices.

J.2 Case-1 Training Grid: Batch Size, Learning Rate, and the Batch-Tier Audit

Case 1 fits data-scaling laws on 51 corpora to test whether compression ranks the fitted exponents, which requires a consistent training grid: the same (P, D) factorial per corpus, with specified batch schedules and learning-rate rules. These choices reduce scheduled differences across cells but do not by themselves exclude optimization interactions. This subsection documents that grid.

The panel is a full 6×5 factorial in (P, D) per corpus, with GPT-2-style models at $d_{\text{model}} \in \{256, 512, 768, 1024, 1536, 2048\}$ ($n_{\text{head}} = d_{\text{model}}/64$, 4–22 layers, 16M–1B parameters) and token budgets $D \in \{12, 40, 130, 250, 700\}$ M at sequence length 512, with longer floor runs at $D \in \{2, 5, 10, 20\}$ B.

Batch size. Tokens per optimizer step follow the critical-batch law of [Zhang et al. \(2025\)](#), whose central finding is that the critical batch size scales with the *data* budget and is largely independent of model size: we set $B_{\text{tok}}^* = 22.91 D^{0.47}$, apply the paper’s $0.8 \times$ safety buffer, and floor to a power of two in sequences (minimum 16). This P -independence lets each fixed D share one batch across all six sizes, so the loss-versus- P comparison of [Section J.5](#) does not mechanically vary the batch schedule with P .

Learning rate. A width rule, $\text{lr} = 0.0015 \times 512/d_{\text{model}}$, with a two-point sweep at $\{1.0, 0.6\} \times$ the anchor and the better validation loss kept per cell. For $D \geq 10$ B, the rule is capped at the anchor. Without this cap, the width rule sets $\text{lr} = 0.003$ for the 256-wide model. This setting is stable at short budgets but diverged in 20/21 cells at $D = 10$ B and 7/15 at $D = 20$ B. Wider models remained stable at $\text{lr} \leq 0.0015$.

J.3 Case-1 Corpus Sets, Filters, and Confidence Intervals

The slope correlation $\rho = 0.84$ uses the 38 corpora retained after the language and fit-quality filters. The floor correlation $\rho = 0.75$ uses the 26 corpora whose long runs consumed at least 11B tokens. The fixed-model-size analysis in [Section J.4](#) uses 29 language corpora. These analyses have different eligibility criteria.

Targets. The observational targets are literature exponents and entropies for C4 ([Krajewski et al., 2025](#)), FineWeb ([Volkova et al., 2026](#)), and the loss-to-loss corpora of [Brandfonbrener et al. \(2024a\)](#). The experimental targets are data-scaling exponents α_D that we fit across 51 Web, Code, and Math corpora. Of these, 26 also provide long-run floor fits: their $D = 20P$ runs train for at least 11B tokens rather than merely requesting that budget.

Corpus-set sensitivity. Removing protein/DNA, LM-generated synthetic data, and instruction/SFT sets leaves 41 language corpora. Requiring at least 16 grid points and joint-fit $R^2 \geq 0.85$ then removes the_stack, Wikipedia RU, and Wikipedia AR, leaving 38. The slope association is $\rho = 0.65$ on all 51 corpora and $\rho = 0.84$, $R^2 = 0.60$ on the 41 language corpora; the fit-quality filter changes the strength but not the conclusion. The four literature points agree in sign ($\rho = 0.95$) but are nearly indistinguishable and therefore uninformative for ranking.

The floor: published references. Against published entropies, $\rho = 0.88$ and $R^2 = 0.88$ (bootstrap 95% CI [0.56, 0.97], $n = 19$), now spanning English, code, Chinese, Japanese, Korean, Russian, French, and German literature values. The Chinese, Japanese, Korean, French, and Russian entries all come from one table and one method: context-aware PPM entropy rates extrapolated to infinite length on same-language newspaper and literary corpora ([Takahira et al., 2016](#), Table 1, stretched-exponential fit), following the same-language-proxy convention already used for the English rows above (e.g. [Gao et al., 2020](#)’s Pile-CC value is reused for FineWeb, C4, and The Pile Deduplicated); German uses an older, pre-compression letter- prediction estimate ([Küpfmüller, 1954](#)). Wikipedia ZH is the highest point on the $\hat{h}$ axis ($\hat{h} = 8.9$, literature 3.1 bits/char, the mean of the four non-romanized Chinese corpora in [Takahira et al.](#)’s Table 1). It sits on the fitted line rather than off it: dropping it lowers the fit to $\rho = 0.86$ and $R^2 = 0.86$ ($n = 18$). Earlier versions of this panel scored Chinese with a 0th-order unigram character entropy (9.6 bits/char), which measures the size of the Hanzi inventory rather than the conditional predictability the other rows report; the wider bootstrap interval reflects the shorter reference range that follows from the corrected value. The Russian row likewise replaces two single-pass compression rates on a Bible translation (1.92 and 2.00 bits/char, [Behr et al., 2003](#)) with the extrapolated rate from the same table as every other non-Latin row. Neither substitution moves the rank statistic: Spearman is 0.88 before and after both. The observational floor comparison remains a ranking result throughout.

The floor: neural fits. Against our own fits we use the 26 corpora whose $D = 20P$ run consumed at least 11B tokens; the other 25 request $D = 20B$ but exhaust their tokenized pool at $\approx 0.9B$, and a run $23\times$ short of budget is not a converged floor. This budget criterion excludes severely truncated runs; a separate loss-tail convergence criterion is not documented. On those 26, $\rho = 0.75$ in nats/token (bootstrap CI [0.50, 0.90], permutation $p = 1 \times 10^{-5}$, $Q^2 = 0.41$).

Unit diagnostics. Character/token conversion remains associated with the floor residual, but no comparable association is detected for the slope residual. On the 26 long-run corpora, converting both axes to bits per character introduces the shared denominator chars_per_token and lowers the floor rank correlation from 0.75 to 0.61. The pooled bits/character fit reaches $R^2 = 0.94$ only with multiscript leverage extending to $\hat{h} = 8.9$, and its floors come from capped runs.

The unit effect is not removed completely by returning to nats per token. After regressing the raw floor on $\hat{h}$, the residual still correlates with chars_per_token (Spearman $\rho = 0.40$, Pearson 0.45, permutation $p = 0.022$). Adding $\log(\text{chars_per_token})$ raises R^2 from 0.483

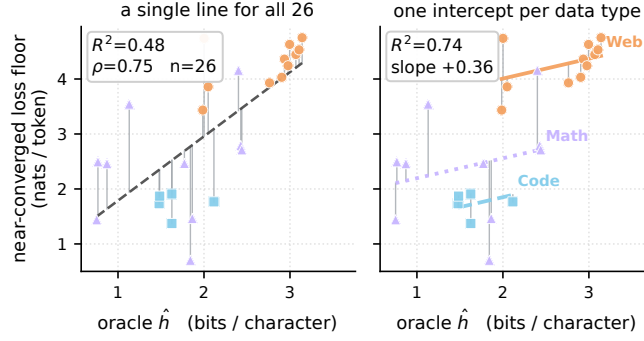


Figure 5: **The floor panel’s unexplained variance is the offset between data types, not noise around a single relationship.** Both sides show the same 26 corpora and the same $\hat{h}$ axis as Panel 4 of Figure 2; grey segments are residuals, whose squared lengths are what R^2 summarizes. Left: one intercept for every corpus, $R^2 = 0.483$. Right: one intercept per data type with a single shared slope of $+0.362$ on $\hat{h}$, $R^2 = 0.742$, with fitted lines drawn over each type’s own $\hat{h}$ range. The residuals that shrink are between-type, so most of the gain is the ordering of the three levels (3.283, 1.834 and 1.126 nats/token for web, math and code). Residuals within a type shrink much less, matching the weak within-type associations reported in Section J.3. The rank statistic quoted in the panel, $\rho = 0.752$, is computed on the pooled points and is unaffected by this refit.

to 0.691 (bootstrap 95% CI [0.53, 0.86]; permutation $p = 0.0006$ against a shuffled covariate). The headline remains the unadjusted rank statistic $\rho = 0.75$.

Why the pooled R^2 is low: data type, not training depth. A single line through all 26 corpora is a misspecification, because the three data types sit at different loss levels. Fitting one shared slope with a per-type intercept raises R^2 from 0.483 to 0.742, with slope $+0.362$ and intercepts 3.283 (web), 1.834 (math) and 1.126 (code) nats/token (Figure 5). What the shared slope buys beyond the intercepts is modest: within web the association is $\rho = +0.64$ ($R^2 = 0.250$, $n = 11$), within math $+0.41$ (0.055, $n = 10$), and within code -0.10 (0.000, $n = 5$), where $\hat{h}$ spans only 27% of its pooled range. So $\hat{h}$ separates the loss levels of different data types and orders web corpora moderately, and it does not resolve differences inside code or math at this sample size.

Training depth is not the limiting factor. The published floor is a minimum over each corpus’s own converged cells, so corpora are read at different model sizes and token budgets. Reading every floor instead at one common (P, D) cell removes that inhomogeneity, and bounds what depth can contribute: the *shallowest* matched cell, 16M parameters at 2B tokens, gives the highest matched R^2 at 0.568, while the deepest matched cells, at 19B tokens, give 0.477 to 0.491. Under a training-depth explanation the ordering would run the other way. Holding the corpus set fixed at the corpora with a complete model ladder and reading the floor off each size in turn is equally flat, moving R^2 by -0.014 across a $62\times$ range in parameters.

One corpus is a documented exception. AutoMathText (py-code) trains to the largest negative residual on this panel, and it is also the most repetitive corpus of the seven probed for duplication: its distinct word-8-gram ratio is 0.7255, below every other corpus measured and below both controls chosen to separate the plausible confounds, namely another LLM-synthesized corpus at 0.9845 and another pure-code corpus at 0.9084. That measurement is independent of the residual. Setting it aside leaves $\rho = 0.768$ and $R^2 = 0.525$ on the remaining 25, and $R^2 = 0.804$ with per-type intercepts. The panel and the headline retain all 26 corpora.

For $\hat{\alpha}_{LZ}$ on the same 26 corpora, the analogous residual correlation is Pearson $\rho = -0.33$ with permutation $p = 0.17$. A log-log slope is algebraically less sensitive than a level to rescaling the token axis. Unit conversion can therefore contribute to the difference between the floor and slope results, but the present test does not identify it as the mechanism.

Distinct information in the compression statistics. They correlate at $\rho = -0.82$, but on the same 41 corpora $\hat{h}$ ranks α_D at only $\rho = -0.54$, compared with 0.84 for $\hat{\alpha}_{LZ}$. The single-point alternatives are weaker still: -0.38 for the corpus-growth exponent, the closest analogue to a raw gzip ratio, and $+0.22$ for the three-point closed form of [Section P.4](#). Conversely, $\hat{\alpha}_{LZ}$ retains partial $\rho = 0.74$ after controlling for $\hat{h}$. The two measurements are related, but the floor readout is a substantially weaker predictor of how fast a corpus scales.

Headline exponent association. On the selected 38-corpus set, $\rho = 0.84$ and $R^2 = 0.64$, with Fisher and bootstrap 95% CIs $[0.72, 0.92]$ and $[0.71, 0.92]$, permutation $p < 10^{-4}$, and leave-one-out $Q^2 = 0.59$. Both the corpus set and the eight-point trimmed log-log zlib estimator were chosen on this panel. Six of the ten corpus exclusions were introduced when 18 corpora were added, and $\hat{\alpha}_{LZ}$ was one of six readouts compared in [Section S.1](#).

The full estimator/corpus-set comparison is therefore more informative than its best cell ([Table 9](#)). All four LZ-slope estimators rank α_D at $\rho \in [0.63, 0.84]$ with $Q^2 \in [0.13, 0.59]$, and agree with one another at $\rho \geq 0.93$ (≥ 0.81 after controlling for $\hat{h}$). In each of the 12 cells, none of 20,000 label permutations matches the observed statistic; these are unadjusted cell-wise counts, and no selection-adjusted p -value is available for the chosen estimator. The three-point readout and the sharper predictive compressors (order-16 PPM and LZ 100k) do not show a comparable corpus-set-wide association. [Proposition B.8](#) allows such finite-scale estimands to differ but does not predict their empirical ordering.

Leave-one-parent-family-out evaluation. Leaving out each of the panel’s 27 parent families gives the recorded results below. Estimator selection used the same corpus panel, so this analysis measures internal robustness. The slope predictor’s MAE is close to the domain mean’s, while its rank correlation is higher.

Predictor	MAE	Q^2	Spearman ρ
Compression slope	0.0514	0.584	0.795
Intercept	0.0785	-0.057	—
Domain mean	0.0508	0.317	0.535

Later measurement on known corpora. After the estimator was fixed, a 270-cell remeasurement of the nine widest- $\hat{\alpha}_{LZ}$ corpora on one Chinchilla grid gives $\rho = 0.87$, $R^2 = 0.90$, and exact permutation $p = 0.0023$ over all 9! relabelings. Applying the earlier line to these newly measured exponents gives $R^2 = 0.56$, or 0.47 under a matched $D \leq 700M$ ceiling. Because the corpora already appeared in the original panel, this is a later outcome measurement rather than a new-corpus test.

Consistent-training sensitivity. Replacing the capped-pool exponents for those nine corpora puts all 41 language corpora on one training protocol. The rank statistic is unchanged: $\rho = 0.839$ versus 0.838, with bootstrap CIs $[0.69, 0.92]$ versus $[0.71, 0.91]$ on the same corpora and $\hat{\alpha}_{LZ}$ span $[0.12, 0.45]$. The apparent R^2 increase from 0.60 to 0.81 is not resolved: the paired change is $+0.20$ with 95% CI $[-0.12, +0.38]$, and removing the two highest- $\hat{\alpha}_{LZ}$ points reduces both fits to $R^2 \approx 0.50$ at $\rho = 0.813$. The evidence supports stable ranking under this protocol change, not a gain in explained variance.

Dependence on prior outcomes. The superseded and remeasured exponents correlate at $\rho = 0.65$. After controlling for the old values, the $\hat{\alpha}_{LZ}$ association is partial $\rho = 0.76$ with exact $p = 0.015$, compared with 0.22 and $p = 0.30$ for order-16 PPM on the same nine corpora. The partial correlation after controlling for $\hat{h}$ is 0.74, rising to 0.80 on the English Web/Code/Math subset. These small, overlapping comparisons support the estimator ordering but do not replace the headline panel or a prospective corpus test.

J.4 The Exponent Link Across the Tested Model-Scale Range

No weakening is detected from 16M to 1B parameters. We fit $L(D) = E_P + A_P D^{-\alpha_{D,P}}$ separately at each model size, excluding the $D = 20B$ cells used for the floor proxy, and correlate

$\alpha_{D,P}$ with $\hat{\alpha}_{LZ}$ over the 29 language corpora. For the selected zlib estimator, Spearman correlations at $P = 16\text{M}/51\text{M}/124\text{M}/253\text{M}/512\text{M}/1\text{B}$ are 0.88/0.91/0.92/0.96/0.93/0.89. The fitted trend is +0.005 per model size across this $64\times$ range; the three largest sizes average 0.93, compared with 0.90 for the three smallest. Order-16 PPM gives lower correlations, 0.34–0.46, but also a nonnegative trend (+0.016 per model size; [Section S.1](#)). Neither result extrapolates beyond 1B.

Each fit must estimate E_P freely. Reusing the largest model’s floor forces its residual toward zero and creates an artificial trend of -0.111 per model size ($0.75 \rightarrow 0.20$). After controlling for the correlated floor readout (Pearson -0.70 between the two measurements), $\hat{\alpha}_{LZ}$ still ranks $\alpha_{D,P}$ at partial Pearson 0.73–0.85 at every size; the reverse partial correlations are -0.23 to -0.43 . No fit reaches its bound, median per-cell $R^2 \approx 0.97$, using only the five shared data budgets gives Spearman 0.95–0.96, and leave-one-out Spearman spans 0.88–0.96.

J.5 The Model Axis, Measured: Ranking with Data-Axis Overlap

Every Case-1 result so far is on the data axis (α_D). The theory distinguishes that from the model axis (α_{head}): the data axis has no geometric target, while the head bound is conditional. Transposing the factorial—fitting $L(P) = E_D + A_D P^{-\alpha_P}$ across six model sizes at each fixed D —shows a weaker positive association with $\hat{\alpha}_{LZ}$ ($\rho \approx 0.75$ rather than 0.84), with substantial overlap with the data-axis result.

Eligible data budgets. A model exponent is not meaningful where the sweep is not parameter-limited. We therefore require at most half the corpora to have loss non-monotone in P . The $D \leq 40\text{M}$ budgets fail because 98% and 84% of corpora are data-starved; at $D = 10\text{B}$ the largest model regresses for 57% of corpora because of the batch-tier artifact in [Section J.2](#), rather than divergence (dropping P_5 restores $\rho = +0.46$).

Result and data-axis overlap. Across the five eligible budgets, $\hat{\alpha}_{LZ}$ ranks α_P at $\rho = +0.53$ to $+0.79$. At the least noisy budget, $D = 5\text{B}$, the correlation is $+0.75$ ($n = 26$, exact permutation $p = 5 \times 10^{-5}$), below the data-axis correlation 0.84. The two fitted exponents correlate at $\rho = +0.68$; controlling for α_D reduces the model-axis association to partial $\rho = +0.31$, showing substantial overlap without assigning a fraction of signal to either axis.

The result is compatible with, but not predicted by, the one-sided achievability bound, which does not order corpora. Descriptively, $\alpha_P \geq 2\hat{\alpha}_{LZ}$ for all 24 corpora with median slack $+0.227$, and $2/\alpha_P \in [3, 100]$ for 18 of 24. This is not a test of the theorem because $\hat{\alpha}_{LZ}$ is not a dimension estimate.

Fit controls. The joint Chinchilla fit’s α_P column cannot be used because its bound at 1.5 binds for 9 of 38 corpora. We refit each budget with bound 4.0; none reaches it at the headline budget. The leave-one- P -out SD is 0.17 of the between-corpus SD there and at most 0.43 at any eligible budget, compared with 0.93 where the per-domain measurement becomes noise ([Section S](#)). Pairwise rank agreement across eligible budgets is $+0.40$ to $+0.91$, with median exponent ratios near one.

J.6 Prospective Case-1 Pilot

The exponent pilot improves on the intercept, whereas the floor mapping does not. For FineWeb-2 ZH, FineWeb-2 AR, and Lucie, the fixed line $\alpha_D^{\text{pred}} = 1.182\hat{\alpha}_{LZ} - 0.046$ gives predicted/observed exponents 0.089/0.099, 0.156/0.252, and 0.245/0.361: MAE 0.074 versus 0.111 for the intercept, or $1.33\times$ the calibration residual RMS. The readouts were computed before the neural outcomes existed and the line was not refit. With only two parent families and no ex-ante decision threshold, this is a prospective calibration diagnostic, not the family-generalization test; the points remain outside [Figure 2](#) and the calibration panel.

The two non-Latin floor points lie outside the published floor range. Their fixed mapping has MAE 5.194 nats/token versus 0.684 for the intercept and 0/2 family-block interval coverage; the gap persists across pooling, refitting, and unit changes. The scope of the empirical cross-script floor mapping is limited accordingly; the entropy-rate result and the certificate in [Theorem B.1](#) and [Proposition B.3](#) do not depend on it.

J.7 Six-Family Blind Transfer and Measurement Checks

On the five families with valid scaling fits, the fixed map has MAE 0.02891, compared with 0.04821 for the intercept and 0.04490 for the domain baseline (ratios 0.600 and 0.644). Spearman correlation is 0.800, and all five residuals fall within the fixed interval half-width 0.103. The result passes all five criteria of the frozen transfer rule in Table 6. With five evaluable families, the exact two-sided permutation $p = 0.133$ and $Q^2 = 0.325$ are descriptive statistics, as specified in that protocol; Table 2 reports the five eligible residuals at recorded precision.

The test comprised BigPatent, Hacker News, FLAN, YouTube Commons, Amazon Books, and FinePDFs English, each trained on a 6×5 (P, D) factorial, for 180 cells. Before any neural outcome existed, the protocol fixed the family set, the rounded map $\hat{\alpha}_D = 1.182\hat{\alpha}_{LZ} - 0.046$, the requirement of at least 16 cells and joint-fit $R^2 \geq 0.85$, both baselines, the family-block prediction half-width of 0.103, and five success criteria. The map and 38-corpus calibration artifact were not refit.

The evaluation rule is recorded in pre-train/case1_blind_eval_v1/preregistration.json (lines 72–103). The corrected execution protocol, pre-train/case1_blind_family_v2/preregistration.json, was frozen at 2026-08-13T21:20:18Z and retained the same evaluator and predictions. Its freeze manifest pins the evaluation protocol to SHA-256 d2e79c7fd57bcb7c5665110900c11ae306e207414271f16a0052b4b1ca38dc3a. The prediction lock records the roster and domain assignments before training.

All 180/180 cells passed the technical checks and all six outcomes were revealed once. An earlier attempt was quarantined before reveal because its seed/resume semantics failed the integrity check; no scientific result from that attempt was used. The corrected run preserved every scientific choice, restarted each attempt from step zero, and used atomic no-overwrite promotion. The exponent metrics therefore describe the five eligible families from this six-family test.

The excluded family, re-measured. On a denser grid at double the data budget, the family the gate excluded agrees with the frozen prediction. FinePDFs English originally fit $\alpha_D = 0.562$ at $R^2 = 0.850$, against a predicted 0.160 and a fixed interval $[0.058, 0.263]$. That fit carried the two marks of a misfit: the lowest R^2 of the six families, which is what the gate caught, and $\alpha_P = 1.44$ against the fitter’s bound of 1.5, where the other five sit between 0.22 and 0.52. We rebuilt its tokenized pool to 1.6B tokens with the original bytes as an exact hash-checked prefix, added a $D = 80M$ rung to resolve the knee between 40M and 130M, added a 1.4B rung, and retrained the whole 6×7 grid on one common validation clock. Every fit window lands inside the frozen interval: 0.092 on the original five rungs read on the new clock, 0.099 on all seven, 0.132 for $D \geq 130M$, 0.174 for $D \geq 80M$, 0.195 for $D \geq 40M$, and 0.208 for $D \geq 80M$ with the 1.4B rung dropped. Replacing the anchor cell with two further training seeds moves the exponent by 0.0003. No window returns more than 0.208, so 0.562 does not reproduce.

This is a post-reveal measurement. It does not change the frozen decision, which rests on the five eligible families, and the six-family metrics above are unchanged. Two of the 52 runs, the seed replicates of the largest cell, were lost to preemption and cannot be regenerated, because the protocol pins the training harness by content hash and that content was never committed. Under the classification rule declared for this diagnosis the result is MIXED rather than PRE-ASYMPTOTIC: the shift from the original-rung fit to the asymptotic one is 0.083 against a declared threshold of 0.10. The record is pre-train/reports/case1_finepdf_partial_readout.md.

What the fit-quality gate should check besides R^2 . With five data budgets the joint fit does not pin α_D independently of the window. Dropping each family’s smallest budget, leaving four, raises every fitted exponent by two to four times while *raising* R^2 : BigPatent goes from 0.239 at $R^2 = 0.93$ to 0.588 at 0.98, Hacker News from 0.139 to 0.510, Amazon Books from 0.111 to 0.494. One window is applied to every corpus, so the rankings are unaffected

and the exponents are protocol-relative; the point is that a higher R^2 is not evidence of a better-determined exponent. Two of the six fits also put a parameter at a bound, FLAN at $E = 10^{-3}$ against a minimum observed loss of 1.35 and FinePDFs at $\alpha_P = 1.44$ against 1.5, and FLAN carries the largest residual among the eligible families (-0.063). A boundary check belongs beside $R^2 \geq 0.85$ in the gate.

Table 6: Frozen prospective protocol and observed outcome. All five decision criteria passed.

Item	Recorded specification
Prediction	$\hat{\alpha}_D = 1.182\hat{\alpha}_{LZ} - 0.046$, fixed before outcomes
Families and grid	Six named families; 6×5 cells per family; 180 cells total
Fit eligibility	At least 16 cells and joint-fit $R^2 \geq 0.85$; one family excluded
Baselines	Global calibration mean 0.182787; calibration domain mean with global-mean fallback
Prediction interval	Half-width 0.103: higher empirical 95th percentile of the maximum absolute residual per held-out parent family
Decision rule	At least five evaluable families; $MAE \leq 0.0785$; MAE ratios to both baselines ≤ 1 ; $\rho \geq 0.7$; interval coverage ≥ 0.8
Reported outcome	Five eligible families; MAE 0.02891; $\rho = 0.800$; interval coverage 5/5

The fixed domain means, computed from the frozen 38-row calibration artifact, are 0.093573 for web/prose, 0.228049 for math, 0.247706 for code, 0.162158 for reference/science, and 0.139885 for other. The locked roster assigns Amazon Books, Hacker News, and YouTube Commons to web/prose; FinePDFs to reference/science; and BigPatent and FLAN to other. The interval uses leave-one-parent-family-out calibration fits, with one maximum absolute residual per family, and is descriptive rather than a calibrated coverage guarantee. The thresholds concern prediction error, direction, and coverage; the exact rank p -value is not a separate pass criterion.

Measurement repeatability. Resampling barely changes the family ordering. Before reveal, the DMO-Corpus statistic was recomputed on the same six source corpora with window seeds 0–19. Across families, the propagated $\hat{\alpha}_D$ SD is 0.0010–0.0067; seed zero reproduces every reported value exactly. The rank correlation with the reported six-family ordering has mean 0.997 and descriptive 2.5–97.5% range [0.970, 1.000]. A second sensitivity draws 100 contiguous 100-document blocks with replacement in each of 20 replicates; its mean rank correlation is 0.986, with range [0.883, 1.000]. These are repeatability diagnostics for the first 10,000 valid documents, not population confidence intervals.

Sequence-structure control. Shuffling word order lowers $\hat{\alpha}_{LZ}$ in all six families while preserving their rank order. The control first reproduced the reported DMO-Corpus scores, then shuffled maximal non-whitespace tokens within each document while preserving their multiset, whitespace slots, document character length, sampling grid, and seed. The resulting $\hat{\alpha}_{LZ}$ was lower in all six families: the median original-minus-shuffled difference was 0.02574, with exact two-sided sign $p = 0.03125$. As an implementation-sensitivity baseline, XOR relabeling preserved byte length and equality structure and changed $\hat{\alpha}_{LZ}$ by median absolute 0.00341. The sequence contrast is therefore larger than this byte-label baseline, but the original and shuffled family orderings remain identical (Spearman $\rho = 1$). The control shows that the finite-grid statistic responds to sequential organization, but it does not isolate a dependency scale or explain the cross-family neural ordering.

Same-tokenizer, same-clock floor diagnostic. The floor gap persists with tokenizer and clock matched. Among the five families with valid fits, validation-tail floor versus neural E has $\rho = 0.30$ (exact two-sided $p = 0.683$ over all $5! = 120$ permutations), Pearson $r = 0.159$, in-sample $R^2 = 0.025$, and leave-family-out MAE 2.064 bits/token versus 1.414 for the intercept. The reported Q^2 is -0.968 .

The protocol was fixed before reveal to test whether the earlier floor gap was merely a unit or tokenizer mismatch. It reads the exact GPT-NeoX unsigned-16-bit token streams used for training, measures raw-DEFLATE rates at 128, 2,048, and 32,768 tokens, and applies

the same three-point extrapolation. Its primary analysis uses the fixed 5% validation tail; the full bin is a sensitivity analysis. After reveal, neural E was refit in bits per GPT-NeoX token from each family’s same 30 primary cells. The full-bin sensitivity likewise detects no rank association and performs worse than the intercept ($\rho = -0.10$, exact $p = 0.950$, $Q^2 = -0.589$). The gate-failing family is excluded from both associations because its fit failed the primary criterion.

J.8 An Equal-Target Training Proxy on the Same Five Families

A proxy that trains on the candidate corpus predicts the level worse and the ordering no worse. The comparison uses cells that were already trained for the six-family test, so it adds no compute: for each family we take the 16M-parameter row of its own 6×5 grid, fit $L = E + AD^{-\alpha}$ on the cheapest budgets alone, and use α as the prediction. The target, the five families, and the pooling are the ones the frozen rule already fixed.

Predictor	MAE	Spearman ρ	exact one-sided p	tokens trained per corpus
Compression slope $\hat{\alpha}_{LZ}$	0.0289	0.800	0.067	none
Fixed intercept	0.0482	—	—	none
Domain mean	0.0449	0.866	0.100	none
16M model, 5 budgets	0.0836	0.900	0.042	1.13B
16M model, 3 budgets	0.0851	1.000	0.008	0.18B
16M model, 2-point slope	0.1151	0.900	0.042	0.05B

The proxy’s error is bias, not noise. Its predictions are 0.036, 0.057, 0.072, 0.098 and 0.106 against observed exponents of 0.111 to 0.239: a single model size reads the joint exponent two to three times too low on every family, because the joint fit pools six sizes and the one-size fit cannot separate E from A over five budgets. Rescaling it needs a calibration panel, which is the cost the raw-text map already paid once.

Ranking is where the proxy holds up. The three-budget proxy orders the five families exactly and the five-budget proxy misses by one adjacent swap, against $\rho = 0.800$ for $\hat{\alpha}_{LZ}$; at $n = 5$ one swap is worth 0.1 in ρ , so these orderings are not separable. This matches [Magnusson et al. \(2025\)](#), who find that ranking data recipes from single-scale runs is about as compute-efficient as extrapolating a scaling law. The two claims that survive the comparison are the level and the cost: 0.0289 against 0.0836 on the same target, at zero training tokens against 1.13B per candidate.

The three training-free rows reproduce the frozen protocol’s recorded MAEs (0.02891, 0.04821, 0.04490) from full-precision $\hat{\alpha}_{LZ}$; the script is `pre-train/scripts/case1_blind_cheap_proxy_baseline.py` and the figure is generated by `hurst/make_main_transfer.py`.

J.9 A Within-Corpus Intervention: Multi-Scale Block Shuffle

This experiment tests whether moving sequential structure within a fixed corpus changes both $\hat{\alpha}_{LZ}$ and the fitted exponent. It does on all three tested corpora. The design removes between-corpus variation but changes several order-dependent features at once, so it does not identify a unique mechanism. For a corpus X , the intervention $T_\ell(X)$ cuts each document into ℓ -character blocks and permutes the blocks within the document. Document length, the block multiset, and every block’s internal character string are unchanged; block order and cross-block context change, and retokenization can change token frequencies. We apply T_ℓ at $\ell \in \{64, 256, 1024, 4096\}$ plus an unshuffled control to three corpora spanning the exponent range and the three Case-1 domains — StarCoder (code), FineWeb-Edu (web), and Proof Pile 2 arXiv (math). For each arm, we rerun the DMO-Corpus readout, retrain a fifteen-cell (three model sizes $\times$ five data budgets) grid per arm from scratch, and refit $\alpha_D(\ell)$.

Arm ordering. Across all three corpora, the fitted exponent rises with ℓ (0.032 at $\ell=64$ to 0.055 at the control on web; 0.085 to 0.130 on code; 0.053 to 0.098 on math), and $\hat{\alpha}_{LZ}$ rises with it under every estimator. Per corpus the Spearman correlation over the five arms is 1.00 on web and math and 0.70 on code under the selected trimmed-zlib slope (0.90/0.70/1.00

under PPM, 0.90/0.60/0.90 under LZ3); code’s $\ell=4096$ arm is near its control. The function `block_shuffle_text` in `pre-train/block_shuffle_exp.py` (lines 76–81) leaves documents of length at most 2ℓ unchanged. Longer documents are split into ℓ -character blocks, including any shorter final block, and all blocks are permuted together. Only 38.8k of the 530k code documents exceed 8k characters, so the 4096-character arm changes a small subset.

Stratified inference. Five-arm rank tests have coarse resolution. To combine the three within-corpus associations, we use a stratified exact permutation: relabel arms within each corpus $((5!)^3 = 1,728,000$ relabelings, statistic the mean per-corpus ρ). This reference distribution assumes exchangeable arm labels under a no-association null and does not quantify training-seed or new-corpus uncertainty. The analysis script `pre-train/scripts/blockshuffle_pooled.py` (lines 59–74) enumerates the upper tail $\Pr_{\text{perm}}(\bar{\rho}_{\text{perm}} \geq \bar{\rho}_{\text{obs}})$ and reports a one-sided probability. Reconstruction from the listed untied five-arm correlations yields 224, 560, and 2435 qualifying relabelings out of 1,728,000 for trimmed zlib, PPM, and LZ3, respectively. These counts reproduce the recorded probabilities: $p = 0.00013$ (trimmed zlib), 0.00032 (PPM), and 0.00141 (LZ3). The absolute two-sided probabilities are twice these values. The script does not record when the tail was chosen. A common slope with corpus fixed effects (both variables centred within corpus, fifteen arms, eleven degrees of freedom) is significant for all three estimators ($t = 6.5\text{--}7.4$, within- R^2 0.80–0.83); the per-corpus slopes agree within 29% across code, web, and math for PPM and LZ3 but span $2.2\times$ for trimmed zlib. Thus, within these three corpora, every tested estimator tracks the arm ordering, while the across-corpus magnitudes are more similar for the two predictive-match estimators. This is consistent with the estimator sweep of [Section S.1](#).

Floor-side behaviour. The floor estimate is the intervention’s control on code and web — $\hat{h}$ is flat across their arms — but not on math, where displacing the long-structured text changes the compression curve’s curvature enough to move the three-point floor extrapolation ($\hat{h}_{\text{LZ}}$ falls from 1.83 to 1.30 at $\ell=64$). The exponent result does not rest on floor flatness, and the floor theory rests on [Theorem B.1](#) rather than on this invariance, but the clean two-coordinate separation is a two-of-three-corpora observation. The five arms are five points of one ordered intervention, not independent draws; the observed mean rank association is unusual under independent random relabeling of the five arms within each corpus.

Lag decomposition. Block shuffling preserves block contents and changes their relative positions. The observed compression gains shift toward larger windows; this does not establish preservation of every original dependency. Writing $G_{\text{comp}}(N) = C_N - C_{2N}$ for the marginal gain (bits/char) from doubling the compression window at N , the association with α_D across the 15 arms, with corpus fixed effects, changes sign with N : positive at short lags ($t = +7.3$ at $N=64$, $+7.7$ at $N=128$), insignificant at $N=512\text{--}1024$, and negative from $N=2048$ on ($t = -5.4$ to -6.8). The cumulative long-window gain $G_{\text{comp},\geq}(N) = \sum_{M\geq N} G_{\text{comp}}(M)$ is negatively signed at every $N \geq 256$, peaking at $N=512$ ($t = -10.5$, within- $R^2 = 0.91$). The block multiset is unchanged, while original cross-block adjacency is disrupted. The measured curves show greater compression gains at long windows after shuffling. On code, for example, $G_{\text{comp}}(16,384)$ rises from 0.024 in the control to 0.128 at $\ell = 64$. The pattern therefore implicates the location of predictable structure, not simply its total amount. The nine N values are correlated measurements from the same curve summarized by $\hat{\alpha}_{\text{LZ}}$, so this is a decomposition of the intervention result rather than independent evidence; the sign-change location is approximate.

Matched-slope control. On two StarCoder pairs, mixtures with the same $\hat{\alpha}_{\text{LZ}}$ as a uniform-shuffle arm also have fitted exponents closer to that arm than to a document-mixture prediction. To construct the pairs, we shuffle a Bernoulli- q fraction of documents at $\ell=64$ and leave the rest intact. The resulting bimodal spectrum has a trimmed slope determined by the endpoint arm curves. We tune $q=0.674$ to match the $\ell=256$ arm and $q=0.364$ to match $\ell=1024$. The realized slopes are 0.2192 versus the target 0.2222 and 0.2384 versus 0.2378; both training grids completed 15/15 cells. We compare two benchmark predictions: a scalar-only interpolation assigns each mixture its slope-matched arm’s exponent, while a simple document-mixture heuristic uses the corresponding mix of the endpoint exponents,

0.008–0.020 away. The observed values are $\alpha_D = 0.1101$ against 0.1075 slope-matched and 0.0996 mixture-predicted at $q=0.674$, and 0.1298 against 0.1333 and 0.1134 at $q=0.364$. The respective errors are 0.0026 versus 0.0105 and 0.0035 versus 0.0164, making the matched-slope predictions $4.1\times$ and $4.7\times$ closer, at the grid’s refit-noise scale and with mixed residual signs. The values and completion counts are recorded in `pre-train/reports/blockshuffle_matched_slope.md`. This two-pair, one-corpus check is consistent with the scalar readout but does not establish that it is generally sufficient. Across the observational corpus set, no single $G_{\text{comp}}(N)$ band out-ranks the aggregated slope.

Extension across model sizes. The control-minus-shuffled exponent gap remains positive through 512M parameters. We extend the maximum-contrast pair (control versus $\ell=64$) from 16M–124M to 253M and 512M parameters on both corpora and refit $L(D) = E_P + A_P D^{-\alpha_{D,P}}$ at each model size with a free floor. The reported quantity is the control-minus-shuffled exponent gap at each size.

model size	16M	51M	124M	253M	512M
gap, code	0.033	0.069	0.073	0.081	0.077
gap, web	0.018	0.023	0.024	0.028	0.032

The recorded gap is smallest at 16M, grows, and then plateaus on code while continuing to rise through 512M on web. One possible explanation is that larger models use order-dependent information that the shuffle disrupts. These fits do not isolate that mechanism or quantify training-seed variation. Across the measured $32\times$ size range, the observed endpoint gap persists alongside the rank stability in [Section J.4](#). The measured window ends at 512M; no decreasing trend is visible in these point estimates.

J.10 One Anchor Run Yields an Empirical Curve Extrapolator

Construction. The calibrated readouts predict the fitted floor E and exponent α_D , but not the amplitude A in $L(D) = E + AD^{-\alpha_D}$. One anchor run (D_0, L_0) supplies it: $\hat{L}(D) = \hat{E} + (L_0 - \hat{E})(D/D_0)^{-\hat{\alpha}_D}$. We take the $D_0 = 12\text{M}$ -token cell of each corpus/model-size row as the anchor and predict its larger-budget cells. The $\hat{\alpha}_{LZ} \rightarrow \alpha_D$ map and every calibration constant are fitted without the predicted corpus. Because $\hat{h} \geq L_0$ in 29% of rows, the method also calibrates the floor without the held-out corpus rather than using $\hat{h}$ directly.

Accuracy and range. Across all 882 predicted cells, median relative error is 14.0% and the 90th percentile is 26.2%, compared with 32.1% for carrying the anchor loss forward and 5.9% when the full curve is refitted. Error grows with distance: 3.7% at $2\text{--}5\times$ the anchor budget, 17.6% at $5\text{--}15\times$, and 66.8% beyond $30\times$. On this panel, the useful range is therefore roughly $10\times$ the anchor budget, where median errors are about 10–15%. Longer extrapolation still requires a direct sweep.

Filtering trades coverage for task accuracy. On the 11 Paloma subsets the $\tau = 10$ arm is 0.068 bits per byte worse than the random draw at 150M and 0.001 worse at 300M, and neither larger size has a bits-per-byte readout. What separates this sweep from the quadrant grid is the dose: splitting at both medians leaves each quadrant about a quarter of its pool, where this one keeps a tenth.

K Related Work

Predicting scaling from trained models. Conventional scaling laws are fit to several trained model sizes ([Hestness et al., 2017](#); [Kaplan et al., 2020](#); [Hoffmann et al., 2022](#)), with results that depend on the fitting protocol ([Krajewski et al., 2025](#); [Volkova et al., 2026](#)). Recent methods reduce the new sweep by transferring information from other trained models ([Ruan et al., 2024](#); [Brandfonbrener et al., 2024a](#); [Mayilvahanan et al., 2025](#)), while prescriptive scaling uses trained-model evaluations to estimate attainable capability boundaries ([Zhang et al., 2026](#)). [Cagnetta et al. \(2026\)](#) explain data scaling through the growth of a usable prediction horizon. When this mechanism dominates, conditional-entropy and token-correlation decays give $\alpha_D = \gamma_{\text{CE}}/(2\beta_{\text{corr}})$; their entropy exponent is estimated from converged short-context losses of trained models. DMO-Corpus supplies a calibrated forecast from raw text before a model is trained on the candidate corpus.

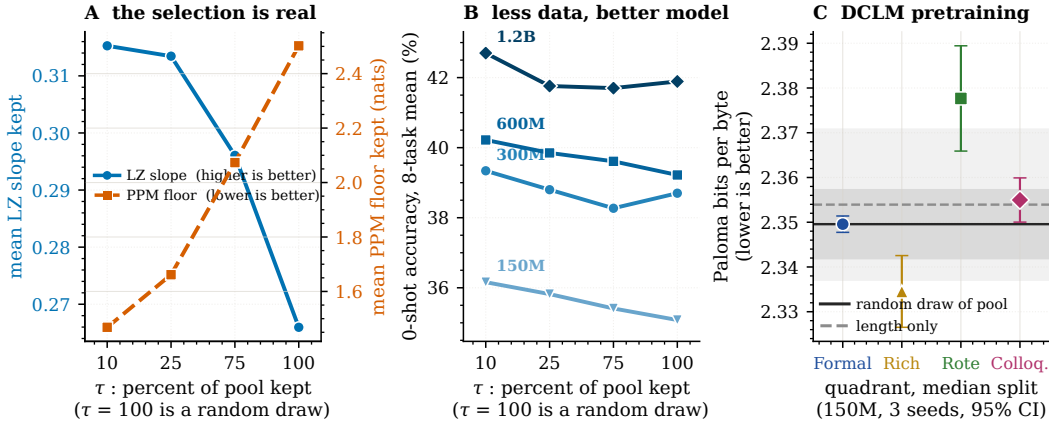


Figure 6: **Selecting the top $\tau\%$ of a pool by the two document scores trains a better model than not selecting at all, at four model sizes.** (A) The selection does what it claims. Mean LZ rate slope and PPM floor of the documents each arm actually trained on, read back from the selected-document records: the slope rises 0.65 standard deviations and the floor falls 1.03 as τ tightens, and the $\tau = 10$ and $\tau = 100$ arms share 0.1% of their documents. (B) Mean 0-shot accuracy over eight tasks. Every arm draws its 200,000 documents uniformly from its own top- τ pool, so $\tau = 100$ is a random draw of the whole pool at the same budget, not a ranked prefix of it; $\tau = 10$ beats that random draw at all four sizes, by 0.9 points on average, and accuracy falls as τ loosens at every size. Counting each task once, averaged over the four sizes, six of eight improve at 0-shot and all eight at 5-shot (sign test $p = 0.008$); the movers are SCIQ (+4.2 points) and ARC-Easy (+1.7), which between them supply 0.74 of the 0.88-point mean. Each cell is one run. (C) The same two scores used as a median split on DCLM, run as a separate six-arm grid at three seeds against a random draw of that pool. **RICH** trains 0.015 bits per byte better than the random draw ($t = 2.7$); a length-only filter blind to both scores does not (+0.004); **FORMAL** matches the draw to 0.000 and **ROTE** is 0.028 worse. A median split leaves each quadrant about a quarter of the pool where (B) keeps a tenth, so selection strength is what separates the two designs.

Raw corpus statistics. The closest precedent is Pandey (2024), who measure gzip compression ratios on fixed-length sequences generated by probabilistic grammars and fit scaling parameters as functions of that compression level. DMO measures how the compression rate changes with window length and tests the resulting fixed predictor on new natural-text families. Other routes connect corpus statistics to an exponent under additional assumptions: Dębowski (2025) chain Zipf, Heaps, and Hilberg laws to neural scaling at one literature value of the Hilberg exponent, and Han et al. (2026) transfer a fitted exponent across corpora through a gzip ratio. The distinctive test here is transfer of a frozen compression-to-exponent map, followed by interventions on sequence order within the same corpus.

Representation and geometry. Manifold models obtain $\alpha \propto 1/d$ from regularity assumptions on the learner and data geometry (Sharma & Kaplan, 2022); resolution-limited, random-feature, and quantization models derive exponents from spectra or skill frequencies (Bahri et al., 2024; Maloney et al., 2022; Michaud et al., 2023). Approximation rates on manifolds can depend on intrinsic rather than ambient dimension (Shaham et al., 2018; Chen et al., 2019; Nakada & Imaizumi, 2020; Schmidt-Hieber, 2020). Our direct-head result specializes this principle to softmax cross-entropy on a supplied, smooth predictive state with regular conditional logits (Proposition B.4).

Dimensionality reduction and finite-sample tests probe geometry in a chosen metric representation (Tenenbaum et al., 2000; Roweis & Saul, 2000; Belkin & Niyogi, 2003; Narayanan & Mitter, 2010; Fefferman et al., 2016). Prior studies use such estimates to predict image difficulty and representation compression (Pope et al., 2021; Ansuini et al., 2019), and to measure dimension in language representations (Du & Tanaka-Ishii, 2024). Our symbolic obstruction identifies the geometric information that a text-only measurement leaves undetermined, and our conformal bridge shows how an independent expansion rate resolves it.

Coding, memory, and recurrence. Probabilistic prediction and code length are operationally linked under the usual coding construction (Délétang et al., 2024; Huang et al., 2024). Compression quantities also constrain context-length scaling in the settings analyzed by Shi et al. (2025). Universal-coding and Brudno-type limits identify entropy rates (Brudno, 1983); Ornstein–Weiss and Wyner–Ziv connect LZ match lengths to recurrence (Ornstein & Weiss, 1993; Wyner & Ziv, 1989). Under dynamical

assumptions, quantitative recurrence recovers local dimension (Boshernitzan, 1993; Barreira & Sausol, 2001; Galatolo, 2007), and entropy–expansion formulas give $d_{\text{geo}} = h_X/\lambda$ in the conformal setting (Ledrappier & Young, 1985; Young, 1982). A finite- N Hilberg residual concerns memory and mutual-information decays (Hilberg, 1990; Dębowski, 2011; Ebeling & Pöschel, 1994; Lin & Tegmark, 2017).

Data selection. Current curation pipelines use classifiers, deduplication, importance resampling, or reference-model perplexity (Li et al., 2024; Xie et al., 2023b; Marion et al., 2023; Brandfonbrener et al., 2024b). The closest line to ours predicts good data by training first and reading the result: Liu et al. (2024) train 1M-parameter proxy models on Dirichlet-sampled mixtures for 1B tokens and regress loss on the mixture weights, and Magnusson et al. (2025) rank data recipes from small runs, finding that eight scaling-law extrapolation variants are no more compute-efficient than ranking single-scale experiments directly. Xie et al. (2023a) reweight domains from a small proxy pair, and Ye et al. (2024) fit loss as a function of mixture proportions so that a mixture can be chosen without training on it. All of these spend training compute on the candidate data first, and the amount can be small: on our own prospective families, 0.18B tokens on a 16M-parameter model orders them as well as the raw-text readout does, while predicting the exponent three times worse (Section J.8). DMO reads the candidate text itself, so it supplies a calibrated exponent at no training cost, and the cost of a decision does not grow with the number of candidates. Jiang et al. (2025) fit per-domain scaling laws during training and allocate samples using the expected marginal loss reduction: the scaling exponent times the remaining reducible loss, divided by the number of samples. DMO-Corpus provides its raw-text forecast before training; the anchor experiment tests how one subsequent loss measurement can turn that forecast into a curve extrapolation.

L Discussion

A training decision from a raw-text measurement. DMO predicts the rate of improvement with more data before training on the candidate corpus. That prediction, frozen before the test, transfers to new parent families, and changing sequence order moves both the compression slope and the fitted training exponent (Section 4). The compression slope and the floor become a loss prediction in the one-anchor experiment, where a low-data run supplies the amplitude that the raw text leaves undetermined.

The exponent and the remaining reducible loss jointly determine the value of more data. Both quantities appear in the derivative of the fitted law $L(D) = E + AD^{-\alpha_D}$, where the marginal improvement is

$$-\frac{dL}{dD} = \frac{\alpha_D}{D} (L(D) - E).$$

This is the natural quantity for comparing further training budgets. Comparing budgets this way is supported so far for corpus ranking and, with one anchor, for loss extrapolation over roughly ten times the anchor budget. Beyond that budget, a prospective allocation experiment can test whether initializing training decisions with these predictions reduces the tokens needed to reach a target loss.

Measure what the model learns at each scale. The loss decomposition in Section 3 suggests the next direct connection between compression and training: we would measure how much a model gains when its available context doubles, at several training budgets. That training-side gain, paired with the compressor’s own gain $G_{\text{comp}}(N)$, would show which scales contain accessible structure and which scales the learner has begun to use. Code, web, and math already provide an intervention that changes this scale profile within each corpus. Matching the two measurements on the same data and prediction units would test the proposed mechanism directly.

What the manifold picture earns, and what it forbids. The geometric result and the corpus result answer different questions. A supplied predictive state of dimension d_{pred} admits the best-in-class direct-head rate $\alpha_{\text{head}} \geq 2/d_{\text{pred}}$. The compression slope instead predicts the fitted data-axis exponent through a frozen empirical calibration, so ranking corpora by that slope ranks their measured α_D under this protocol, not their latent dimension.

Three things the frame forbids, and the prohibitions are its content. A circle and a k -torus emit the same symbol law, so no statistic of the symbols recovers dimension uniformly: any estimator’s expected absolute error is at least $(k-1)/2$ on one of them, at every sample size (Theorem B.7). The two halves of the geometric picture never compose over the class we state them on, because an expanding source leaves its own state in the future of the symbol stream while a predictive state has to live in the past (Proposition C.2). And information alone never fixes the data exponent: Proposition B.10 gives processes with identical block entropy at every length that differ only in how hard their next

symbol is to compute, and on five of them the deployed readout returns the same $\hat{\alpha}_{LZ}$, a spread of 0.0005 against a start-index noise of 0.0009 (Section Q).

Each of the three says where the measurement has to come from instead. Geometry needs expansion on the same clock, $d_{\text{geo}} = h_X / \lambda$, taken on the source rather than on its output. The data exponent needs a trained grid, which is why the map from $\hat{\alpha}_{LZ}$ to α_D is fit once and then frozen. That is the division of labour the paper is built on. The cost that used to fall on every candidate corpus is now one pass over its raw text, which gives the exponent and a floor ranking, plus one anchor run for the amplitude.

M Notation

Symbol	Where	Meaning
$\mathcal{M}_{\text{geo}}, d_{\text{geo}}, d_{\text{geo,amb}}$	Assumption A.1	geometric source manifold, Hausdorff dimension, and ambient dimension
$\mathcal{M}_{\text{pred}}, \nu, d_{\text{pred}}, d_{\text{pred,amb}}$	Assumption A.7	predictive-state manifold, its law, representation-relative dimension, and ambient dimension
μ, T	Assumption A.2	invariant data measure; generating dynamics
$\mathbb{P}_X$	Assumption A.3	law of the stationary symbolic source
$h_X, h_\mu(T)$	Assumptions A.3 and A.4	symbolic entropy rate; measure-theoretic entropy of the geometric dynamics, equal under the generating code
h_m, p_m, q_m	Proposition B.2	optimal loss using the last m symbols; true conditional law given those symbols; a predictor using that same context
$G_{\text{comp}}(N)$	Section J.9	compression gain $C_N - C_{2N}$ from doubling the character-window length; bits per character
$\mathcal{I}$	Section 2	sampled start indices for the compression measurement
$c_N(x, i), C_N$	Section 2	sampled raw-DEFLATE rate and its mean at character length N
$C(N)$	Equation (1)	functional notation for the finite-range curve template; not an unspecified asymptotic diagonal $w = w(N)$

Symbol	Where	Meaning
α_s, A_H	Equation (1)	structural exponent and amplitude of the Hilberg template; fit parameters, not source invariants (Remark H.3)
$\hat{h}, \hat{a}_{LZ}$	Section 2 and Remark B.9	empirical floor and rate-slope rankers; neither denotes a source invariant
$\hat{H}_0^{(3)}, \hat{a}_s^{(3)}$	Section P.4	explanatory readouts from three rates on the grid N, gN, g^2N ; $\hat{h}$ is the first of these evaluated at 128, 2048, 32768 characters, while the deployed slope $\hat{a}_{LZ}$ is a separate eight-length regression (Remark H.3)
α_{1NN}	Section D	exponent of the nearest-neighbour learning curve in the fixed-participation-ratio construction
γ_{CE}	Section I	tail exponent of the conditional entropy, $H_n^{\text{cond}} - h_X \asymp n^{-\gamma_{CE}}$
ρ	Section D	Spearman rank correlation; the subscripted $\rho_{\text{geo}}, \rho_{\text{pred}}$ below are metrics, not correlations
$U_h, \kappa, \bar{Z}$	Equation (8)	one-sided finite-sample upper bound on h_X , its window-overhead factor $b/(b - s_{\text{max}})$, and the mean per-symbol codelength it is built from
η	Proposition B.4	arbitrary positive slack in the approximation-rate exponent; the training learning rate is written lr
$\alpha_D, \alpha_P, \alpha_{D,P}$	Remark B.9	fitted data-axis exponent, fitted model-axis exponent, and data-axis exponent refit at one model size P
α_{head}	Proposition B.4	best-in-class cross-entropy exponent, $\sup\{a : \Delta\mathcal{L}(P) = O(P^{-a})\}$

Symbol	Where	Meaning
λ	Assumption A.4	conformal Lyapunov exponent (geometric side-input)
d_{PR}	Section D	participation ratio $(\sum_i \zeta_i)^2 / \sum_i \zeta_i^2$ of a spectrum; a finite-resolution effective dimension, distinct from the exact $\dim_{\text{H}}(\mathcal{M}_{\text{pred}})$ of Assumption A.7
$\beta_{\text{app}}, \beta_{\text{dyn}}, \gamma$	Remark E.1 , Assumption A.4 , and Lemma A.10	target smoothness, dynamical smoothness margin, and Hölder exponent of the source-to-state factor
P, D, N, N_{tot}	Section A.1	parameter count; training-token budget; text length; total certificate-corpus length
w, m, ℓ, b	Section A.1	LZ window; predictive context length; recurrence block length; certificate window length in source symbols
$\ell_{\text{utf}}, \bar{\ell}_{\text{utf}}$	Lemma B.11	UTF-8 codeword length in bytes and its mean; the bytes-per-character factor relating the byte and character clocks
$\rho_{\text{geo}}, \rho_{\text{pred}}, r, g$	Section A.1 and Lemma A.10	source and predictive metrics; metric radius; three-point grid ratio N_{i+1}/N_i

N Deferred Proofs and Supporting Results

Scope and organization. This section first gives the proofs deferred from the main-results narrative, including the exact wrapped-code construction behind the finite-sample certificate. It then proves the supporting results stated in the technical pillars and closes with the deferred oracle proofs. The preceding formal sections state the entropy, geometry, and head-rate results.

N.1 Proofs Deferred from the Main Results

Proof of [Lemma A.10](#). Set $s = \dim_{\text{H}} \mathcal{M}_{\text{geo}}$ and fix $q > s/\gamma$. Then $\gamma q > s$, so $\mathcal{H}^{\gamma q}(\mathcal{M}_{\text{geo}}) = 0$. For every $\delta > 0$, let $\{U_j\}_{j \geq 1}$ be a countable relative cover of $\mathcal{M}_{\text{geo}}$, with $U_j \subseteq \mathcal{M}_{\text{geo}}$ and $\text{diam}(U_j) \leq \delta$. The Hölder condition implies

$$\text{diam}(\Phi(U_j)) \leq K \text{diam}(U_j)^\gamma,$$

and $\{\Phi(U_j)\}_{j \geq 1}$ covers $\Phi(\mathcal{M}_{\text{geo}})$. Consequently,

$$\mathcal{H}_{K\delta^\gamma}^q(\Phi(\mathcal{M}_{\text{geo}})) \leq K^q \mathcal{H}_\delta^{\gamma q}(\mathcal{M}_{\text{geo}}).$$

Letting $\delta \downarrow 0$ gives $\mathcal{H}^q(\Phi(\mathcal{M}_{\text{geo}})) = 0$, hence $\dim_{\text{H}} \Phi(\mathcal{M}_{\text{geo}}) \leq q$. Since this holds for every $q > s/\gamma$, $\dim_{\text{H}} \Phi(\mathcal{M}_{\text{geo}}) \leq s/\gamma$. Finally, $\mathcal{M}_{\text{pred}} = \Phi(\mathcal{M}_{\text{geo}})$ identifies the image with the predictive-state support used in [Proposition B.4](#). $\square$

Proof of Theorem B.1. Let $L_{N,w}^{\text{WZ}} = NC_{\text{WZ}}(N;w)$ be the codelength. If the initial window is sent verbatim, Kraft's inequality gives $\mathbb{E}L_{N,w}^{\text{WZ}} \geq H(X_1^N)$. If it is instead treated as an initial database available to both encoder and decoder, the conditional form of the same inequality gives

$$\mathbb{E}L_{N,w}^{\text{WZ}} \geq H(X_{w+1}^N | X_1^w) = H(X_1^N) - H(X_1^w).$$

For fixed w , division by N and the entropy-rate identity $N^{-1}H(X_1^N) \rightarrow h_X$ show that $h_w \geq h_X$.

The expected-rate sliding-window theorem of [Wyner & Ziv \(1994\)](#) gives the converse bound

$$\limsup_{w \rightarrow \infty} \limsup_{N \rightarrow \infty} \mathbb{E}C_{\text{WZ}}(N;w) \leq h_X.$$

Together with $h_w \geq h_X$, this implies $h_w \rightarrow h_X$. The argument concerns the expected iterated limit; it neither asserts that a fixed- w sample-path limit exists under [Assumption A.3](#) nor controls a diagonal $w = w(N)$.

When $h_X > 0$, the match-length assertion is the theorem quoted as [Lemma G.2](#): under the same stationary ergodic hypothesis, $L_N / \log_2 N \rightarrow 1/h_X$ in probability. The waiting-time formulation in the same lemma gives $\ell^{-1} \log_2 R_\ell \rightarrow h_X$ almost surely for the recurrence time R_ℓ of an ℓ -block. This is the pathwise recurrence statement of Ornstein–Weiss; Shannon–McMillan–Breiman and Kac's lemma describe its mechanism, but the asserted convergence follows from the cited recurrence theorems rather than from that heuristic alone. $\square$

Proof of Proposition B.2. Let $\mathcal{F}_m = \sigma(X_{-m}, \dots, X_{-1})$ and $\mathcal{F}_\infty = \sigma(X_j : j \leq -1)$. Because $\mathcal{F}_m \uparrow \mathcal{F}_\infty$, conditioning reduces entropy and hence $h_m = H(X_0 | \mathcal{F}_m)$ is non-increasing. For every $a \in \Sigma$, the martingale convergence theorem gives

$$\mathbb{P}(X_0 = a | \mathcal{F}_m) \rightarrow \mathbb{P}(X_0 = a | \mathcal{F}_\infty) \quad \text{a.s.}$$

The alphabet is finite, so the entropy of these conditional probability vectors is bounded and continuous. Dominated convergence therefore yields

$$\lim_{m \rightarrow \infty} h_m = H(X_0 | \mathcal{F}_\infty).$$

Stationarity gives $h_m = H(X_{m+1} | X_1, \dots, X_m)$, whose limit is h_X by [Assumption A.3](#). Hence $H(X_0 | \mathcal{F}_\infty) = h_X$. Since $\mathcal{F}_m \subseteq \mathcal{F}_\infty$, the definition of conditional mutual information now gives

$$I(X_0; X_{-\infty}^{m-1} | X_{-m}^{-1}) = H(X_0 | \mathcal{F}_m) - H(X_0 | \mathcal{F}_\infty) = h_m - h_X.$$

Finally, let $q(\cdot | X_{-m}^{-1})$ be any predictor based on the last m symbols and let $p_m(\cdot) = \mathbb{P}(X_0 \in \cdot | \mathcal{F}_m)$. Its population log-loss decomposes as

$$\mathbb{E}[-\log_2 q(X_0 | X_{-m}^{-1})] = h_m + \mathbb{E}[\text{KL}_2(p_m \| q)] \geq h_m,$$

with equality for $q = p_m$ almost surely. Thus the optimal m -context loss is h_m , whereas the infinite-past optimum is h_X , proving the final claim. $\square$

N.1.1 Exact Wrapper for the Finite-Sample Certificate

The proposition fixes the source alphabet, document sampling law, and an injective symbol encoder $e : \Sigma \rightarrow \bigcup_{j \leq L_{\text{max}}} \{0, \dots, 255\}^j$ with prefix-free range and maximum codeword length L_{max} bytes. Apply e symbolwise, let Defl be a fixed lossless DEFLATE encoder, and put $q_b = \lceil \log_2(8L_{\text{max}}b + 1) \rceil$. For $y \in \Sigma^b$, the wrapper E_b uses one of two encodings:

1. If $|\text{Defl}(e(y))| < 8L_{\text{max}}b$, emit a one-bit compressed flag, the q_b -bit binary representation of the compressed length, and then $\text{Defl}(e(y))$.
2. Otherwise emit a one-bit literal flag followed by an $8L_{\text{max}}$ -bit index for each of $y_1, \dots, y_b$, for $8L_{\text{max}}b$ literal bits.

The flag distinguishes the two cases; the compressed header fixes the remaining length, and all literal encodings have one common length. The range is therefore prefix-free. The byte identity map and the UTF-8 character map give the two unit-specific instantiations in the proposition.

The coverage proof has three steps: Kraft's inequality lower-bounds the population mean, empirical Bernstein controls the sampled offsets within a realized corpus, and McDiarmid controls replacement of the independent documents.

Proof of Proposition B.3. All entropies in this proof are in bits, and all rates are in bits per source symbol.

Population lower bound. Fix an offset u , and let Y_u be the corresponding length- b window of source symbols. After deleting the symbols at its inserted separator positions, Y_u consists of contiguous pieces of lengths $a_1, \dots, a_j$ taken from distinct documents. For a stationary source,

$$H(X_1^a) = \sum_{i=1}^a H(X_i | X_1^{i-1}) \geq ah_X \quad (a \geq 1).$$

The documents are independent, and reinserting separators at known positions is injective. Consequently, if s_u is the number of inserted separator symbols in the window, then

$$H(Y_u) = \sum_{j=1}^J H(X_1^{a_j}) \geq (b - s_u)h_X \geq (b - s_{\max})h_X.$$

The wrapper is a code on Σ^b . Its literal branch is injective because $|\Sigma| \leq 2^{8L_{\max}}$, so $8L_{\max}$ bits index Σ ; in the compressed branch the prefix-free range of e makes a concatenation of codewords parse uniquely, so $y \mapsto e(y)$ is injective on Σ^b , and Defl is lossless, so y is recovered from $\text{Defl}(e(y))$. Both branches are therefore lossless on Σ^b . The range of E_b is prefix-free: the first bit separates the two branches; literal codewords all have the same length $1 + 8L_{\max}b$; and in the compressed branch the fixed-width header determines the number of remaining bits. Kraft's inequality therefore implies that the expected length is at least $H(Y_u)$. Writing Z_u for this length divided by b and then averaging over a uniform offset U gives

$$\frac{h_X}{\kappa} \leq \mathbb{E}_{\text{docs}, U}[Z_U]. \quad (25)$$

Offsets, window lengths, document lengths, and N_{tot} are counted in source symbols throughout, so stationarity is used only for the source process and no shift invariance is required of the byte image $e(Y_u)$.

Offset-sampling error. For a realized corpus $\mathcal{D}$, put $\bar{z}(\mathcal{D}) = \mathbb{E}[Z_U | \mathcal{D}]$. Conditional on $\mathcal{D}$, the sampled rates $Z_1, \dots, Z_M$ are i.i.d. in $[0, B_b]$: the literal branch has rate $8L_{\max} + 1/b$, and the compressed branch is entered only when $|\text{Defl}(e(y))| < 8L_{\max}b$, giving rate at most $8L_{\max} + q_b/b$, both at most B_b . The one-sided empirical Bernstein inequality, with failure probability $\delta/2$, therefore gives

$$\bar{z}(\mathcal{D}) \leq \bar{Z} + \varepsilon_M. \quad (26)$$

Document-sampling error. It remains to compare $\bar{z}(\mathcal{D})$ with its expectation over documents. Replacing document k can change only those windows whose starting offsets lie within $b - 1$ positions of that document. There are at most $L_k + b - 1$ such offsets, and each rate changes by at most B_b . Thus $\bar{z}$ has bounded difference

$$|\bar{z}(\mathcal{D}) - \bar{z}(\mathcal{D}^{(k)})| \leq \frac{B_b(L_k + b - 1)}{N_{\text{tot}} - b + 1} = c_k.$$

The lower-tail form of McDiarmid's inequality, again with failure probability $\delta/2$, yields

$$\mathbb{E}_{\mathcal{D}} \bar{z}(\mathcal{D}) \leq \bar{z}(\mathcal{D}) + \sqrt{\frac{1}{2} \ln(2/\delta) \sum_k c_k^2} = \bar{z}(\mathcal{D}) + \varepsilon_{\text{doc}}. \quad (27)$$

On the intersection of the events in (26) and (27), (25) gives

$$h_X \leq \kappa \mathbb{E}_{\mathcal{D}} \bar{z}(\mathcal{D}) \leq \kappa(\bar{Z} + \varepsilon_M + \varepsilon_{\text{doc}}) = U_h(b, M).$$

The union bound assigns this event probability at least $1 - \delta$. The bounded-difference argument uses both the fixed document lengths and independence across documents, which is why those conditions cannot be dropped from the stated coverage guarantee. $\square$

Proof of Proposition B.4. The canonical map $\Lambda: \mathcal{M}_{\text{pred}} \rightarrow \mathbb{R}^{|\Sigma|}$ is Lipschitz by Assumption A.8. The density and support hypotheses of Lemma E.3(a) are supplied by Assumption A.7. Applying that lemma to the coordinates of Λ in parallel (with the fixed alphabet size absorbed into the constant) shows that, for every $\eta > 0$, there is an $f_P \in \mathcal{F}_P$ such that

$$\varepsilon_P := \sup_{s \in \mathcal{M}_{\text{pred}}} \|f_P(s) - \Lambda(s)\|_\infty = O_\eta \left(P^{-1/(d_{\text{pred}} + \eta)} \right). \quad (28)$$

Fix s , set $p_s = p(\cdot | s) = \text{softmax}(\Lambda(s))$, and write $\delta = f_P(s) - \Lambda(s)$. If $q = \text{softmax}(f_P(s))$, then

$$q_i = \frac{p_{s,i} e^{\delta_i}}{\sum_j p_{s,j} e^{\delta_j}}, \quad \text{KL}_2(p_s \| q) = \frac{1}{\ln 2} \left(\ln \mathbb{E}_{i \sim p_s} e^{\delta_i} - \mathbb{E}_{i \sim p_s} \delta_i \right).$$

The random variable δ_i takes values in an interval of length at most $2\|\delta\|_\infty$. Hoeffding's lemma consequently gives the uniform bound

$$\text{KL}_2(p_s \| q) \leq \frac{\|\delta\|_\infty^2}{2 \ln 2} \leq \frac{\varepsilon_P^2}{2 \ln 2}. \quad (29)$$

This estimate is uniform over the simplex and therefore introduces no lower bound on the coordinates of p_s . Integrating (29) with respect to ν and using the infimum in (9) yields

$$\Delta \mathcal{L}(P) \leq \mathbb{E}_{s \sim \nu} \text{KL}_2(p(\cdot | s) \| \text{softmax}(f_P(s))) = O_\eta \left(P^{-2/(d_{\text{pred}} + \eta)} \right).$$

Because the bound holds for every $\eta > 0$, the definition of α_{head} gives

$$\alpha_{\text{head}} \geq \sup_{\eta > 0} \frac{2}{d_{\text{pred}} + \eta} = \frac{2}{d_{\text{pred}}}.$$

The constants may depend on η , so this argument does not assert the endpoint bound $O(P^{-2/d_{\text{pred}}})$. $\square$

Proof of Lemma B.5. By definition of conditional entropy,

$$\mathbb{E} H(p(\cdot | S_t)) = H(X_{t+1} | S_t).$$

Predictive sufficiency identifies the conditional law given S_t with the conditional law given $X_{\leq t}$ almost surely, and hence

$$H(X_{t+1} | S_t) = H(X_{t+1} | X_{\leq t}).$$

For a stationary finite-alphabet process, the latter is the limit of the finite-past conditional entropies and therefore equals the entropy rate h_X ; this is also the $m \rightarrow \infty$ identity proved in Proposition B.2. $\square$

Proof of Theorem B.7. Fix integers $q, k \geq 2$. On the circle $S^1 = \mathbb{R}/\mathbb{Z}$, let

$$T_1 x = q^k x \pmod{1}$$

with normalized Lebesgue measure and the partition into the q^k half-open intervals of length q^{-k} . On the torus $\mathbb{T}^k$, let

$$T_2(y_1, \dots, y_k) = (qy_1, \dots, qy_k) \pmod{1}$$

with Haar measure and the partition into the q^k half-open cubes of side length q^{-1} . Partition boundaries have measure zero. Both partitions are generating, and both maps are computable conformal expanding maps with absolutely continuous invariant measure, so the two systems satisfy Assumptions A.1, A.2 and A.4.

For the first system, the cylinder determined by any word of length ℓ is an interval of Lebesgue measure $(q^k)^{-\ell}$. For the second, such a word specifies ℓ successive base- q digits in each of the k coordinates, and its cylinder has Haar measure $q^{-k\ell} = (q^k)^{-\ell}$. Hence, in both systems,

$$\mathbb{P}(X_1 = u_1, \dots, X_\ell = u_\ell) = (q^k)^{-\ell}$$

for every word $(u_1, \dots, u_\ell)$. Their coded processes are therefore the same i.i.d. uniform process on q^k symbols, although

$$\dim_{\text{H}}(S^1) = 1, \quad \dim_{\text{H}}(\mathbb{T}^k) = k.$$

To see the consequence for consistency directly, choose $0 < \varepsilon < (k-1)/2$. If a symbolic estimator $\hat{d}_N$ were consistent on both systems, then under their common law both

$$\mathbb{P}(|\hat{d}_N - 1| < \varepsilon) \longrightarrow 1, \quad \mathbb{P}(|\hat{d}_N - k| < \varepsilon) \longrightarrow 1.$$

The two events are disjoint, so their probabilities cannot both tend to one. This proves the obstruction. [Theorem H.1](#) strengthens the same construction to the finite- N minimax lower bound $(k-1)/2$. $\square$

Proof of Proposition B.8. Equation (10) is the $O(1/\log_2 N)$ expected-redundancy result of [Savari \(1997\)](#) for the stated LZ78/LZW source classes. It is not a matching lower bound: if $r_N = O(1/\log_2 N)$, for example, the same statement remains true when r_N is of order N^{-c} for some $c > 0$. Thus these results alone do not determine a leading residual.

For a nondegenerate binary memoryless source under the phrase-code convention fixed in [Louchard & Szpankowski \(1997\)](#), that paper identifies the sharper expansion

$$\mathbb{E}[C_{\text{LZ78}}(N)] - h_X = \frac{A_{\text{LZ}} + \delta(N)}{\log_2 N} + O\left(\frac{\log_2 \log_2 N}{(\log_2 N)^2}\right),$$

including the nonzero leading logarithmic contribution and its bounded oscillation. Since $1/\log_2 N$ is not of order $N^{-1/d}$ for any fixed finite d , this source class already rules out the proposed polynomial form as a universal LZ78 leading law. These theorems are tied to LZ78/LZW; neither applies by substitution to the deployed fixed-window statistic. Its coverage statement must therefore come from [Proposition B.3](#), whose proof does not use a redundancy expansion. $\square$

Proof of Lemma B.11. (i) $\ell_{\text{utf}}(X_t)$ is a bounded measurable function of a stationary ergodic process, so Birkhoff's theorem gives $N^{-1} \sum_{t \leq N} \ell_{\text{utf}}(X_t) \rightarrow \bar{\ell}_{\text{utf}}$ almost surely, and the uniform bound $\ell_{\text{utf}} \leq 4$ upgrades this to L^1 .

(ii) The range of ϕ is prefix-free, so a concatenation of UTF-8 codewords parses uniquely and ϕ is injective on Σ_{ch}^N . Composing it with a lossless L whose range is prefix-free makes $x_1^N \mapsto L(\phi(x_1^N))$ an injective map with prefix-free range on the block alphabet Σ_{ch}^N , so Kraft's inequality gives $\mathbb{E}[L(\phi(X_1^N))] \geq H(X_1^N)$. Stationarity gives $H(X_1^N) = \sum_{i \leq N} H(X_i | X_1^{i-1}) \geq N h_X$, because each conditional entropy is at least the limit h_X of that non-increasing sequence.

(iii) Shannon–McMillan–Breiman on the character source gives $-N^{-1} \log_2 \mathbb{P}(X_1^N) \rightarrow h_X$ almost surely. Writing

$$-\frac{1}{K_N} \log_2 \mathbb{P}(X_1^N) = \frac{N}{K_N} \left(-\frac{1}{N} \log_2 \mathbb{P}(X_1^N) \right)$$

and applying (i), with $\bar{\ell}_{\text{utf}} \geq 1$, gives the stated limit. $\square$

N.2 Proofs of Supporting and Appendix-Only Results

Proof of Lemma E.3. Part (a): [Nakada & Imaizumi \(2020\)](#) prove that, for every

$$\overline{\dim}_M \text{supp } \nu < d' < d_{\text{pred,amb}},$$

deep ReLU networks achieve

$$\inf_{\hat{f} \in \mathcal{N}_P} \|\hat{f} - f\|_{L^\infty(\nu)} \leq C_{f,d'} P^{-\beta_{\text{app}}/d'}, \quad (30)$$

for an ambient β_{app} -Hölder target. The strict inequality in their hypothesis matters: their theorem does not, by citation alone, give one Big-O bound at the endpoint $d' = \overline{\dim}_M \text{supp } \nu$.

For the compact C^k_{pred} manifold $\mathcal{M}_{\text{pred}}$ of [Assumption A.7](#), the upper Minkowski, Hausdorff, and intrinsic dimensions all equal d_{pred} . The upper bound $d' < d_{\text{pred,amb}}$ is part of [Assumption A.7](#). Given $\eta > 0$, choose $0 < \eta' < \min\{\eta, d_{\text{pred,amb}} - d_{\text{pred}}\}$ and set $d' = d_{\text{pred}} + \eta'$ in (30). Each scalar coordinate of the Lipschitz logit map extends from $\mathcal{M}_{\text{pred}}$ to the ambient unit cube by McShane's theorem. Moreover, the strictly positive density in [Assumption A.7](#) implies $\text{supp } \nu = \mathcal{M}_{\text{pred}}$. If u is continuous on $\mathcal{M}_{\text{pred}}$, every nonempty relatively open subset of $\mathcal{M}_{\text{pred}}$ has positive ν -measure, and hence

$$\|u\|_{L^\infty(\nu)} = \sup_{s \in \mathcal{M}_{\text{pred}}} |u(s)|.$$

Thus (30) is a uniform bound on $\mathcal{M}_{\text{pred}}$. Since $P \geq 1$ and $\eta' < \eta$, its Lipschitz specialization is the stronger estimate $O_{\eta'}(P^{-1/(d_{\text{pred}}+\eta')}) = O_{\eta}(P^{-1/(d_{\text{pred}}+\eta)})$. When $\beta_{\text{app}} > 1$, this application requires the additional ambient or Whitney extension hypothesis stated in Remark E.1; Lipschitz regularity alone does not supply it.

Part (b): Throughout this part $\beta_{\text{app}} \in \mathbb{N}$ is an integer smoothness order. Yarotsky (2017) proves both bounds by a VC-dimension argument on piecewise-polynomial networks (his Theorem 4); the metric is the sup-norm and the variable is the parameter count.

For completeness, the counting reduction is as follows. Place $q \asymp \varepsilon^{-d_{\text{pred}}/\beta_{\text{app}}}$ disjoint β_{app} -smooth bumps on a regular grid and assign an independent sign to each bump. These 2^q functions remain in a fixed multiple of the unit $W^{\beta_{\text{app}},\infty}$ ball. A network that approximates every one of them in sup norm to less than half the bump height realizes, after thresholding at the bump centres, all 2^q binary labelings. Its VC dimension is therefore at least q . Substitution of the piecewise-polynomial network VC bounds in Theorem 4 of Yarotsky (2017) gives $P = \Omega(\varepsilon^{-d_{\text{pred}}/(2\beta_{\text{app}})})$ without a depth restriction and $P = \Omega(\varepsilon^{-d_{\text{pred}}/\beta_{\text{app}}} \log^{-2p_{\text{dep}}-1}(1/\varepsilon))$ when $L_{\text{net}} \leq C_{\text{depth}} \log^{p_{\text{dep}}}(1/\varepsilon)$. These are minimax statements over the target ball; they do not give a lower bound for each individual logit map.

Assumption A.8 enters this lemma only by fixing the target class of the downstream application ($\beta_{\text{app}} = 1$, the L_{logit} -Lipschitz logit map of Proposition B.4); the rates themselves depend only on the smoothness–dimension pair $(\beta_{\text{app}}, d_{\text{pred}})$ and the network class. $\square$

Oracle composition and finite-sample propagation.

Proof of Theorem C.1. By the conformal bridge and the ideal LZ limit,

$$d_{\text{geo}} = \frac{h_X}{\lambda}. \quad (31)$$

The Hölder factor has image $\Phi(\mathcal{M}_{\text{geo}}) = \mathcal{M}_{\text{pred}}$, so Lemma A.10 gives

$$d_{\text{pred}} = \dim_{\text{H}} \Phi(\mathcal{M}_{\text{geo}}) \leq \frac{d_{\text{geo}}}{\gamma}. \quad (32)$$

Because $d_{\text{pred}} > 0$ and $\gamma > 0$, (32) implies $d_{\text{geo}} > 0$; then (31) and $\lambda > 0$ imply $h_X > 0$. Applying Proposition B.4 and taking reciprocals in (32) yields

$$\alpha_{\text{head}} \geq \frac{2}{d_{\text{pred}}} \geq \frac{2\gamma}{d_{\text{geo}}} = \frac{2\gamma\lambda}{h_X}.$$

It remains to identify the reference loss. The state S_t is past measurable and predictively sufficient, so Lemma B.5 gives $\mathbb{E}H(p(\cdot | S_t)) = h_X$: the excess cross-entropy of the head is measured above the same Bayes floor the symbolic limit identifies.

Two different implications can fail. Without λ , the equal-law systems in Theorem B.7 have identical symbolic data but different d_{geo} , so a symbolic-only rule cannot recover (31) uniformly over the stated class. Without a sufficient factor, Assumption A.7 supplies no comparison between $\mathcal{M}_{\text{geo}}$ and $\mathcal{M}_{\text{pred}}$, and hence no replacement for (32). A separately supplied direct predictive state still supports Proposition B.4, and the symbolic entropy-rate result remains available in either case. $\square$

Proof of Corollary C.3. Let

$$E_h = \{h_X \leq U_h\}, \quad E_\lambda = \{\lambda \geq \lambda_L\}.$$

No independence is needed: the union bound gives

$$\mathbb{P}(E_h \cap E_\lambda) = 1 - \mathbb{P}(E_h^c \cup E_\lambda^c) \geq 1 - \delta_h - \delta_\lambda.$$

On $E_h \cap E_\lambda$, the conformal bridge and positivity give

$$d_{\text{geo}} = \frac{h_X}{\lambda} \leq \frac{U_h}{\lambda} \leq \frac{U_h}{\lambda_L}.$$

This proves the geometric conclusion without any predictive-state or factor hypothesis. If a fixed factor (Φ, γ) satisfying Lemma A.10 is also supplied, its deterministic Hölder comparison implies

$$d_{\text{pred}} \leq \frac{d_{\text{geo}}}{\gamma} \leq \frac{U_h}{\gamma\lambda_L}.$$

If the direct-head assumptions [Assumptions A.7](#) and [A.8](#) additionally hold, taking reciprocals and applying [Proposition B.4](#) gives

$$\alpha_{\text{head}} \geq \frac{2}{d_{\text{pred}}} \geq \frac{2\gamma\lambda_L}{U_h}.$$

Every applicable conclusion holds on the same event. If λ is known exactly, set $\lambda_L = \lambda$ and let E_λ be the sure event. $\square$

O Case-Study Details and Overlap Checks

O.1 The Document-Level Quadrant Scorer

The corpus-level rate-slope ranker $\hat{\alpha}_{LZ}$ of [Section 2](#) aggregates sampled windows at the eight dyadic lengths from 64 through 8192 characters. A single document is usually too short for that measurement. The document-level scorer for the selection experiments therefore uses the Higuchi fractal dimension D_{Hig} ([Higuchi, 1988](#)) of the mean-centred UTF-8 byte-value stream (adaptive $k_{\text{max}} \leq 32$), reported as

$$r_{\text{Hig}} = 2 - D_{\text{Hig}}.$$

This is a persistence-oriented score under a fixed encoding, not intrinsic geometric roughness. A separate corpus-level Higuchi calculation on 200 KB byte streams, with $k_{\text{max}} = 64$, correlates with $\hat{\alpha}_{LZ}$ at Spearman $\rho = 0.76$ on 13 corpora. Aggregating the document scores gives $\rho = 0.58$ in the 18-corpus comparison of [Section O.1](#).

The second document score is the whole-document bits-per-character compression rate c_i^{doc} : gzip for the Case 2 DCLM partition and order-16 PPM on OpenThoughts reasoning data. On the three OpenThoughts pools where both were run, the two compressors rank documents near-identically (Spearman $\rho = 0.97\text{--}0.99$, $n \geq 173\text{k}$ each), and their median splits agree on 93–97% of documents.

The four cells of the within-pool median split on $(c_i^{\text{doc}}, r_{\text{Hig}})$ define DMO-Doc, a separate document-level variant. Both coordinates rank documents within a pool. They are their own statistics, c_i^{doc} next to $\hat{h}$ and r_{Hig} next to the corpus-level rate slope $\hat{\alpha}_{LZ}$, and they carry neither of those estimands.

What the quadrants contain. We sampled 2,000 documents from DCLM-Baseline-1.0 web text and 2,000 from the OpenThoughts-114k reasoning dataset, scored each with the two coordinates above, and split at the within-dataset medians after min–max normalization within each dataset.

Quadrant	compression score	r_{Hig}	Observed tendency
FORMAL	low	high	coherent or structured text
RICH	high	high	diverse, difficult, or rare text
ROTE	low	low	repetitive or simple text
COLLOQUIAL	high	low	fragmented or templated text

Up to eight documents per cell were read.

Results. The DCLM cells hold 413/497/961/129 documents for FORMAL/RICH/ROTE/COLLOQUIAL, so the recorded split is unbalanced on both coordinates and its thresholds were not kept; the grouping stands as a description of this sample. FORMAL holds coherent narrative and technical prose, ROTE repetitive low-information text such as transcripts, polls and reference lists, RICH high-diversity, high-surprise fragments such as email threads and niche technical content, and COLLOQUIAL XML-encoded, templated or fragmented scrapes.

On OpenThoughts the FORMAL cell holds 65 of the 2,000 sampled documents: concise, well-structured problems with clean chains of thought. RICH holds harder and longer problems with denser reasoning, ROTE repetitive or trivially simple ones, and COLLOQUIAL rare, structurally malformed examples. These are readings of a few documents per cell. What tests the selection is the training outcome in the selection experiments.

Aggregate stability and the corpus-to-document gap. Split-half agreement measures how stable the corpus aggregates are to document resampling. For each of 18 corpora, we draw 4000 documents and split them into disjoint halves. We run the corpus-level readout (3000 windows) and the document-level scorer on each half, aggregate within each half, and Spearman–Brown-correct the half-to-half agreement across corpora. The resulting reliability estimates are high at this sample size.

Under a classical additive measurement-error model with uncorrelated errors, an observed Pearson correlation is attenuated by $\sqrt{r_{11}r_{22}}$, which is the model’s ceiling for it. We apply the analogous correction to Spearman correlations as a diagnostic; the raw values are what describe agreement between the reported corpus rankings.

comparison	raw ρ		model-based	
	ρ	reliability	reference	corrected ρ (95% CI)
$\hat{h}$ vs. aggregate c_i^{doc}	+0.84	0.99/0.99	0.99	+0.85 [0.53, 0.98]
$\hat{a}_{\text{LZ}}$ vs. aggregate r_{Hig}	+0.58	0.97/0.99	0.98	+0.59 [0.04, 0.87]

The compression-score aggregate correlation changes little after this correction (0.84 to 0.85); the raw value is below the model-based reference of 0.99. The two aggregates share an empirical ordering, although their levels differ by roughly 1.2 bits/char because a 32 KB window amortizes its dictionary over much more text than a median document. We use c_i^{doc} only to rank documents within a pool.

For persistence the raw correlation is 0.58 against a model-based reference of 0.98, and correcting split-half instability moves it to 0.59: the disagreement is in what the two statistics measure, not in how many documents were drawn. Across 20k bootstrap resamples the reference exceeds the raw correlation every time, and 99% put the corrected value below 0.90. The selection experiments are evaluated by their own training outcomes.

O.2 Case 2: Document-Score Quadrant Pretraining

Case 2 asks whether the separate whole-document selector changes pretraining outcomes. We partition one DCLM pool at the c_i^{doc} and r_{Hig} medians into four document-level ($c_i^{\text{doc}}, r_{\text{Hig}}$) score quadrants (FORMAL/RICH/ROTE/COLLOQUIAL, Section O.1) and pretrain transformers of four reported sizes (150M, 300M, 600M, 2B) on each quadrant at the Chinchilla budget $D = 20P$ tokens. Realized tokens, read out of each cell’s own training record, are 3.00B at 150M and 6.00B at 300M, exactly $20P$. At 600M and 2B the quadrant pool binds first and every cell consumes the whole of its own pool, 11.2–11.9B tokens, a 7% spread across quadrants; that is $D/P = 19.2$ at 600M and 5.8 at 2B. The protocol equalizes realized token exposure across the four quadrants of a given size and uses single-epoch sliding-window packing (stride = sequence length). This removes a direct token-count imbalance, but it does not eliminate all selection-correlated differences in document length, composition, or optimization. Held-out evaluation uses the accuracy suite and Paloma bits-per-byte throughout.

Why 1.2B is not reported. At 1.2B the four quadrants did not train on the same data. The grid restarted many times and some restarts switched pretokenized pools, so we charge each run’s wall clock to the pool named by the job that produced it. All sixteen reported cells trained entirely on one pool, dclm_quadrants_40b_tok. The four 1.2B cells did not, and they disagree with each other: ROTE took 53% of its training from that pool, FORMAL 9%, and RICH and COLLOQUIAL none, the rest coming from dclm_quadrants_43b_fast, a looser selection that a matched re-run puts 0.45 bits per byte behind. A quadrant contrast whose arms consumed different pools in different proportions is not a quadrant contrast. Wall clock is a proxy for optimizer steps and the two pools do not run at the same rate, so these shares are approximate.

The checkpoints say the same thing a second way. Reading global_step out of every surviving checkpoint and comparing the exported weights against it, 16 of the 20 cells end at 98.2–100% of their cosine schedule with the learning rate annealed below 10^{-6} , and all four 2B cells were scored from exactly those checkpoints. At 1.2B, RICH was scored at 55.4% of schedule with the learning rate still at 4.3×10^{-4} and it is the best number in that row, FORMAL at 88.5%, and COLLOQUIAL from a checkpoint that no longer exists. Two of the row’s cells have fully annealed seed replicates on the same pool, and those replicates land 0.018 and 0.020 worse than the cells they replicate: a checkpoint taken partway through a cosine does not beat the annealed endpoint of its own recipe by twice the seed standard deviation, in both quadrants. The four reported sizes still span 150M to 2B because 1.2B was interior to that range; results in Section 5.1.

Shared selection and uncertainty. The same scored partition is used at every model size. The 4/4 sign pattern therefore repeats the training comparison at four scales, not the document selection on four independent samples; we report it without a binomial p -value. Each cell is a single run. A six-arm mini-grid at 150M with three training seeds supplies the noise scale: the pooled single-run standard deviation is 0.0086 bits per byte, and that is the error bar drawn in Figure 4. The persistence gap reproduces in the mini-grid at 0.024 bits per byte, 4.9 standard errors. The separation margin between the worst high-persistence cell and the best low-persistence one does not: 0.021 from single runs

against 0.005 at three seeds. A difference of two extremes carries the seed noise twice, so the margin at 2B (0.003) sits below the floor and we do not read it.

The mini-grid also supplies the random-selection and length-only baselines the published grid lacks. They cannot be added to the grid itself: the tokenized pool behind it was deleted, and the surviving, more loosely filtered pool trains about 0.45 bits per byte worse. Built instead from one pool at 150M over three seeds, the six arms read **ROTE** 0.028 worse than random ($t(8) = 5.1$, $p = 0.0009$), **RICH** 0.015 better ($p = 0.03$), **COLLOQUIAL** 0.005 worse, **FORMAL** level with it, and length-only selection 0.004 worse. The persistence contrast inside the grid is therefore two effects of unequal size, **ROTE** falling behind the random draw and **RICH** pulling ahead of it. That the length-only arm, blind to both document scores, does not clear the draw is what separates the selection from a length proxy; **RICH** documents are longer than **COLLOQUIAL** ones, so the control was needed.

The accuracy suite (task axis). The same checkpoints are evaluated 0-shot on the multiple-choice suite; [Table 4](#) reports the per-quadrant means, with the protocol and the observed numerical alignment between the two document scores and the tested evaluation axes in its caption and in [Section 5.1](#). The grid does not establish that either score governs an axis causally.

O.3 Benchmark-Overlap Audits for Case 2

We audit benchmark overlap with three nested tests: strict equality after NFKC, case, and whitespace normalization; formatting-insensitive equality of deterministic word/math tokens; and token 5-gram Jaccard or directional containment of at least 0.80, with at least ten shingles. Rare shingles retrieve candidates, and every reported pair is then rescored exactly.

Case 2 scope. Against 2,640 Paloma items, the extant scored DCLM source-pool JSONLs contain no strict- or canonical-exact pair and three near matches, all in Wikitext-103: two evaluation items in **FORMAL** and one in **ROTE**. The `dclm_quadrants_40b_tok` bins actually consumed by the published training ladder were deleted. This is therefore an audit of the surviving scored source pool, not a token-for-token contamination certificate for those bins.

P Data and Measurement Details

DMO-Corpus implementation details. For a requested document budget, the loader scans at most four times that many records, keeps strings longer than 100 characters, reads the declared text field with fallbacks text and then content, and joins documents with `\n\n`. If the joined corpus is too short, it is repeated until it spans twice the largest requested window. In the three-point floor calculation, the implementation returns C_{32768} when $|C_{128} + C_{32768} - 2C_{2048}| < 10^{-9}$; otherwise it clips only negative extrapolates to zero. The committed three-point value for the one legacy row without a 100k-document artifact is the only artifact fallback. The $\hat{\alpha}_{LZ}$ path uses Python random seed zero. Historical $\hat{h}$ values are fixed realized Monte Carlo outputs, not retrospectively seeded reruns.

P.1 Data Availability

All experiments use public data and models, as distributed; exact sources (Hugging Face dataset paths, benchmark versions, model families) are given below for replication.

Case 1 fits data-scaling laws on the 51 corpora of [Table 8](#), which gives the Hugging Face dataset and subset behind each label, resolving a corpus label to the exact shard trained on; several labels are different subsets of one release, as with the five AutoMathText configurations and the three Proof Pile 2 splits, and the table is the only place that distinction is unambiguous.

Table 8: The 51 corpora in the Case 1 panel, with the Hugging Face dataset and subset each was drawn from. A dash in the last column marks a corpus with no source publication of its own, identified by path instead.

Corpus	Hugging Face dataset	Subset	Source
AutoMathText (arxiv)	math-ai/AutoMathText	arxiv-0.50-to-1.00	(Zhang et al., 2024)
AutoMathText (arxiv-full)	math-ai/AutoMathText	arxiv-full	(Zhang et al., 2024)

continued on the next page

Table 8, continued

Corpus	Hugging Face dataset	Subset	Source
AutoMathText (code)	math-ai/AutoMathText	code-0.50-to-1.00	(Zhang et al., 2024)
AutoMathText (py-code)	math-ai/AutoMathText	code-python-0.50-to-1.00	(Zhang et al., 2024)
AutoMathText (web)	math-ai/AutoMathText	web-0.50-to-1.00	(Zhang et al., 2024)
C4 English	allenai/c4	en	(Raffel et al., 2020)
CodeParrot Clean	codeparrot/codeparrot-clean	–	–
CodeXGLUE (python)	google/code_x_glue_ct_code_to_text	python	(Lu et al., 2021)
DCLM Baseline	mlfoundations/dclm-baseline-1.0	–	(Li et al., 2024)
Falcon RefinedWeb	tiiuae/falcon-refinedweb	–	(Penedo et al., 2023)
FineMath (3plus)	HuggingFaceTB/finemath	finemath-3plus	(Lozhkov et al., 2025)
FineMath (4plus)	HuggingFaceTB/finemath	finemath-4plus	(Lozhkov et al., 2025)
FineWeb	HuggingFaceFW/fineweb	default	(Penedo et al., 2024)
FineWeb-Edu	HuggingFaceFW/fineweb-edu	default	(Penedo et al., 2024)
FineWeb-HQ	epfml/FineWeb-HQ	–	(Messmer et al., 2025)
FreeLaw (CourtListener)	pile-of-law/pile-of-law	courtlistener_opinions	(Henderson et al., 2022)
GitHub Jupyter	codeparrot/github-jupyter-code-to-text	–	–
InfiWebMath (3plus)	HuggingFaceTB/finemath	infiwebmath-3plus	(Lozhkov et al., 2025)
Isabelle Proofsteps	xu3kev/proof-pile-2-proofsteps	–	(Azerbaiyev et al., 2024)
Open-Genome (DNA)	LongSafari/open-genome	stage2	(Nguyen et al., 2024)
OpenCodeInstruct	nvidia/OpenCodeInstruct	–	(Ahmad et al., 2025a)
OpenCodeReasoning	nvidia/OpenCodeReasoning	split_0	(Ahmad et al., 2025b)
OpenCoder (algorithmic)	OpenCoder-LLM/opc-annealing-corpus	algorithmic_corpus	(Huang et al., 2025)
OpenCoder (synth-QA)	OpenCoder-LLM/opc-annealing-corpus	synthetic_qa	(Huang et al., 2025)
OpenMathInstruct-2	nvidia/OpenMathInstruct-2	–	(Toshniwal et al., 2025)
OpenMathReasoning (CoT)	nvidia/OpenMathReasoning	default	(Moshkov et al., 2025)
OpenWebMath (full)	open-web-math/open-web-math	–	(Paster et al., 2024)
OpenWebText	Skylion007/openwebtext	–	(Gokaslan & Cohen, 2019)
Project Gutenberg EN	manu/project_gutenberg	–	–
Proof Pile 2 (algebraic-stack)	aklein4/proof-pile-2-fixed	algebraic-stack	(Azerbaiyev et al., 2024)
Proof Pile 2 (arxiv)	aklein4/proof-pile-2-fixed	arxiv	(Azerbaiyev et al., 2024)
Proof Pile 2 (open-web-math)	aklein4/proof-pile-2-fixed	open-web-math	(Azerbaiyev et al., 2024)
PubMed (full text)	armanc/scientific_papers	pubmed	(Cohan et al., 2018)
SlimPajama-6B	DKYoon/SlimPajama-6B	default	(Soboleva et al., 2023)

continued on the next page

Table 8, continued

Corpus	Hugging Face dataset	Subset	Source
SmolLM	HuggingFaceTB/smolLM	cosmopedia-v2	(Ben Allal et al., 2024)
cosmopedia-v2	corpus		
Stack Edu Python	zayduhri/stack-edu-python	–	(Ben Allal et al., 2025)
StackExchange	HuggingFaceH4/stack-exchange-preferences	–	–
StackMathQA	math-ai/StackMathQA	stackmathqafull-1q1a	(Zhang, 2024a)
StarCoderData	bigcode/starcoderdata	–	(Li et al., 2023)
TemplateGSM	math-ai/TemplateGSM	templatgsm-7473-1k	(Zhang, 2024b)
The Pile	EleutherAI/the_pile_	–	(Gao et al., 2020)
Deduplicated	deduplicated		
The Stack	bigcode/the-stack	–	(Kocetkov et al., 2023)
The Stack Dedup	bigcode/the-stack-dedup	–	(Kocetkov et al., 2023)
TinyStories	roneneldan/TinyStories	–	(Eldan & Li, 2023)
UniRef50 (protein)	agemagician/uniref50	–	(Suzek et al., 2015)
Wikipedia AR	wikimedia/wikipedia	20231101.ar	–
Wikipedia EN	wikimedia/wikipedia	20231101.en	–
Wikipedia JA	wikimedia/wikipedia	20231101.ja	–
Wikipedia KO	wikimedia/wikipedia	20231101.ko	–
Wikipedia RU	wikimedia/wikipedia	20231101.ru	–
Wikipedia ZH	wikimedia/wikipedia	20231101.zh	–

P.2 Supporting Measurement Figures

The main text reports the headline numbers; the figures below collect the corresponding measurements (Figures 7 and 8), and each caption states its scope and interpretation.

P.3 Synthetic Data Generator

To evaluate the sequence analysis algorithms, synthetic sequences of length $N = 200,000$ were generated, each isolating one structural property: a **fractal backbone** (deterministic L-system over the alphabet $\Sigma = \{A, B, C, +, -, /\}$ with production rules $A \rightarrow AB+C$, $B \rightarrow CA-B$, $C \rightarrow B/AC$, giving strong long-range deterministic correlations); **pure chaos** (i.i.d. uniform over the same alphabet, the maximum-entropy baseline at matched character diversity); **linear refinement** (mixtures S_p selecting per position between backbone and chaos through one fixed random mask at purity $p \in \{0.5, 0.7, 0.9\}$; the static mask makes the true signal at 50% a strict subset of that at 70% and 90%, simulating progressive denoising); and **long-period order** (a random length-1000 block repeated $N/1000$ times: high apparent entropy in short windows, strict global periodicity, an adversarial control against false-positive structure detection).

P.4 Derivation of the Explanatory Three-Point Readouts

Abbreviate $H_0 := H_0^{(w_0)}$. Let $N_2 = gN_1$ and $N_3 = g^2N_1$ with $g > 1$, and suppose the three values follow the nonconstant pure-power template $C_i = H_0 + A_H N_i^{-\alpha_s}$ with $A_H \neq 0$ and $\alpha_s > 0$. Writing $a = A_H N_1^{-\alpha_s}$ and $x = g^{-\alpha_s}$ gives $C_i = H_0 + ax^{i-1}$. The explanatory estimators are

$$\hat{H}_0^{(3)} = \frac{C_1 C_3 - C_2^2}{C_1 + C_3 - 2C_2}, \quad \hat{\alpha}_s^{(3)} = \frac{|\log((C_1 - C_2)/(C_2 - C_3))|}{\log g}.$$

Adjacent differences cancel the floor and their ratio cancels the amplitude:

$$\frac{C_1 - C_2}{C_2 - C_3} = \frac{a(1 - x)}{ax(1 - x)} = \frac{1}{x} = g^{\alpha_s}.$$

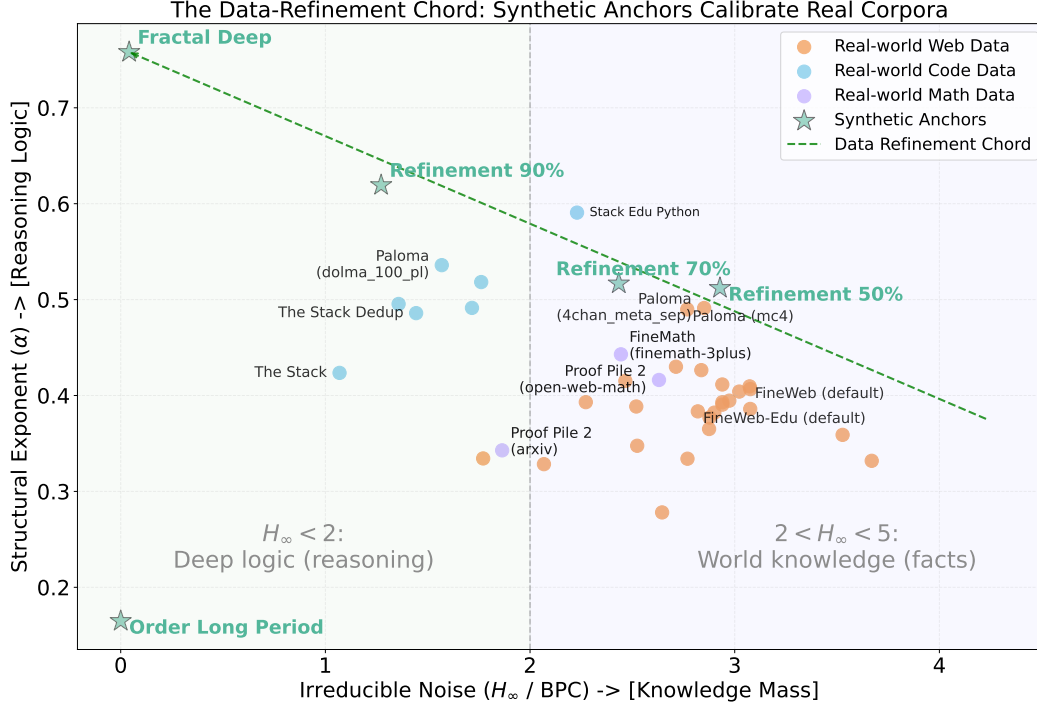


Figure 7: **Synthetic controls in the $(\hat{h}, \hat{\alpha}_s^{(3)})$ plane.** Stars mark five 6-symbol sources evaluated by the same three-point routine: a deterministic L-system, three nested corruption levels ($p = 50/70/90\%$) generated with one fixed mask, and a repeated 1,000-symbol block whose residual exponent falls once the measured windows exceed the period. The dashed line connects the nested denoising sequence; dots mark the 36 recorded real-corpus readouts in Figure 3. Generator details are in Section P.3.

Substitution in the slope definition gives $\hat{\alpha}_s^{(3)} = \alpha_s$. For the floor estimator,

$$\begin{aligned} C_1 C_3 - C_2^2 &= a H_0 (x - 1)^2, \\ C_1 + C_3 - 2C_2 &= a (x - 1)^2, \end{aligned}$$

so $\hat{H}_0^{(3)} = H_0$. Under the stated template, the denominator is nonzero and the adjacent-difference ratio is positive, so both estimators are defined. Outside it, the slope estimator still requires a positive adjacent-difference ratio. If measurement noise makes that ratio smaller than one, the absolute value reverses the sign of its logarithm and keeps the reported slope nonnegative; it does not make a nonpositive ratio admissible.

Status of the ansatz. The power-law residual form is a heuristic scaffold for this algebra, not a universal finite- N law for a Lempel–Ziv code. Known redundancy bounds do not justify a geometric power law, and the exact leading correction is logarithmic in the stated memoryless LZ78 case (Proposition B.8). The Case 1 floor retains a high-sample three-point extrapolation. Its deployed $\hat{\alpha}_{LZ}$, however, is the separate eight-point slope of the compression rate itself. It is neither $\hat{\alpha}_s^{(3)}$, the Hilberg parameter, nor $1/d$.

Q The Entropy-Matched Parity Family

Proposition B.10 is verified numerically before it is used. Brute-force enumeration over all 2^{k+N} prefixes at $k = 4$ reproduces $H(X_{1:N}) = N + \min(N, k)$ for $N = 1, \dots, 6$ and every support order $1, \dots, 4$, in 24 of 24 cases to within 10^{-9} ; the leakage rank at $k = 16$ is $\min(N, k)$ for every order, as the triangular argument requires. The check is pre-train/scripts/parity_family.py and its record is pre-train/reports/parity_family_identity.json.

Rendering. Each process step is written as one of " 0", " 1", " 2", " 3", which the GPT-NeoX tokenizer maps to exactly one token each (470, 337, 374, 495), verified before the corpora were

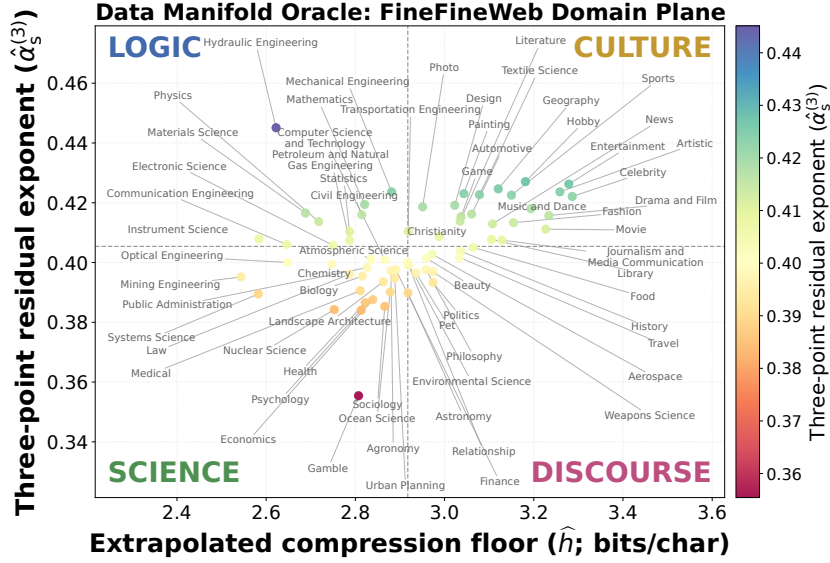


Figure 8: **Within-corpus domain comparison on FineFineWeb.** We apply the three-point statistic of Figure 1 once per domain; dashed lines mark the medians. Technical domains tend toward the low- $\hat{h}$ side and cultural domains toward the high- $\hat{h}$ side, so the same exploratory $(\hat{h}, \hat{\alpha}_s^{(3)})$ coordinates give a similar descriptive partition among domains within one corpus. All 66 non-outlying domains of the 67 scored are named, with leader lines where needed.

generated. One token is therefore one process step and the parity window is a token window, with no byte-pair grouping in between. Generation is chunked with a k -symbol carry so the parity is continuous across chunk boundaries; a break there would insert an unpredictable step per chunk and would not affect the arms equally.

The readout on the five arms.

$ a $	$\hat{\alpha}_{LZ}$	start-index SD	$\hat{h}$	predicted α_D
1	0.1981	0.0004	1.3792	0.1881
2	0.1979	0.0007	1.3784	0.1879
4	0.1979	0.0001	1.3783	0.1879
8	0.1984	0.0001	1.3789	0.1885
16	0.1982	0.0009	1.3783	0.1883

The cross-arm spread in $\hat{\alpha}_{LZ}$ is 0.0005 against a largest start-index SD of 0.0009. Both readouts land mid-range for natural corpora, $\hat{\alpha}_{LZ}$ near the panel median of 0.180 and $\hat{h}$ inside the code and math band, so these are not pathological inputs to the instrument. The protocol, the locked per-arm predictions and the decision rule are in pre-train/case1_parity_v1/preregistration.json, frozen before any cell was trained; success requires the fitted spread to exceed three times the largest leave-one-cell-out refit SD, with monotonicity in the support order reported separately.

R Calibration Certificate on a Source With a Known Exponent

Section D.2 gives the paired result. This section records the full sweep, the control that failed, and the window reach. All measurements use the deployed grid of Table 1 unless stated: raw DEFLATE level 9, wbits= -15, eight character lengths from 64 to 8192, the 300-window schedule, seed 0. Sources have 10^6 facts and 4M steps in the first sweep, 8M steps in the other two.

The unmodified family: monotone and over-determined. Seven Santa Fe sources at $\beta = 0.3, \dots, 0.9$ give

β	$\gamma_{\text{CE}} = 1 - \beta$	$\hat{\alpha}_{\text{LZ}}$	$\hat{\alpha}_s^{(3)}$	$\hat{\alpha}_s^{(3)}$ (floor-subtracted)	$\hat{h}$
0.30	0.70	0.311	0.536	0.547	0.31
0.40	0.60	0.253	0.548	0.566	0.57
0.50	0.50	0.235	0.610	0.602	0.91
0.60	0.40	0.208	0.648	0.628	1.30
0.70	0.30	0.183	0.686	0.647	1.76
0.80	0.20	0.158	0.708	0.661	2.25
0.90	0.10	0.135	0.741	0.692	2.70

$\hat{\alpha}_{\text{LZ}}$ orders γ_{CE} at Spearman +1.00. The ordering is real and it does not identify a mechanism: $h_X = H(K)$ and β are both functions of the same Zipf exponent, so $\hat{h}$ is rank-collinear with β and the floor-attenuation route and the memory route push the readout the same way. The paired arm in [Section D.2](#) is what separates them.

The three-point column is the one that fails. It rises monotonically as the true exponent falls, Spearman -1.00 against the known answer, and subtracting the fitted floor does not repair it. Neither column recovers the value $1 - \beta$. This is the measurement behind treating [Figure 1](#) as an exploratory display and deploying the eight-point slope instead.

A matched-entropy control that does not work. Appending an i.i.d. field U_t whose alphabet is chosen per β holds the per-step entropy rate at 10.63 bits and leaves $\hat{\alpha}_{\text{LZ}}$ flat: 0.161–0.168 across $\gamma_{\text{CE}} = 0.10$ –0.70, Spearman -0.64 . The construction cannot be read, because the compensating field shrinks from $|U| = 930$ to $|U| = 1$ as β rises and so changes the string in the direction that cancels the effect being measured. It is recorded because it was run.

Reach of the deployed grid. At 8192 characters a Santa Fe source has revealed only tens of distinct facts, so none of these measurements sits in the asymptotic N^β regime. Extending the grid to 2^{19} characters roughly halves every measured slope, 0.086 against 0.152 on the structureless control, without changing an ordering. The certificate reports what this grid resolves.

S Estimator Sensitivity

S.1 Compressors Ablation

This ablation compares how several compression algorithms behave as the available context length increases, using the rate-slope ranker $\hat{\alpha}_{\text{LZ}}$ and floor ranker $\hat{h}$. The comparison spans standard dictionary methods and more elaborate predictive models (figure in the supplementary note); it is a finite-sample comparison, not an ordering certified by an entropy theorem.

Evaluated algorithms: LZ (zlib Deflate; 32KB window + Huffman, the baseline), LZMA (larger dictionary + arithmetic coding), Zstd (large window + FSE), order-16 PPM (a conditional context model), and PAQ (context mixing).

Estimator selection. Lower finite-length rates can indicate a tighter coding model, but the *exponent* readout is a different question: what matters is how well $\hat{\alpha}_{\text{LZ}}$ ranks the fitted data-scaling exponent α_D . We choose it using a like-for-like comparison over the same $n = 38$ corpora of the Case 1 panel (identical subset, identical fitted α_D ; only the compressor behind $\hat{\alpha}_{\text{LZ}}$ varies), and repeat it on two wider corpus sets. Six of the ten exclusions were introduced when the panel expanded by eighteen corpora ([Table 9](#)).

Every cell of the LZ-slope family has an unadjusted per-cell permutation $p < 10^{-4}$: none of 20,000 label permutations matched the observed statistic in any of the twelve cells. No cell of the lower block clears $p < 10^{-2}$ on every corpus set. Because the estimator/corpus-set comparison uses this same panel, the per-cell values are not a selection-adjusted p -value for the selected maximum.

The four LZ-slope variants agree with one another at $\rho \geq 0.93$ (≥ 0.81 after controlling for $\hat{h}$), and each ranks α_D at $\rho \geq 0.63$ on all three corpus sets. The ordering runs opposite to compressor sharpness: the *coarser* dictionary methods rank α_D better than the sharper predictive model. The chosen statistic is the finite-scale slope of the compression rate itself, not an entropy estimate or the residual exponent in (1). PPM’s lower finite-length rate therefore need not improve this empirical ranking. This is why the deployed predictor uses the many-point zlib slope, and why the exponent is treated as an *empirical* ranker under a fixed estimator. The floor ranker $\hat{h}$ is likewise a finite-sample point estimate. Universal coding identifies the source invariant h_X asymptotically, but this fixed-window statistic does not

Table 9: **Estimator $\times$ corpus-set sensitivity of the exponent association.** Each cell reports Spearman $\rho(\hat{\alpha}_{LZ}, \alpha_D)$ / Pearson R^2 / leave-one-out Q^2 against *identical* fitted α_D ; only the compressor behind $\hat{\alpha}_{LZ}$ and the corpus set vary. Sets: the selected 38-corpus set; that set with the six exclusions added during the panel expansion restored; and no exclusions at all. The bold row marks the selected estimator; the upper and lower blocks separate the LZ-slope family from the alternative estimators.

$\hat{\alpha}_{LZ}$ estimator	$\rho(\hat{\alpha}_{LZ}, \alpha_D)$ / R^2 / Q^2 by corpus set		
	selected ($n=38$)	plus six ($n=44$)	all ($n=48$)
zlib 8-pt trimmed slope (selected)	.84 / .64 / .59	.76 / .48 / .43	.65 / .24 / .18
zlib slope, untrimmed	.84 / .64 / .59	.74 / .47 / .41	.63 / .23 / .17
Zstd slope	.84 / .65 / .59	.75 / .48 / .42	.64 / .21 / .13
LZMA slope	.76 / .55 / .51	.69 / .42 / .38	.67 / .28 / .24
PPM (order-16, 100k)	.46 / .42 / .33	.47 / .34 / .26	.36 / .18 / -.01
zlib 3-point readout	.32 / .19 / .12	.33 / .18 / .11	.24 / .00 / -1.37

inherit that consistency. The raw-rate slope $\hat{\alpha}_{LZ}$ and the three-point residual exponent $\hat{\alpha}_s^{(3)}$ agree only moderately on corpus rankings ($\rho \approx 0.52$), so the choice genuinely matters, and it was made on this panel, which is why the comparison is reported in full rather than summarized by its winner. We use the trimmed 8-point zlib slope for every reported *corpus-level* exponent.

The corpus-set comparison above is in-sample. A chronological check uses nine corpora whose α_D values were fitted *after* the estimator choice was fixed. On that small, not-independently-sampled panel, the LZ-slope family gives $\rho \in [0.80, 0.92]$ while PPM gives $\rho = 0.53$ and $R_{out}^2 < 0$; the gap survives partialling out the overlap with their superseded exponents (details with the Case 1 statistics).

T Additional document-selection results

T.1 Case 2: pretraining comparisons and data audit

The high- r_{Hig} quadrants (FORMAL, RICH) have numerically lower mean Paloma bits per byte than the low- r_{Hig} quadrants (ROTE, COLLOQUIAL) at every tested size. Averaged over the two quadrants on each side, the gap is 0.032 bits per byte at 150M and 0.023 at 2B, and does not shrink monotonically in between. Quadrant COLLOQUIAL improves most with scale, closing most of its gap to FORMAL by 2B. Task accuracy follows the compression-score axis: the left quadrants score numerically higher at every size, while differences within each side are small (Table 4).

The error bars in Figure 4 are one single-run seed standard deviation, 0.0086 bits per byte, measured in the 150M six-arm three-seed mini-grid; individual departures such as RICH at 600M sit inside them. All sizes reuse one scored partition and each cell is a single run. This grid compares the four quadrants against each other; the random and length-only baselines live in the mini-grid instead (Section O.2). The surviving scored DCLM source pool contains no strict or canonical exact Paloma matches, but the deleted training bins prevent a bin-level audit (Section O.3).