PREDICTING NEURAL SCALING LAWS WITHOUT TRAINING: A DATA MANIFOLD ORACLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural scaling laws set large-scale pretraining budgets, yet estimating a data scaling exponent requires fitting dozens of compute-intensive proxy models, one sweep per candidate corpus. 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. We measure how the DEFLATE rate of a corpus falls as the compression window grows, and map that one curve to the data scaling exponent α_D before any model is trained on the candidate corpus. We froze the measurement and its calibration on 38 pretraining corpora, then predicted five held-out corpus families before their training grids were run, reaching an MAE of 0.029 without GPU overhead against 0.084 for a proxy model trained on the candidate corpus itself. We pair that predictor with a theoretical boundary: symbol-level compressibility cannot uniquely resolve the dimension of a continuous source, at any sample size. By turning scaling law estimation from a multi-GPU training challenge into a data-profiling task, we replace a training sweep per candidate corpus with one compression pass.

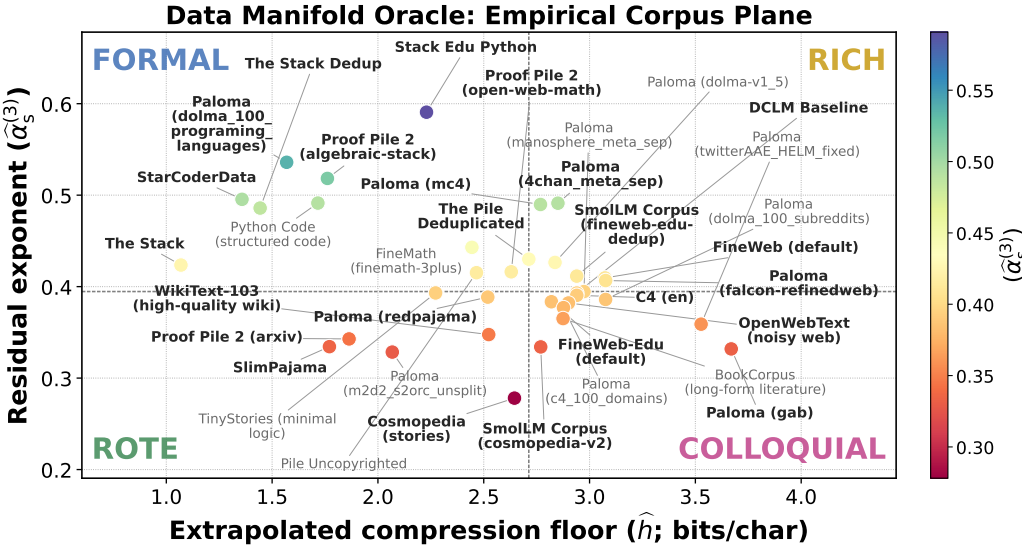


Figure 1: **Explanatory DMO plane.** We plot each corpus at its asymptotic floor $\widehat{h}$ and residual exponent $\widehat{\alpha}_s^{(3)}$ of Figure 3, at 10,000 documents each. The predictor deployed in Section 4 is the eight-point slope $\widehat{\alpha}_{\mathrm{LZ}}$ of Table 1; this plane displays a related pair of coordinates for explanation. Median splits name the four corners for the text in them: a low floor compresses well and a high residual exponent keeps gaining with window length, so FORMAL (low floor, high exponent) is coherent narrative and technical prose, RICH (high, high) diverse high-surprise fragments, ROTE (low, low) repetitive low-information text such as transcripts and reference lists, and COLLOQUIAL (high, low) templated or fragmented scrape. Section 5 reuses these names for a document-level split.

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

1 INTRODUCTION

A training corpus is chosen before anyone knows its data-scaling exponent. Neural scaling laws (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022) would supply that exponent, and whether a statistic read off the raw text can stand in for the training sweep they require is unsettled. Settling it takes a particular order: fix the statistic and its calibration first, then train the grids that could refute them.

The quantity being predicted has a direct training interpretation. At fixed parameter count N, loss depends on the training-token budget D as

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

where E is the loss floor and A sets the amplitude, so doubling the data multiplies the excess loss $L - E$ by $2^{-\alpha_D}$. Fitting them means training a family of models at several scales, over 400 for the Chinchilla exponents (Hoffmann et al., 2022) and 512 proxy runs 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.

A raw-text compression curve can predict that exponent. 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 fixes what a symbol-only measurement can never recover; 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, from the candidate text alone.

What the manifold oracle measures. The source manifold generates the symbols; a sufficient predictive state of dimension d_{pred} predicts the next one. Under the conformal assumptions of Assumption F.4, the entropy rate h_X and an independently measured expansion rate λ on the same clock recover the source dimension, $d_{\text{geo}} = h_X/\lambda$, and d_{pred} bounds a direct head's model-axis exponent, $\alpha_{\text{head}} \geq 2/d_{\text{pred}}$ (Section 3). Neither orders the fitted data-axis exponents, so we 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 symbol-only estimator is consistent for the dimension (Theorem G.7).

Measuring structure across scales. Longer text windows give a compressor access to more repeated structure. We measure the mean DEFLATE rate on random windows and summarize its dependence on length by a long-window level $\widehat{h}$ and a log–log slope $\widehat{\alpha}_{\text{LZ}}$; Figure 1 places each corpus on a related pair of coordinates and names its four corners. DMO-Corpus is $\widehat{\alpha}_{\text{LZ}}$ under a frozen linear calibration, a map with a precedent in the gzip statistic Pandey (2024) fit to scaling-law parameters on controlled grammars. We measure a slope across substring lengths instead and put it through a prospective test: estimator and calibration frozen on one panel, then six new parent families predicted before their neural outcomes existed (Sections 2 and 4).

From prediction to intervention. On the 38-corpus panel that chose the estimator, $\widehat{\alpha}_{\text{LZ}}$ ranks the fitted exponents at $\rho = 0.84$, jointly and at each model size from 16M to 1B, a reading we take for sensitivity before the prospective test evaluates the frozen map without refitting. The level $\widehat{h}$ ranks 26 near-converged neural floors at $\rho = 0.75$ but fails the same prospective checks, so it stays a ranker. Changing sequence order tests the same connection within one corpus: we permute text blocks in code, web, and math, retain the strings inside each block, and retrain, and both $\widehat{\alpha}_{\text{LZ}}$ and α_D move with block length, which a cross-corpus association alone cannot show. Combining $\widehat{h}$ and $\widehat{\alpha}_{\text{LZ}}$ 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.

Sections 2 to 4 define DMO-Corpus, state what the slope can and cannot mean, and test it: on the five of six new families that passed a fit gate fixed in advance, the frozen map predicts α_D to mean absolute error 0.029 against 0.045 for the better baseline (Table 2). A separate application scores documents rather than corpora: DMO-Doc selects pretraining data inside a pool with two related document-level features (Section 5).

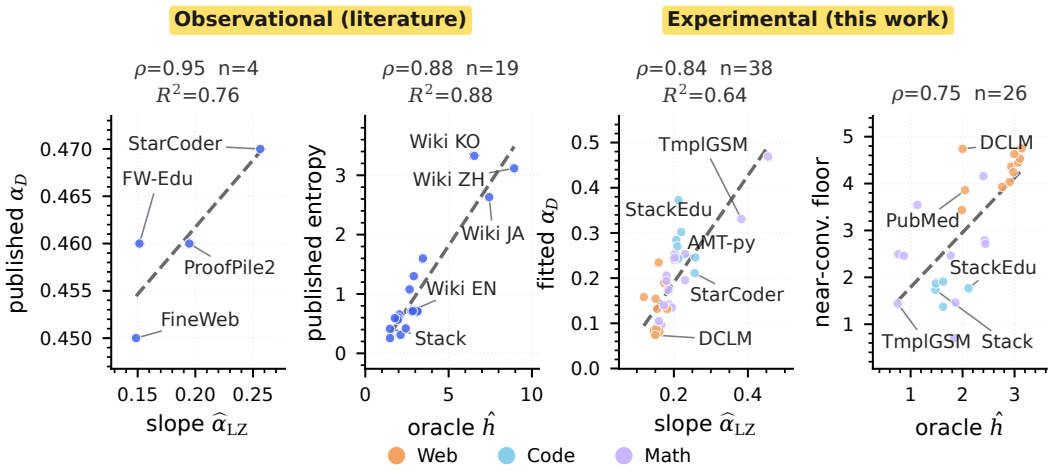


Figure 2: **Predictive raw-text statistics versus external references and fitted scaling quantities.** The left pair compares compression statistics with published scaling exponents and entropy references; the right pair shows our 38-corpus exponent calibration and 26 near-converged neural floor fits. Panels 1 and 3 use the slope $\widehat{\alpha}_{\mathrm{LZ}}$; panels 2 and 4 use the floor ranker $\widehat{h}$ in bits/character, against entropy in bits/character and neural floors in nats/token. Colors denote data type, and insets give Spearman ρ. Panel 3's dashed line is the frozen map $\widehat{\alpha}_D = 1.182\widehat{\alpha}_{\mathrm{LZ}} - 0.046$; the other dashed lines are ordinary least squares. Some entropy references share same-language proxies, and the neural floor fits require a $D = 20N$ run of at least 11B tokens. Corpus values and fitting details are in Table 9 and Section B.3; prospective transfer is tested separately in Table 2.

2 DMO-Corpus: Measuring a Compression Curve

We measure how fast a compressor's rate in bits per character falls as its window grows. The target is a rate too: the data-scaling exponent α_D is the rate at which loss falls as the training-token budget grows. Between those two rates we fit a calibration on existing training grids, and then freeze it.

2.1 The Prediction Target

Those exponents are fitted quantities, so we fix what produces them: a training grid. With N the model parameter count and D the training-token budget, the primary grids fit the additive form of Hoffmann et al. (2022),

$$L(N, D) = E + A_N N^{-\alpha_N} + A_D D^{-\alpha_D}. \tag{1}$$

Our target is α_D as fitted on the training grid of Table 1: it depends on the model family, tokenizer, optimization schedule, and budget range of that fit, so a forecast is evaluated under one specified training protocol; a per-size refit with a free floor gives a different, size-wise exponent (Section C).

2.2 The Compression Curve

Against that target we measure a curve. We concatenate corpus documents and sample character windows uniformly by starting position, with replacement. Let $\ell_C(x_{i:i+n})$ be the bits the fixed compressor produces on the window encoded as UTF-8. The compressed rate and its sample mean are

$$c_n(x, i) = \frac{\ell_C(x_{i:i+n})}{n}, \qquad C_n = \frac{1}{m_n} \sum_{j=1}^{m_n} c_n(x, i_j),$$

where m_n is the number of windows sampled at length n. Figure 3 sketches that measurement, and the DMO plane beside it, redrawn in Figure 1, names the four kinds of text Section 5 reuses. The measured object is this curve: a finite, compressor-specific statistic, not a latent state (Theorem G.7). That is why the map from curve to exponent is fit on trained grids instead of derived.

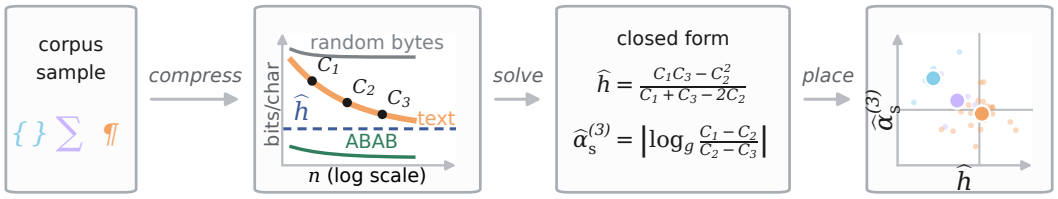


Figure 3: **Mapping a corpus to two coordinates without GPUs.** The curve panel is schematic: one fixed compressor encodes random bytes, repeated ABAB, and text across a geometric grid of window lengths. Random bytes remain near 8 bits per character, ABAB collapses toward zero, and text falls toward 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 $\widehat{h} = (C_1C_3 - C_2^2)/(C_1 + C_3 - 2C_2)$ and the residual exponent $\widehat{\alpha}_{\mathrm{s}}^{(3)} = |\log_g[(C_1 - C_2)/(C_2 - C_3)]|$. The final panel plots these recorded coordinates for 36 real corpora from the same 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 $\widehat{\alpha}_{\mathrm{LZ}}$ is a separate statistic, an eight-length regression of the raw rate, and is the input to the frozen map $\widehat{\alpha}_D = 1.182\widehat{\alpha}_{\mathrm{LZ}} - 0.046$ (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 $\widehat{\alpha}_{\mathrm{LZ}} \to \alpha_D$ map.

2.3 The Floor and the Slope

We take two statistics from the curve. The first is a floor. It starts from a finite-range Hilberg template (Hilberg, 1990),

$$C(n) = H_0 + A_{\mathrm{H}} n^{-\alpha_{\mathrm{s}}}. \tag{2}$$

For three lengths equally spaced on a log scale, write the measured rates as C_1, C_2, C_3 in increasing length order. The floor estimate is the template's long-window level,

$$\widehat{h} = \max\left\{0, \frac{C_1C_3 - C_2^2}{C_1 + C_3 - 2C_2}\right\}, \tag{3}$$

measured to test whether a loss floor could be forecast the same way; it fails the prospective test of Section 4.

The second is a slope. Over a fixed eight-length geometric grid $\mathcal{N}$ from 64 to 8192 characters,

$$\widehat{\alpha}_{\mathrm{LZ}} = |\mathrm{OLS\ slope}(\log C_n \text{ on } \log n)|\,, \qquad n \in \mathcal{N}, \tag{4}$$

with all lengths weighted equally, whatever their window counts. On the 38-corpus panel $\widehat{\alpha}_{\mathrm{LZ}}$ runs from 0.119 to 0.454 (median 0.180), and a synthetic source with no long-range structure reads 0.152 under the same grid, near the bottom of that range (Section M.2). Fitted to total rates, the slope is not biased by a floor-subtraction error, and the template exponent α_{s} is a different feature from $\widehat{\alpha}_{\mathrm{LZ}}$ (Section 3). Both are finite-window statistics of raw text, fixed with their compressor and grid before transfer; rerunning the slope at five start-index seeds moves it by at most 0.0022. Table 1 gives the budgets, Sections B.1 and Q the full specification.

2.4 The Fixed Prediction Rule

A calibration turns that slope into a prediction, and DMO-Corpus is that map, frozen:

$$\widehat{\alpha}_D = 1.182\,\widehat{\alpha}_{\mathrm{LZ}} - 0.046. \tag{5}$$

We chose the compressor, the grid, and these two coefficients on one panel of slopes and fitted exponents, so agreement there includes estimator selection and we read it as sensitivity; we then froze the map before measuring any new family or training its evaluation grid.

3 The Data Manifold Oracle: Information, Geometry, and Prediction

We froze the compression slope in Section 2 without saying what it measures. Here we read it three ways: as a property of the geometry behind the symbols, as a statistic of the symbols alone, and as an exponent of entropy decay. The geometry needs a scale the symbols do not carry, and the compressor's own floor attenuates the entropy reading; those two failures are why the map to α_D is calibrated rather than derived.

3.1 Source Dimension from Entropy and Expansion

Entropy and an independently measured dynamical scale combine into a single ratio. Under the conformal expanding-source assumptions of Assumption F.4, the source dimension is

$$d_{\text{geo}} = \frac{h_X}{\lambda}, \tag{6}$$

where h_X is the entropy rate in bits per symbol step and λ the average base-2 logarithmic expansion per step, and ideal universal compression identifies h_X in its asymptotic regime (Theorem G.1 and Proposition G.6). With that scale in hand, dimension also fixes a guaranteed rate for a learner given the state. For N-parameter deep-ReLU logit heads supplied with a smooth sufficient predictive state of dimension d_{pred}, Lipschitz conditional logits admit a best-in-class approximation exponent $\alpha_{\text{head}} \geq 2/d_{\text{pred}}$ (Proposition G.4 and Theorem H.1). Those results count parameters; α_D is fitted along the token budget.

3.2 The Same Symbols Can Have Different Source Geometries

The scale in (6) has to be measured, for a reason more data cannot remove: even complete knowledge of the symbolic process would not identify the dimension of its continuous source. Fix integers $q, k \geq 2$, equip the circle and the k-torus with their uniform invariant measures, and partition the circle into q^k equal intervals and the torus into q^k equal cubes. On those partitions the maps $x \mapsto q^k x \bmod 1$ and $y \mapsto qy \bmod 1$ generate the same i.i.d. uniform process on q^k symbols, while their source dimensions are 1 and k.

Because both sources emit that one law, any estimator reading a finite symbol sequence has the same output distribution under each, and cannot use the symbols to choose which geometric dimension generated them. Formally,

$$\max_{i \in \{1,2\}} \mathbb{E}_{W_i} |\widehat{d}_n - d_i| \geq \frac{k-1}{2}.$$

The bound holds at every sample size, so no amount of symbolic data closes the gap, and since every compression measurement is a function of the symbols, the same obstruction reaches the complete compression curve (Theorem G.7 and Proposition L.2).

3.3 What the Symbols Do Identify

The entropy rate of the symbol stream, unlike the source dimension, is within reach. Ideal universal coding identifies it under stationarity and ergodicity, and a multi-document procedure gives a finite-sample upper bound U_h with $\mathbb{P}(h_X \leq U_h) \geq 1 - \delta$ (Theorem G.1 and Proposition G.3). The floor extrapolation $\widehat{h}$ inherits no such guarantee, the formal counterpart of its prospective failure in Section 4.

An exact decomposition gives the compression curve and the training target a common language. A predictor q_m that uses the last m symbols has population log-loss $L(q_m) - h_X = (h_m - h_X) + \mathbb{E}\,\mathrm{KL}(p_m \,\|\, q_m)$, where h_m is the conditional entropy given that context and p_m the true conditional law: information beyond the context, plus error within it (Proposition G.2). More data shrinks the second term; the compression curve records how much of the first each doubling of the window releases. The slope reads that tail only through the floor, however. A stationary source's conditional entropy can decay as $n^{-\gamma_{\text{CE}}}$, the tail Cagnetta et al. (2026) measure on trained models, but a total compressed rate also contains a floor, so even for the idealized curve $C(n) = H + An^{-\gamma_{\text{CE}}}$ the local log–log slope stays below γ_{CE} and changes with window length (Section A). With no formula for α_D, we treat the slope's relation to α_D as an empirical hypothesis tested by frozen calibration (Propositions G.2 and G.8): the prediction rests on the slope and on a calibration we measured rather than derived.

4 Predicting Data Scaling from Raw Text

Three tests decide the frozen map: transfer to families the calibration never saw, an intervention inside a single corpus, and the training decision a forecast can carry. Table 1 collects the settings they share.

Table 1: **Experimental settings for the deployed predictor.** Counts and lengths are fixed before the new-family test. Full window counts, training schedules, and calibration range appear in Sections B.1 and B.2.

Component	Setting
Compression	Headerless DEFLATE, level 9, `wbits`$= -15$
Slope sample	10,000 documents; 8 dyadic lengths, 64–8192 characters
Floor sample	100,000 documents; 128, 2048, 32768 characters
Frozen calibration	$\widehat{\alpha}_D = 1.182\widehat{\alpha}_{\mathrm{LZ}} - 0.046$; selected on 38 corpora
Primary neural grid	GPT-2-style, 16M–1B parameters; sequence length 512
Data budgets	12, 40, 130, 250, 700 million training tokens
Long-run floor analysis	Separate runs at 2, 5, 10, 20 billion tokens

Table 2: **Prospective transfer: recorded numerical precision.** We apply 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	$\widehat{\alpha}_{\mathrm{LZ}}$	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

4.1 Prospective Transfer to New Families

We froze the compression statistic and its calibration before training on five new corpus families (Table 2). Each family gets a 6×5 grid of model sizes and token budgets; evaluating an exponent requires at least 16 of its 30 cells and a joint fit with $R^2 \geq 0.85$, which all five meet.

On these five, the frozen map has mean absolute error 0.029, against 0.048 for the fixed intercept and 0.045 for the domain mean. Its largest error is on Dolmino FLAN, and Table 2 gives each family's prediction beside its scaling-fit quality.

Those residuals also clear the five criteria fixed before any neural outcome existed, with rank correlation 0.800 and all five outcomes inside the fixed intervals of half-width 0.103; at five families the exact two-sided $p = 0.133$ carries no decision, and the test bounds transfer to new parent families rather than to corpora at large (Section B.7).

Training on the candidate corpus does not do better. An uncalibrated exponent fitted from the 16M-parameter row of the same grid, $L = E + AD^{-\alpha}$ over five budgets and 1.1B tokens per corpus, has mean absolute error 0.084 on these five families against 0.029 for the raw-text map. That gap is bias: one model size underestimates the joint exponent on every family, by two to three times, and rescaling it would need a calibration panel of its own. It ranks the five families as well as the slope does, so what the raw text yields without training is the value of the exponent and not only its ordering (Section B.8).

The 38-corpus panel we fitted the map on, with its scatterplots in Figure 2, chose the estimator as well as the coefficients, so we read it for sensitivity rather than transfer; Section C reports the leave-one-family-out, model-size and compressor checks.

4.2 Moving Structure across Scales Changes Data Scaling

Agreement across corpora may still reflect domain or preprocessing differences, so we perturb a single corpus: we cut its documents into blocks, permute their order, recompute the compression curve, and train fresh neural grids. The block lengths are 64, 256, 1024, and 4096 characters, with unshuffled text as the control, and we run it on StarCoder (Li et al., 2023), FineWeb-Edu (Penedo et al., 2024), and Proof Pile 2 arXiv (Azerbayev et al., 2024), with three model sizes and five training-token budgets in each of the fifteen arms.

Table 3: **Floor prediction transfers less well than exponent prediction.** The first two rows are within-panel associations; the third is the prospective diagnostic, scored by leave-family-out prediction error. Sections B.3, B.6 and B.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 $\geq$ 11B 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

The permutation preserves document length and each block's internal character string, and changes cross-block context and, after retokenization, token frequencies: what it moves is sequential organization, without isolating a single mediator (Section B.9).

Both quantities move with it. From the 64-character shuffle to the original order, α_D rises from 0.032 to 0.055 on FineWeb-Edu, 0.085 to 0.130 on StarCoder, and 0.053 to 0.098 on Proof Pile 2 arXiv, and the compression slope rises with it in all three. An exact stratified test that relabels arms within each corpus gives upper-tail $p = 0.00013$ under a no-association null that does not model training-seed variation.

The endpoint gap persists at larger model sizes across the same arms, and the curves show where the shuffle acts: writing $G_{\text{comp}}(n) = C_n - C_{2n}$ for the bits per character saved when the window doubles, the 64-character shuffle of code raises $G_{\text{comp}}(16{,}384)$ from 0.024 to 0.128, a 5.3$\times$ increase, while α_D falls from 0.130 to 0.085; across the fifteen arms, short-window gains associate positively with α_D and gains from 2048 characters onward negatively, after centering within each corpus (Section B.9).

If the slope carries the response, differently built texts with equal slopes should have similar exponents: mixtures of shuffled and intact StarCoder documents matched to a uniform-shuffle arm's slope land 0.0026 and 0.0035 from that arm's exponent, against 0.0105 and 0.0164 from a weighted average of the two ingredient exponents, 4.1$\times$ and 4.7$\times$ closer (Section B.9).

4.3 One Anchor Run Predicts the Return to More Data

A forecast of α_D says which corpora merit a scaling study; inside the calibration panel a separate experiment extends it to a whole loss curve. With the floor calibrated from the panel without the held-out corpus and one trained anchor at 12M tokens fixing the amplitude the calibration leaves free, median relative error over 882 larger-budget cells falls from 32%, when the anchor loss is carried forward, to 14%. No new family tests that extension (Section B.10).

The floor estimate itself does not transfer. A matched-token diagnostic that applies the same three-point formula to fixed validation-tail token streams, under a protocol set before reveal, has weaker rank agreement and larger error than the intercept, and the earlier cross-script frozen-floor test also fails (Sections B.6 and B.7).

Inside that same panel the estimate does correlate; Table 3 sets that against its prospective failure, and Section B.3 the two unit mismatches that weaken the in-panel reading.

5 Extension: Pretraining Data Selection

DMO-Corpus ranks corpora; a separate application of related document-level features ranks documents inside one pool. DMO-Doc scores each document by its whole-document compression rate and Higuchi persistence $r_{\text{Hig}} = 2 - D_{\text{Hig}}$. Aggregated over corpora these correlate with the DMO-Corpus floor estimate and compression slope at $\rho = 0.84$ and 0.58 and estimate neither (Section P.1). We split a DCLM (Li et al., 2024) pool at both medians into four groups, reusing the corner names of Figure 1: Formal, Rich, Rote and Colloquial.

The two high-persistence groups train better models at all four sizes. We pretrained 150M–2B at $D \approx 20N$ through 600M, with 2B capped by its own pool at 11.2–11.9B tokens; held-out

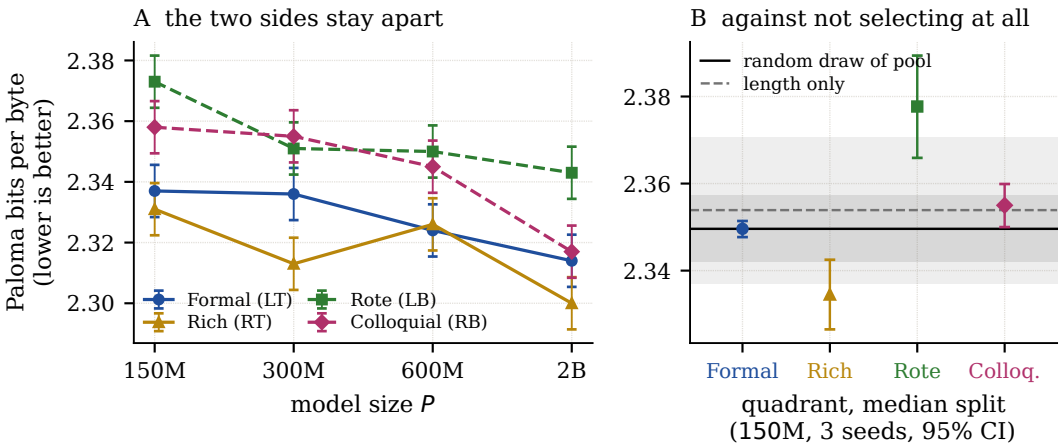


Figure 4: **The high-persistence side trains better models at every size from 150M to 2B, and its stronger quadrant beats a random draw of the same pool. (A)** Held-out mean Paloma bits per byte, lower is better, for DCLM split at the medians of the two document scores and pretrained at $D \approx 20N$ 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 $D \approx 6N$. 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. Fitted slopes are -0.022 to -0.038 bits per byte per decade of N and no pair separates by more than 1.2 standard deviations, so the quadrant effect is an offset rather than a change in how each pool scales. Each cell is one run, and error bars are one single-run seed standard deviation, 0.0086 bits per byte, measured from the grid in (B). **(B)** The same median split run as a six-arm grid at 150M over three seeds, against a random draw of the same pool. RICH trains 0.015 bits per byte better than that draw ($t = 2.7$); a length-only filter blind to both document scores does not (+0.004); FORMAL matches the draw to 0.000 and ROTE is 0.028 *worse*. This grid returns the persistence gap of (A) at 0.0243, 4.9 standard deviations. Shaded bands are the 95% intervals of the two reference arms. A fifth size, 1.2B, was trained and is not plotted; Section P.2 gives the reason.

Paloma (Magnusson et al., 2024) loss is lower on that side by 0.032 bits per byte at 150M and 0.023 at 2B (Figure 4A). A quadrant contrast cannot say whether selection beats no selection, so we ran a six-arm grid at 150M over three seeds: it returns that gap at 0.024, the stronger quadrant beats a random draw of the same pool, and a length-only filter does not (Figure 4B).

Against that same draw, task accuracy separates on the compression median instead, and moves further at a tighter dose: the top-scoring tenth of a FineWeb sample (Penedo et al., 2024) beats a draw of the same size at four sizes from 150M to 1.2B, by 0.6 to 1.1 points of mean 0-shot accuracy, and accuracy falls as τ loosens, while tightening it costs coverage, 0.068 bits per byte at 150M (Figure 5; per-arm record, accuracy suite and overlap audits in Sections D.2 and P.3 and Table 6).

6 RELATED WORK

Forecasts from trained grids. Fitting α_D conventionally means training model families (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022), at a cost set by corpus and protocol (Krajewski et al., 2025; Volkova et al., 2026). Cheaper variants still start from trained models, transferring information across them (Ruan et al., 2024; Brandfonbrener et al., 2024a) or estimating attainable capability boundaries from their evaluations (Zhang et al., 2026). DMO-Corpus forecasts that exponent from raw text, once calibrated.

Forecasts from raw text. From text alone, Pandey (2024) fit linear maps from gzip compressibility to scaling-law parameters on controlled probabilistic context-free grammars, and Han et al. (2026) transfer scaling predictions through source and target gzip ratios. We measure a slope across substring lengths, not a single ratio, and froze that calibration before testing it on new real-text families.

Forecasts derived from entropy. A second route to the exponent is derivation. Délétang et al. (2024); Huang et al. (2024); Shi et al. (2025) connect prediction, code length, and context-length effects in the settings each analyzes; Dębowski (2025) works through Zipf, Heaps, and Hilberg laws; and Cagnetta et al. (2026) obtain $\alpha_D = \gamma_{\mathrm{CE}}/(2\beta_{\mathrm{corr}})$ from a conditional-entropy estimate that uses trained-model n-gram losses; Hilberg residuals reflect long memory and mutual information (Hilberg,

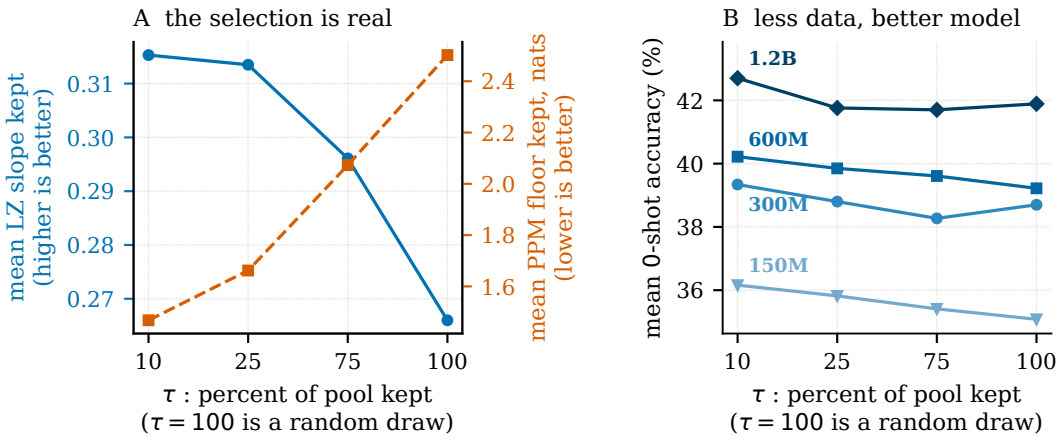


Figure 5: **Keeping the top τ% of a pool by the two document scores beats a random draw of the same size, at four model sizes from 150M to 1.2B. (A)** The selection separates the arms on both scores. Mean LZ rate slope and PPM floor of the documents each arm actually trained on, read back from the selected-document 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 together account for 0.74 of the 0.88-point mean. Each cell is one run. This sweep keeps a tenth of the pool where the median split of Figure 4 leaves each quadrant about a quarter, which is the dose that separates the two designs.

1990; Dębowski, 2011), and Proposition G.8 and Theorem G.7 bound what a finite-window slope says about source dimension.

Forecasts of source geometry. Kolmogorov-complexity and entropy results characterize symbolic information under their stated conditions (Brudno, 1983; Ornstein & Weiss, 1993; Lutz & Lutz, 2018), and quantitative recurrence reaches source geometry only with a supplied dynamical scale (Wyner & Ziv, 1989; Boshernitzan, 1993; Barreira & Saussol, 2001; Saussol et al., 2002; Galatolo, 2007; Ledrappier & Young, 1985; Young, 1982); we supply that scale as an external expansion rate, and our counterexample exhibits two sources with identical symbols and different dimensions (Proposition G.6 and Theorem G.7).

Forecasts of which documents to keep. Rather than the exponent, current pipelines forecast which documents to keep, and proxy classifiers, importance resampling, and reference-model perplexity each need a learned proxy or reference model (Li et al., 2024; Xie et al., 2023; Marion et al., 2023; Brandfonbrener et al., 2024b). The closest line trains first: Liu et al. (2024) regress loss on mixture weights from 1M-parameter proxy runs, Magnusson et al. (2025) rank recipes from small runs, and Jiang et al. (2025) fit per-domain scaling laws online and reweight during training. Because we measure DMO-Corpus before training, each further candidate costs one compression pass. Section 5 selects inside a pool with related document-level features; Section E covers representation geometry and dimension estimation.

7 Conclusion

In this work, we introduced the Data Manifold Oracle (DMO), a training-free framework that predicts neural data scaling exponents directly from raw text compression dynamics. By replacing compute-heavy proxy model sweeps with a single compression pass, DMO achieves high prospective forecasting accuracy on unseen corpus families (MAE 0.029) and establishes a causal link between sequential text structure and scaling behaviors. We further present a theoretical boundary on symbol-level compressibility and demonstrate that our document-level extension, DMO-Doc, effectively enhances pretraining dataset quality across model scales. Ultimately, DMO recasts scaling law estimation from costly compute sweeps to instantaneous, principled raw-data profiling.

AI USE STATEMENT

Generative AI tools were used to edit the manuscript for readability and presentation, including the presentation of proofs; to identify related literature and check reference metadata; to format the paper; and to assist with experiment and analysis code under the authors' specification.

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.

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. Hugging Face blog, 2024. URL https://huggingface.co/blog/cosmopedia.

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.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

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 Délétang, 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.

Ronen Eldan and Yuanzhi Li. TinyStories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*, 2023.

Kenneth Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. Wiley, 2 edition, 2003.

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. Dataset release, 2019. URL https://skylion007.github.io/OpenWebTextCorpus/.

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, Sadhika 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.

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 dataset card, 2025. URL https://huggingface.co/datasets/HuggingFaceTB/finemath.

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.

Jack H. Lutz and Neil Lutz. Algorithmic information, plane Kakeya sets, and conditional dimension. *ACM Transactions on Computation Theory*, 10(2):7:1–7:22, 2018.

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.

Ian Magnusson et al. Paloma: A benchmark for evaluating language model fit. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2024. arXiv:2312.10523.

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.

Bettina Messmer, Vinko Sabolčec, and Martin Jaggi. Enhancing multilingual LLM pretraining with model-based data selection. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2502.10361. Releases the FineWeb-HQ corpus.

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 Brixi, 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, Ebtesam 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. URL https://www.cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama.

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, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2302.03169.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2017. arXiv:1610.01145.

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. Technical report, Math-AI, 2024a. URL https://stackmathqa.github.io/StackMathQA.pdf.

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 Compression measurement details 17

B Case-1 Evidence and Reproducibility 17

C Calibration and Model-Size Sensitivity 25

D Document Selection: Full Record 26

E Related Work on Representation Geometry 26

F Setup: Compression, Source Geometry, and Prediction 27

G Theory: From Compression to Geometry and Model Scaling 30

H Available Conclusions and Required Information 34

I Predictive Dimension Controls Direct-Head Approximation 36

J Local Dimension of the Source Measure 39

K From Symbolic Recurrence to Geometric Dimension 39

L Why Symbols Do Not Identify Geometric Dimension 41

M Controlled Checks of the Theory's Components 43

N Notation 44

O Deferred Proofs and Supporting Results 46

P Case-Study Details and Overlap Checks 51

Q Data and Measurement Details 53

R The Entropy-Matched Parity Family 57

S Calibration Certificate on a Source With a Known Exponent 57

T Estimator Sensitivity 58

A Compression measurement details

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

$$\frac{H(X_{1:n})}{n} - h_X = \frac{1}{n}\sum_{j=0}^{n-1}(H_j^{\mathrm{cond}} - h_X).$$

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

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

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 that estimate's gap from γ_{CE}. In particular, the raw-rate slope $\widehat{\alpha}_{\mathrm{LZ}}$, the template exponent α_{s} in (2), and the conditional-entropy-tail exponent are distinct measurements.

Compressor comparisons. We selected the compressor and grid on the Case 1 calibration panel, and the prospective family test evaluates the selected estimator without refitting. Section T.1 and Table 10 report the agreement among the LZ-slope variants and the full estimator sweep.

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.

B Case-1 Evidence and Reproducibility

This section answers four questions about Case 1: how we measured the neural targets, how the calibration changes across corpus sets and model scales, whether the exponent map transfers to new parent families, and whether changing sequence order changes both the compression statistic and the fitted exponent. Data sources and preprocessing are given in Section Q, and compressor comparisons in Section T.1.

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

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

Its value depends on the reference predictor and on how predictions are pooled across folds, so we fix both: the reference $\widehat{y}_i^{\mathrm{ref}}$ 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: the predictor is refit with one parent family removed, that family's rows are predicted, and the pooled predictions are scored. The reconstruction returns the calibration line $1.1821\widehat{\alpha}_{\mathrm{LZ}} - 0.0461$, in-sample $R^2 = 0.635$ and $\rho = 0.845$, and the five frozen domain means, so the artifact and the reported calibration agree.

B.1 Full Compression Measurement Protocol

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

$$L(N, D) = E + A_N N^{-\alpha_N} + A_D D^{-\alpha_D}. \tag{7}$$

Our target is the fitted α_D, which depends on the model family, tokenizer, training schedule, and range of budgets. At fixed N, the model-size term is constant; the sensitivity analysis instead fits $L(D) =$

$E_N + A'_N D^{-\alpha_{D,N}}$ 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 B.2 gives the training schedules. We report the two exponent analyses separately in Sections B.4 and 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

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

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

$$\widehat{\alpha}_{\mathrm{LZ}} = |\mathrm{OLS\ slope}(\log C_n \text{ on } \log n)|\,, \qquad n \in \{64, 128, \ldots, 8192\}. \tag{8}$$

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

$$\widehat{\alpha}_D = 1.182\widehat{\alpha}_{\mathrm{LZ}} - 0.046. \tag{9}$$

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 T.1 records the estimator comparisons; Section B.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:

$$\widehat{h} = \max\left\{0, \frac{C_{128}C_{32768} - C_{2048}^2}{C_{128} + C_{32768} - 2C_{2048}}\right\}. \tag{10}$$

This expression follows from the finite-range template

$$C(n) = H_0 + A_{\mathrm{H}} n^{-\alpha_{\mathrm{s}}}. \tag{11}$$

The rate slope $\widehat{\alpha}_{\mathrm{LZ}}$ 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 Q. The two empirical outputs are $Q_{\mathrm{DMO}}(X; \mathcal{I}) = (\widehat{h}(X; \mathcal{I}), \widehat{\alpha}_{\mathrm{LZ}}(X; \mathcal{I}))$, where $\mathcal{I}$ denotes the sampled start indices.

B.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 (N, D) factorial per corpus, with specified batch schedules and learning-rate rules. These choices hold the schedule fixed across cells.

The panel is a full 6×5 factorial in (N, D) per corpus, with GPT-2-style models at $d_{\mathrm{model}} \in \{256, 512, 768, 1024, 1536, 2048\}$ ($n_{\mathrm{head}} = d_{\mathrm{model}}/64$, 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 $B^*_{\mathrm{tok}} = 22.91\, D^{0.47}$, apply the paper's $0.8\times$ safety buffer, and floor to a power of two in sequences (minimum 16). This N-independence lets each fixed D share one batch across all six sizes, so the loss-versus-N comparison of Section B.5 does not mechanically vary the batch schedule with N.

Learning rate. A width rule, $\mathrm{lr} = 0.0015 \times 512/d_{\mathrm{model}}$, 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, because the uncapped rate diverges at the narrowest width.

B.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 B.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 $D = 20N$ 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 $[0.56, 0.97]$, $n = 19$), spanning English, code, Chinese, Japanese, Korean, Russian, French, and German. The non-Latin rows come from one source and one method, context-aware PPM entropy rates extrapolated to infinite length (Takahira et al., 2016, Table 1), which continues the same-language-proxy convention the English rows already use; German uses an older letter-prediction estimate (Küpfmüller, 1954). Wikipedia ZH is the highest point on the $\widehat{h}$ axis and lies on the fitted line rather than off it: dropping it gives $\rho = 0.86$, $R^2 = 0.86$ ($n = 18$). Two rows were corrected to that same source, Chinese from a 0th-order character entropy and Russian from single-pass compression rates on one translation, and Spearman is 0.88 before and after both.

The floor: neural fits. Against our own fits we use the 26 corpora whose $D = 20N$ run consumed at least 11B tokens; the other 25 request $D = 20$B but exhaust their tokenized pool at ≈ 0.9B, and a run $23\times$ short of budget is not a converged floor. 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 is associated with the floor residual and not with the slope residual. 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. Returning to nats per token does not remove the effect: after regressing the raw floor on $\widehat{h}$, the residual still correlates with `chars_per_token` at Spearman 0.40 (permutation $p = 0.022$), and adding $\log($`chars_per_token`$)$ raises R^2 from 0.48 to 0.69. 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 have different loss levels. Fitting one shared slope with a per-type intercept raises R^2 from 0.483 to 0.742 (Figure 6). The shared slope adds little beyond the intercepts: within web the association is $\rho = +0.64$ ($n = 11$), within math $+0.41$ ($n = 10$), and within code -0.10 ($n = 5$), where $\widehat{h}$ spans only 27% of its pooled range. So $\widehat{h}$ 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.

Matched cells put the highest R^2, 0.57, at the *shallowest* depth. The published floor is a minimum over each corpus's own converged cells, so corpora are read at different model sizes and token budgets; measuring every floor at one common (N, D) cell removes that inhomogeneity. The shallowest cell, 16M parameters at 2B tokens, beats the deepest matched cells at 19B tokens (0.48–0.49), where a training-depth explanation predicts the reverse.

One corpus is a documented exception. AutoMathText (py-code) has the largest negative residual on this panel and is the most repetitive of the seven corpora probed for duplication, at a distinct word-8-gram ratio of 0.73 against 0.90–0.98 for two controls chosen to separate the plausible confounds. Setting it aside leaves $\rho = 0.77$ on the remaining 25; the panel and the headline retain all 26.

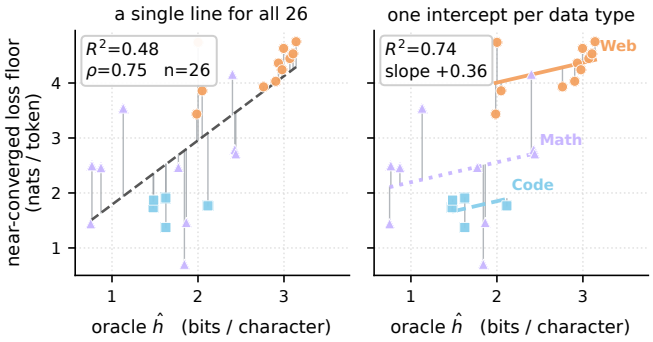


Figure 6: **The floor panel's unexplained variance is the offset between data types, not noise around a single relationship.** Both sides show the same 26 corpora and the same $\widehat{h}$ axis as Panel 4 of Figure 2; grey segments are residuals, whose squared lengths are what R^2 summarizes. Left: one intercept for every corpus, $R^2 = 0.483$. Right: one intercept per data type with a single shared slope of $+0.362$ on $\widehat{h}$, $R^2 = 0.742$, with fitted lines drawn over each type's own $\widehat{h}$ range. The residuals that shrink are between-type, so most of the gain comes from ordering 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 B.3. The rank statistic quoted in the panel, $\rho = 0.752$, is computed on the pooled points and is unaffected by this refit.

For $\widehat{\alpha}_{\mathrm{LZ}}$ on the same 26 corpora the analogous residual correlation is not detected (Pearson -0.33, permutation $p = 0.17$): a log–log slope is algebraically less sensitive than a level to rescaling the token axis, which can contribute to the difference between the floor and slope results.

Distinct information in the compression statistics. They correlate at $\rho = -0.82$, but on the same 41 corpora $\widehat{h}$ ranks α_D at only $\rho = -0.54$, compared with 0.84 for $\widehat{\alpha}_{\mathrm{LZ}}$. 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 Q.4. Conversely, $\widehat{\alpha}_{\mathrm{LZ}}$ retains partial $\rho = 0.74$ after controlling for $\widehat{h}$. The two measurements are related, but the floor statistic $\widehat{h}$ 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 $Q^2 = 0.59$. We chose both the corpus set and the eight-point trimmed log–log `zlib` estimator on this panel. Six of the ten corpus exclusions were introduced when 18 corpora were added, and $\widehat{\alpha}_{\mathrm{LZ}}$ was one of six estimators compared in Section T.1.

The full estimator/corpus-set comparison is therefore more informative than its best cell (Table 10). 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 $\widehat{h}$). 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 estimator and the sharper predictive compressors (order-16 PPM and LZ 100k) do not show a comparable corpus-set-wide association. Proposition G.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. We selected the estimator on this same corpus panel, so this reading is 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-$\widehat{\alpha}_{\mathrm{LZ}}$ 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 $D \leq 700$M 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 $\widehat{\alpha}_{\mathrm{LZ}}$ 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-$\widehat{\alpha}_{\mathrm{LZ}}$ 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 $\widehat{\alpha}_{\mathrm{LZ}}$ 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 $\widehat{h}$ is 0.74, rising to 0.80 on the English Web/Code/Math subset.

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

No weakening is detected from 16M to 1B parameters. We fit $L(D) = E_N + A_N D^{-\alpha_{D,N}}$ separately at each model size, excluding the $D = 20$B cells used for the floor proxy, and correlate $\alpha_{D,N}$ with $\widehat{\alpha}_{\mathrm{LZ}}$ over the 29 language corpora. For the selected `zlib` estimator, Spearman correlations at $N = 16\mathrm{M}/51\mathrm{M}/124\mathrm{M}/253\mathrm{M}/512\mathrm{M}/1\mathrm{B}$ 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 T.1). Neither result extrapolates beyond 1B.

Each fit must estimate E_N freely. Reusing the largest model's floor forces its residual toward zero and creates an artificial trend of -0.111 per model size ($0.75 \to 0.20$). After controlling for the correlated floor statistic (Pearson -0.70 between the two measurements), $\widehat{\alpha}_{\mathrm{LZ}}$ still ranks $\alpha_{D,N}$ 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.

B.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 $L(N) = E_D + A_D P^{-\alpha_N}$ across six model sizes at each fixed D—shows a weaker positive association with $\widehat{\alpha}_{\mathrm{LZ}}$ ($\rho \approx 0.75$ rather than 0.84), with substantial overlap with the data-axis result.

Eligible data budgets. A model exponent is meaningful only where the sweep is parameter-limited. We therefore require at most half the corpora to have loss non-monotone in N. The $D \leq 40$M budgets fail because 98% and 84% of corpora are data-starved; at $D = 10$B the largest model regresses for 57% of corpora because of the batch-tier artifact in Section B.2, rather than divergence (dropping N_5 restores $\rho = +0.46$).

Result and data-axis overlap. Across the five eligible budgets, $\widehat{\alpha}_{\mathrm{LZ}}$ ranks α_N at $\rho = +0.53$ to $+0.79$. At the least noisy budget, $D = 5$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.

Descriptively $\alpha_N \geq 2\widehat{\alpha}_{\mathrm{LZ}}$ holds for all 24 corpora, which the one-sided achievability bound permits without predicting: that bound orders no corpora, and $\widehat{\alpha}_{\mathrm{LZ}}$ is not a dimension estimate.

Fit controls. The joint Chinchilla fit's α_N column cannot be used because its bound at 1.5 binds for 9 of 38 corpora; refitting each budget with bound 4.0 leaves none at the bound. Pairwise rank agreement across eligible budgets is $+0.40$ to $+0.91$.

B.6 Prospective Case-1 Pilot

The exponent pilot improves on the intercept, whereas the floor mapping does not. On FineWeb-2 ZH, FineWeb-2 AR, and Lucie the fixed line $\alpha_D^{\text{pred}} = 1.182\widehat{\alpha}_{\text{LZ}} - 0.046$ gives MAE 0.074 against 0.111 for the intercept, with the compression statistics computed before the neural outcomes existed and no refit. Two parent families and no ex-ante threshold make this a calibration diagnostic rather than the family-generalization test.

The two non-Latin floor points lie outside the published floor range, and their fixed mapping has MAE 5.194 nats/token against 0.684 for the intercept. The empirical cross-script floor mapping is scoped to Latin-script corpora accordingly.

B.7 Six-Family Blind Transfer and Measurement Checks

On the five families, the fixed map has MAE 0.029, compared with 0.048 for the intercept and 0.045 for the domain baseline (ratios 0.60 and 0.64). 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 4. With five 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 residuals at recorded precision.

The test comprised BigPatent, Hacker News, FLAN, YouTube Commons, and Amazon Books, each trained on a 6×5 (N, D) factorial, for 150 cells. Before any neural outcome existed, the protocol fixed the family set, the rounded map $\widehat{\alpha}_D = 1.182\widehat{\alpha}_{\text{LZ}} - 0.046$, 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, the roster, and the domain assignments were all fixed before training. All 150 cells passed the technical checks and all five outcomes were revealed once, so the exponent metrics describe this five-family test.

Table 4: Frozen prospective protocol and observed outcome. All five decision criteria passed.

Item	Recorded specification
Prediction	$\widehat{\alpha}_D = 1.182\widehat{\alpha}_{\text{LZ}} - 0.046$, fixed before outcomes
Families and grid	Five named families; 6×5 cells per family; 150 cells total
Fit eligibility	At least 16 cells and joint-fit $R^2 \geq 0.85$
Baselines	Global calibration mean 0.1828; 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	All five families evaluable; MAE ≤ 0.0785; MAE ratios to both baselines ≤ 1; $\rho \geq 0.7$; interval coverage ≥ 0.8
Reported outcome	Five families evaluable; MAE 0.029; $\rho = 0.800$; interval coverage $5/5$

The fixed domain means, computed from the frozen 38-row calibration artifact, are 0.094 for web/prose, 0.228 for math, 0.248 for code, 0.162 for reference/science, and 0.140 for other. The roster assigns Amazon Books, Hacker News, and YouTube Commons to web/prose; 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.

Measurement repeatability. Resampling barely changes the family ordering. Before reveal, the DMO-Corpus statistic was recomputed on the same source corpora with window seeds 0–19. Across families, the propagated $\widehat{\alpha}_D$ SD is 0.0010–0.0067; seed zero reproduces every reported value exactly. The rank correlation with the reported 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 $\widehat{\alpha}_{\text{LZ}}$ 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 $\widehat{\alpha}_{\mathrm{LZ}}$ was lower in all six families: the median original-minus-shuffled difference was 0.026, with exact two-sided sign $p = 0.03125$. As an implementation-sensitivity baseline, XOR relabeling preserved byte length and equality structure and changed $\widehat{\alpha}_{\mathrm{LZ}}$ by median absolute 0.0034. 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.

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 $p = 0.683$) and leave-family-out MAE 2.064 bits/token against 1.414 for the intercept. The protocol was fixed before reveal to test whether the earlier floor gap was merely a unit or tokenizer mismatch: it reads the same token streams the models trained on, measures raw-DEFLATE rates at three window lengths, and applies the same three-point extrapolation, with neural E refit after reveal in bits per token on each family's primary cells. A full-bin sensitivity gives the same answer.

B.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 this 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 $\widehat{\alpha}_{\mathrm{LZ}}$	0.029	0.800	0.067	none
Fixed intercept	0.048	—	—	none
Domain mean	0.045	0.866	0.100	none
16M model, 5 budgets	0.084	0.900	0.042	1.13B
16M model, 3 budgets	0.085	1.000	0.008	0.18B
16M model, 2-point slope	0.115	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 underestimates the joint exponent by a factor of two to three on every family, because the joint fit pools six sizes and the one-size fit cannot separate E from A over five budgets. Rescaling the proxy needs a calibration panel, the same cost the raw-text map already incurred once.

On ranking the proxy matches $\widehat{\alpha}_{\mathrm{LZ}}$. The three-budget proxy orders the five families exactly and the five-budget proxy misses by one adjacent swap, against $\rho = 0.800$ for $\widehat{\alpha}_{\mathrm{LZ}}$; at $n = 5$ one swap changes ρ by 0.1, 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 comparison supports two claims, on level and on cost: 0.029 against 0.084 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 (0.029, 0.048, 0.045) from full-precision $\widehat{\alpha}_{\mathrm{LZ}}$.

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

This experiment tests whether moving sequential structure within a fixed corpus changes both $\widehat{\alpha}_{\mathrm{LZ}}$ and the fitted exponent. It changes both 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 measurement, 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 ℓ=64 to 0.055 at the control on web; 0.085 to 0.130 on code; 0.053 to 0.098 on math), and $\widehat{\alpha}_{\mathrm{LZ}}$ 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 ℓ=4096 arm is near its control. The shuffle 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 enumerates the upper tail $\Pr_{\mathrm{perm}}(\bar{\rho}_{\mathrm{perm}} \geq \bar{\rho}_{\mathrm{obs}})$ 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. We did 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 ($t = 6.5$–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 T.1.

Floor-side behaviour. The floor estimate is the intervention's control on code and web — $\widehat{h}$ 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 ($\widehat{h}_{\mathrm{LZ}}$ falls from 1.83 to 1.30 at ℓ=64). 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, changes their relative positions, and shifts the compression gains it produces toward larger windows. Writing $G_{\mathrm{comp}}(n) = C_n - C_{2n}$ 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 ($n \leq 128$), insignificant at n=512–1024, and negative from n=2048 on. On code, $G_{\mathrm{comp}}(16{,}384)$ rises from 0.024 in the control to 0.128 at $\ell = 64$. The pattern implicates the location of predictable structure, not simply its total amount. The nine n values are correlated measurements from the same curve summarized by $\widehat{\alpha}_{\mathrm{LZ}}$, so this decomposes the intervention result rather than replicating it.

Matched-slope control. On two StarCoder pairs, mixtures with the same $\widehat{\alpha}_{\mathrm{LZ}}$ as a uniform-shuffle arm also have fitted exponents closer to that arm than to a document-mixture prediction. We shuffle a Bernoulli-q fraction of documents at ℓ=64 and leave the rest intact, tuning q=0.674 to match the ℓ=256 arm and q=0.364 to match ℓ=1024; both realized slopes land within 0.003 of their targets and both training grids completed all 15 cells. The observed exponents are 0.110 against 0.108 slope-matched and 0.100 mixture-predicted, and 0.130 against 0.133 and 0.113, making the matched-slope predictions $4.1\times$ and $4.7\times$ closer at the grid's refit-noise scale. This two-pair, one-corpus check is consistent with the scalar statistic without establishing that a scalar is generally sufficient. Across the observational corpus set, no single $G_{\mathrm{comp}}(n)$ 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 ℓ=64) from 16M–124M to 253M and 512M parameters on both corpora and refit $L(D) = E_N + A_N D^{-\alpha_{D,N}}$ at each model size with a free floor. The reported quantity is the control-minus-shuffled exponent gap at each size.

Table 5: **Calibration and sensitivity evidence.** Spearman coefficients measure rank agreement with fitted data exponents unless marked partial Pearson. Rows use different corpus sets and answer different questions; they are not independent replications. Full comparisons and uncertainty estimates are in Sections B.3, B.4 and T.1.

Analysis	Sample	Result
Corpus-set sensitivity	All corpora ($n = 51$)	$\rho = 0.65$
	Language ($n = 41$)	$\rho = 0.84$
	Fit-qualified language ($n = 38$)	$\rho = 0.84$, $R^2 = 0.64$
Leave-one-family-out	27 parent families	MAE: slope 0.0514; intercept 0.0785; domain 0.0510
Fixed-size data fits	29 language corpora; 16M–1B	$\rho = 0.88$–0.96
Control for floor estimate	Same fixed-size analysis	Partial Pearson 0.73–0.85
LZ estimator sensitivity	Four slope variants	Pairwise $\rho \geq 0.93$; exponent $\rho \geq 0.63$
Later neural measurement	Nine known corpora; 270 cells	$\rho = 0.87$

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, consistent with larger models using order-dependent information that the shuffle disrupts. Across the measured 32× size range the endpoint gap persists alongside the rank stability in Section B.4.

B.10 One Anchor Run Yields an Empirical Curve Extrapolator

Construction. The calibrated statistics 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) fixes the amplitude: $\widehat{L}(D) = \widehat{E} + (L_0 - \widehat{E})(D/D_0)^{-\widehat{\alpha}_D}$. We take the $D_0 = 12$M-token cell of each corpus/model-size row as the anchor and predict its larger-budget cells. The $\widehat{\alpha}_{\mathrm{LZ}} \to \alpha_D$ map and every calibration constant are fitted without the predicted corpus. Because $\widehat{h} \geq L_0$ in 29% of rows, the method also calibrates the floor without the held-out corpus rather than using $\widehat{h}$ 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 2–5× the anchor budget, 17.6% at 5–15×, and 66.8% beyond 30×. On this panel, the useful range is therefore roughly 10× the anchor budget, where median errors are about 10–15%. Longer extrapolation still requires a direct sweep.

C Calibration and Model-Size Sensitivity

The prospective test of Section 4 fixes the map and asks what it transfers. Inside the 38-corpus panel we fitted the map on, we can examine the same association at much larger n. Table 5 collects that evidence: language and fit-quality filters strengthen the association relative to the full corpus panel, and because we chose the estimator on that panel as well as the coefficients, the strengthening is internal fit, not transfer.

We can separate internal fit from transfer inside the panel as well: leaving out whole parent families reduces leakage from closely related corpora. Under that split the slope predictor's absolute error is essentially the domain mean's; it still improves on the intercept, and its rank correlation is higher than the domain baseline's (Section B.3). The two tests therefore measure different things: prospectively the map has lower absolute error than both baselines, while the family-level leave-one-out gives a rank advantage at a comparable error.

We fit the target jointly across model sizes, so we also refit the data curve separately at each size to check whether pooling creates the association. It does not: rank agreement stays high when each size gets its own data-curve fit, the association survives controlling for the floor estimate, and refitting on

Table 6: **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 P.2; Section D.1 reports the held-out bits-per-byte results.

common data budgets preserves the ordering. Freeing each size's floor matters, however. Fixing it to the largest model's floor manufactures a decline in the correlation (Section B.4), so the stability we report depends on defining the neural target that way.

We also varied the compressor and the window estimator. Alternative LZ slope variants give similar rankings across the tested corpus sets. Predictive compressors, including order-16 PPM, have less consistent associations on those sets. We later measured known corpora, which also support the slope we selected. Those corpora had contributed to estimator selection, so that measurement tests robustness to new neural measurements rather than transfer.

D DOCUMENT SELECTION: FULL RECORD

The two document features and the quadrant partition are specified in Section P.1, which also names the four within-pool cells; Section 5 reports the pretraining outcome.

D.1 CASE 2: QUADRANT PRETRAINING

Section 5 gives the outcome; this section gives the protocol. Splitting DCLM (Li et al., 2024) by the document compression and r_{Hig} medians, we pretrain four sizes (150M–2B) at $D \approx 20N$ through 600M, with 2B capped by the quadrant pool at 11.2–11.9B tokens ($D \approx 6N$), and evaluate held-out loss on Paloma (Magnusson et al., 2024). All four sizes train on the same scored partition. At 150M we ran all 18 quadrant, random and length-only cells over three seeds; there FORMAL matches the random draw. We trained a fifth size, 1.2B, and excluded it: its four quadrants drew from two different document pools in different proportions, and we scored two of its cells from checkpoints stopped mid-schedule. Full task results and overlap checks are in Table 6 and Section P.3.

D.2 SELECTION STRENGTH

A tighter dose raises accuracy further. We rank a 350B-token FineWeb sample (Penedo et al., 2024) by the two document scores and keep the top τ percent. Mean 0-shot accuracy over eight tasks then beats a random draw of the same pool at all four sizes from 150M to 1.2B, by 0.6 to 1.1 points, and accuracy falls as τ loosens (Figure 5). Every arm draws its 200,000 documents uniformly from its own top-τ pool, so $\tau = 100$ is that random draw; the $\tau = 10$ and $\tau = 100$ arms share 0.1% of their documents and the mean scores of what each arm trained on move with τ in the predicted direction. Counting each task once across sizes, six of the eight improve at 0-shot and all eight at 5-shot, sign test $p = 0.008$; the gain is concentrated in SciQ (+4.2 points) and ARC-Easy (+1.7), between them 0.74 of the 0.88-point mean. Bits per byte goes the other way, 0.068 worse at 150M: the gain is in task accuracy, and we report the coverage cost alongside it. The median split leaves each quadrant about a quarter of its pool, where this sweep keeps a tenth; that difference in dose is what separates them.

E RELATED WORK ON REPRESENTATION GEOMETRY

Geometry supplied by a representation. Model-side theories derive exponents after a data manifold or spectrum has been specified (Sharma & Kaplan, 2022; Bahri et al., 2024; Maloney et al., 2022;

Michaud et al., 2023), and nonlinear dimension-reduction, finite-sample testing and empirical probes all require a chosen representation and metric (Tenenbaum et al., 2000; Roweis & Saul, 2000; Belkin & Niyogi, 2003; Narayanan & Mitter, 2010; Fefferman et al., 2016; Bengio et al., 2013; Pope et al., 2021; Ansuini et al., 2019; Du & Tanaka-Ishii, 2024). Each of them starts from a specified representation and metric; Theorem G.7 says why our symbol-only measurement cannot recover source dimension in general. Our approximation result applies manifold rates (Shaham et al., 2018; Chen et al., 2019; Nakada & Imaizumi, 2020; Schmidt-Hieber, 2020) to a supplied predictive state (Proposition G.4).

F 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. We fix the distinctions here before stating the results.

F.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 $\mathcal{M}_{\mathrm{geo}}$. 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 $\mathcal{M}_{\mathrm{geo}}$; in the smooth case, $\mathcal{N}_{\mathrm{cov}}(\varepsilon; \mathcal{M}_{\mathrm{geo}}) = \Theta(\varepsilon^{-d_{\mathrm{geo}}})$.
$d_\mu(x)$	Local mass-scaling exponent of the source measure. Under Assumption F.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 $d_{\mathrm{geo}} = h_X/\lambda$ (Proposition G.6). An additional Hölder sufficient factor relates source geometry to the predictive state, whose direct-head rate obeys $\alpha_{\mathrm{head}} \geq 2/d_{\mathrm{pred}}$. The empirical finite-window slope $\widehat{\alpha}_{\mathrm{LZ}}$ is not any of these dimensions.

F.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 G.1); a finite-context model plateaus above it. Ideal universal coding identifies h_X asymptotically. The deployed $\widehat{h}$ 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 $\alpha_{\mathrm{head}} \geq 2/d_{\mathrm{pred}}$ (Proposition G.4). Against

training data, long memory and estimator choice matter; on the tested panels, $\widehat{\alpha}_{\mathrm{LZ}}$ 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 $\widehat{\alpha}_{\mathrm{LZ}}$ estimates a different quantity (Propositions G.6 and G.8 and Theorem G.7).

F.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 $X_{\le t}$ is well defined; Assumption F.3 adds ergodicity and fixes the entropy rate.

Source-side assumptions.

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

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

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 K.6. Write ρ_{geo} for the Riemannian distance on $\mathcal{M}_{\mathrm{geo}}$.

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

Assumption F.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 $h_X = \lim_{m\to\infty} H(X_0 \mid X_{-m}^{-1})$. This assumption supports the ideal universal-LZ and symbolic return-time limits; it imposes no smooth geometry.

Assumption F.4 (Conformal expanding bridge [**Cond. D**])**.** In addition to Assumptions F.1 and F.2, the symbols arise from a finite generating partition of finite entropy, and $(\mathcal{M}_{\mathrm{geo}}, T, \mu)$ is a $C^{1+\beta_{\mathrm{dyn}}}$ conformal expanding map for some $\beta_{\mathrm{dyn}} > 0$ with absolutely continuous SRB/ACIP measure. The process (X_t) of Assumption F.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 F.3 applies verbatim to that coding. Thus D_xT is a scalar multiple of an isometry, $|\det D_xT| = \|D_xT\|^{d_{\mathrm{geo}}}$, $\|D_xT\| > 1$, and the single positive Lyapunov exponent is

$$\lambda = \int_{\mathcal{M}_{\mathrm{geo}}} \log_2 \|D_xT\|\, d\mu > 0.$$

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

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

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

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 G.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 F.6 (Ergodicity versus mixing). The entropy, return-time, and ideal universal-LZ limits require stationarity and ergodicity, not a mixing rate. Fast mixing is required only when the recurrence-dimension lower bound is upgraded to a two-sided limit (Saussol et al., 2002; Galatolo, 2007); Assumption F.4 supplies that rate.

Predictive-side assumptions.

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

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

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

$$\nu \ll \mathcal{H}^{d_{\mathrm{pred}}}|_{\mathcal{M}_{\mathrm{pred}}}, \qquad 0 < c_1^{\mathrm{pred}} \le \frac{d\nu}{d\mathcal{H}^{d_{\mathrm{pred}}}} \le c_2^{\mathrm{pred}} < \infty.$$

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 $2/d_{\mathrm{pred}}$.

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 F.8 (Prediction regularity)**.** On $\mathcal{M}_{\mathrm{pred}}$, 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 $s \in \mathcal{M}_{\mathrm{pred}}$ 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 $\mathcal{M}_{\mathrm{pred}}$ it also implies a finite bound $\|\Lambda(s)\|_\infty \le B$ and hence $p_i(s) \ge e^{-2B}/|\Sigma|$.

Remark F.9 (Probability floor and the KL constant). Compactness and Lipschitz conditional logits imply a positive probability floor: $\|\Lambda(s)\|_\infty \le B$ gives $p_i(s) \ge e^{-2B}/|\Sigma|$. 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 F.8's scope.

Connecting source and predictive states.

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

$$\rho_{\mathrm{pred}}(\Phi(z), \Phi(z')) \le K \rho_{\mathrm{geo}}(z, z')^\gamma, \qquad 0 < \gamma \le 1.$$

Then

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

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 $\dim_{\mathrm{H}} \Phi(\mathcal{M}_{\mathrm{geo}}) \le \dim_{\mathrm{H}}(\mathcal{M}_{\mathrm{geo}})/\gamma$. See Section O.1.

Remark F.11 (Geometric and predictive states). $\mathcal{M}_{\mathrm{geo}}$ supports the dynamical system in Assumptions F.2 and F.4; $\mathcal{M}_{\mathrm{pred}}$ supports the sufficient statistic used for prediction. Only Lemma F.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 H.1 additionally requires a positive-dimensional smooth sufficient state satisfying Assumptions F.7 and F.8, the conformal source bridge in Assumption F.4, and the Hölder map above. No nontrivial source with independent past and future satisfies all of them at once, for the reason Proposition H.2 proves.

F.4 Requirements and Notation

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

We reserve N 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 N and D for the model and data axes and write α_D for the data-scaling exponent.

G 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.

G.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 G.1 (Symbolic LZ entropy limit (Wyner & Ziv, 1989; Ornstein & Weiss, 1993; Wyner & Ziv, 1994); [**Theorem**]). *Under Assumption F.3, let $C_{\mathrm{WZ}}(n; w)$ 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*

$$h_w := \limsup_{n\to\infty} \mathbb{E}C_{\mathrm{WZ}}(n; w),$$

the ideal universal-LZ floor satisfies

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

The order of limits is fixed: a diagonal $w = w(n)$ and any finite realized codelength lie outside this statement. If $h_X > 0$, the associated LZ match length L_n also obeys $L_n/\log_2 n \to 1/h_X$ 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 $w \to \infty$. The match-length conclusion is the Ornstein–Weiss/Wyner–Ziv recurrence theorem quoted in Lemma K.2. The limit-order and initial-database details are proved in Section O.1.

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

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

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

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 $\widehat{\alpha}_{\mathrm{LZ}}$'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 O.1.

We obtain finite-sample coverage from the certificate below instead, 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 G.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 $L_1, \ldots, L_{J_{\mathrm{doc}}}$ drawn from one common stationary source on Σ, whose entropy rate is h_X. Insert a fixed string $\sigma_{\mathrm{sep}} \in \Sigma^{b_{\mathrm{sep}}}$ at the positions determined by those known lengths; its positions, rather than its symbol values, identify it. Let the resulting total length be $n_{\mathrm{tot}} \geq b$. Sample M offsets independently and uniformly from the legal window starts $\{0, \ldots, n_{\mathrm{tot}} - b\}$. Suppose each window contains at most $s_{\max} < b$ 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 O.1.1. Let $Z_1, \ldots, Z_M$ 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 $q_b = \lceil \log_2(8L_{\max} b + 1) \rceil$, define

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

Then, jointly over document contents and sampled offsets,

$$\mathbb{P}\big[h_X \leq U_h(b, M) := \kappa\big(\bar{Z} + \varepsilon_M + \varepsilon_{\mathrm{doc}}\big)\big] \geq 1 - \delta, \tag{12}$$

with h_X and $U_h(b, M)$ in the same source-symbol unit. The byte identity map gives a bits-per-byte certificate ($L_{\max} = 1$); UTF-8 gives a bits-per-character certificate ($L_{\max} \leq 4$). 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 (12). The exact wrapper and proof are in Section O.1.1.

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

G.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 G.4 (Cross-entropy achievability exponent). *Let Assumptions F.7 and F.8 hold on a predictive-state manifold $\mathcal{M}_{\mathrm{pred}}$ of dimension $d_{\mathrm{pred}} > 0$ with state law ν. For the N-parameter deep-ReLU logit class of Lemma I.3, define*

$$\Delta\mathcal{L}(N) := \inf_{f \in \mathcal{F}_N} \mathbb{E}_{s \sim \nu} \mathrm{KL}_2(p(\cdot \mid s) \,\|\, \mathrm{softmax}(f(s))),$$
$$\alpha_{\mathrm{head}} := \sup\{a : \Delta\mathcal{L}(N) = O(N^{-a})\}. \tag{13}$$

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

$$\mathbb{E}_{s \sim \nu} \mathrm{KL}_2(p(\cdot \mid s) \,\|\, \mathrm{softmax}(f_N(s))) = O_\eta\Big(N^{-2/(d_{\mathrm{pred}} + \eta)}\Big).$$

Consequently $\Delta\mathcal{L}(N)$ obeys the same bound and $\alpha_{\mathrm{head}} \geq 2/d_{\mathrm{pred}}$. 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 $O_\eta(N^{-1/(d_{\mathrm{pred}} + \eta)})$. 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 $\alpha_{\mathrm{head}} \geq 2/d_{\mathrm{pred}}$. See Section O.1 for the endpoint and constant bookkeeping.

Lemma G.5 (Predictive states attain the entropy floor)**.** *For the stationary process of Assumption F.3, if S_t is a measurable function of $X_{\le t}$ and is predictively sufficient, then*

$$\mathbb{E}H\big(p(\cdot \mid S_t)\big) = H(X_{t+1} \mid X_{\le t}) = h_X.$$

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. We record the two-line formal argument in Section O.1.

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

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

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

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 K.7.

Proof. The coding partition of Assumption F.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 $h_\mu(T) = \int_{\mathcal{M}_{\text{geo}}} \log_2 |\det D_x T| \, \mathrm{d}\mu(x)$, 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 K.4 records the same calculation and extends it to the pointwise dimension. □

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 F.5.

G.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 G.7 (Symbolic obstruction to recovering dimension)**.** *Geometric dimension is not a function of the symbolic process law. There exist two systems satisfying Assumptions F.1, F.2 and F.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 O.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 G.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*

$$\mathbb{E}C_{\text{LZ78}}(n) - h_X = O(1/\log_2 n) \qquad \textit{(Savari, 1997).} \tag{14}$$

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

$$\frac{A_{\text{LZ}} + \delta(n)}{\log_2 n} + O\left(\frac{\log_2 \log_2 n}{(\log_2 n)^2}\right), \tag{15}$$

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 G.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 O.1.

Remark G.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 G.4; α_D is a trained model's fitted data-scaling exponent, which can depend on memory and spectral structure; and $\widehat{\alpha}_{\text{LZ}}$ 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 H.1, while the $\widehat{\alpha}_{\text{LZ}}$-to-$\alpha_D$ link is only a calibration. These are the only established pairings, and Theorem G.7 rules out interpreting $\widehat{\alpha}_{\text{LZ}}$ universally as geometric $1/d_{\text{geo}}$.

G.4 Entropy Does Not Determine the Data Exponent

Theorem G.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 G.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*

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

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

$$H\big(X_1^{(a)}, \ldots, X_n^{(a)}\big) = n + \min(n, k). \tag{16}$$

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 Hilberg-type exponent.

Proof. Write $u_j := U_{j-k}$ for $j = 1, \ldots, k$: 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 $U_1, \ldots, U_n$, so $H(U_1^n) = n$.

Step 2: only the first r second coordinates stay random once U_1^n is known. The window at time t is $(U_{t-k}, \ldots, U_{t-1})$, 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 $1 \leq t \leq r$,

$$X_{t,2}^{(a)} = \underbrace{\bigoplus_{i\leq k+1-t} a_i\, u_{t+i-1}}_{\langle v_t, u\rangle} \oplus \underbrace{\bigoplus_{i>k+1-t} a_i\, U_{t-k+i-1}}_{\text{measurable w.r.t. } U_1^n}. \tag{17}$$

Step 3: the rows $v_1, \ldots, v_r$ are independent over $\mathrm{GF}(2)$. The row v_t is supported on coordinates $t, \ldots, k$, and its entry at t is $a_1 = 1$. Stacking $t = 1, \ldots, r$ 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 (17) 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 $H(X_1^{(a)}, \ldots, X_n^{(a)}) = n + r$, which is (16). □

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 partial knowledge of the window lowers the loss, not in how much information the next symbol 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 (16) adds is exact entropy matching, which turns the observation into a statement about what a training-free predictor cannot do.

The deployed statistics cannot separate the family 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, $\widehat{\alpha}_{\mathrm{LZ}}$ spans 0.1979 to 0.1984: a cross-arm spread of 0.0005 against a start-index noise of 0.0009, so the spread is 0.6 of the instrument's own noise floor. The three-point floors agree across arms to 0.001. The frozen map therefore predicts the same exponent, 0.188, for every arm. Section R records the construction, the five measured arms and the locked per-arm predictions, under a decision rule requiring the trained spread to exceed three times the largest leave-one-cell-out refit standard deviation.

G.5 The Character Clock of the Deployed Instrument

We normalize the symbolic results 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 G.11 (Character clock of a UTF-8 encoded source)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ satisfy Assumption F.3 on a finite character alphabet, with h_X in bits per character. Let ϕ be the prefix-free UTF-8 map, write $\ell_{\mathrm{utf}}(a) = |\phi(a)| \in \{1, 2, 3, 4\}$, $\bar{\ell}_{\mathrm{utf}} = \mathbb{E}\ell_{\mathrm{utf}}(X_0)$, and let $K_n = \sum_{t\le n} \ell_{\mathrm{utf}}(X_t)$ be the byte length of $\phi(X_1^n)$. Then:*

(i) *$K_n/n \to \bar{\ell}_{\mathrm{utf}}$ almost surely and in L^1.*

(ii) *For every lossless binary encoder L of byte strings whose range is prefix-free,*

$$\frac{\mathbb{E}\big|\mathsf{L}(\phi(X_1^n))\big|}{n} \;\ge\; \frac{H(X_1^n)}{n} \;\ge\; h_X \qquad (n \ge 1).$$

(iii) *For prefixes anchored at character boundaries,*

$$-\frac{1}{K_n}\log_2 \mathbb{P}(X_1^n) \;\longrightarrow\; \frac{h_X}{\bar{\ell}_{\mathrm{utf}}} \qquad \text{almost surely.}$$

Proof sketch. Birkhoff controls K_n/n, Kraft gives the expected-length inequality, and Shannon–McMillan–Breiman combined with $n/K_n \to 1/\bar{\ell}_{\mathrm{utf}}$ gives the anchored byte rate. See Section O.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.

H 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, independently of the other rows.

The final two rows of Table 7 combine a geometric comparison with a neural approximation result. A γ-Hölder map can increase Hausdorff dimension by at most a factor $1/\gamma$ (Falconer, 2003); if its image is a sufficient predictive state, a direct prediction head inherits the corresponding dimension-dependent rate.

Table 7: **Available conclusions.** Each row lists its required information and the resulting statement. The last two rows are conditional implications, and Proposition H.2 proves that an expanding source with independent past and future supplies no such Hölder factor, so an instance has to come from elsewhere.

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

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

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

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

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

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 $\widehat{\alpha}_{\mathrm{LZ}} \to \alpha_D$ calibration rests on trained grids, not on this chain.*

Proof sketch. Three links yield the chain: the conformal bridge gives $d_{\mathrm{geo}} = h_X/\lambda$, the Hölder factor gives $d_{\mathrm{pred}} \le d_{\mathrm{geo}}/\gamma$, and the direct-state theorem gives $\alpha_{\mathrm{head}} \ge 2/d_{\mathrm{pred}}$. 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 require the state to be measurable with respect to opposite halves of the process, and that is a theorem rather than a gap in our search.

Proposition H.2 (Expanding sources cannot supply the factor). *Let $(\mathcal{M}_{\mathrm{geo}}, T, \mu)$ satisfy Assumption F.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 $\sigma(X_{t+1}, X_{t+2}, \ldots)$ up to a μ-null set. Consequently a state S_t that is past measurable as Assumption F.7 requires and also satisfies $S_t = \Phi(Z_t)$ almost surely as Lemma F.10 requires is measurable with respect to $\sigma(X_{\le t}) \cap \sigma(X_{>t})$. If in addition 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 $d_{\mathrm{pred}} = 0$. Such a source admits no Hölder factor onto a predictive state of positive dimension, which is what the last two rows of Table 7 would need. Without independence the same argument still confines S_t to $\sigma(X_{\le t}) \cap \sigma(X_{>t})$, so those two rows need a source whose intersection σ-algebra carries positive dimension.*

Proof. The source point lies in the future. A generating partition determines Z_t from its forward itinerary $(X_{t+1}, X_{t+2}, \ldots)$ 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 $Z_t \in \sigma(X_{>t})$ mod μ, 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 F.7 excludes. □

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 a past-measurable state. The two halves of Theorem H.1 therefore cannot be joined by a source of the kind Assumption F.4 describes whose past and future are independent. Joining them means replacing 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 F.11 records the other direction, where a filter has a positive-dimensional predictive law on a generally Cantor support.

Corollary H.3 (Finite-sample propagation)**.** *Assume the conformal-source hypotheses of Proposition G.6, and suppose two procedures have the marginal guarantees*

$$\mathbb{P}(h_X \leq U_h) \geq 1 - \delta_h, \qquad \mathbb{P}(\lambda \geq \lambda_{\text{L}}) \geq 1 - \delta_\lambda,$$

where $U_h, \lambda_{\text{L}} > 0$ and $\delta_h + \delta_\lambda < 1$. Then, with probability at least $1 - \delta_h - \delta_\lambda$,

$$d_{\text{geo}} \leq \frac{U_h}{\lambda_{\text{L}}}.$$

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

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

and if the direct-head assumptions Assumptions F.7 and F.8 also hold, then

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

No independence is required. For known exact λ, set $\lambda_{\text{L}} = \lambda$ 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 F.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 empirical scores and the certificate require different inputs. The instrument of Section 2 returns $(\widehat{h}, \widehat{\alpha}_{\text{LZ}})$ 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 G.3 gives $h_X \leq U_h$, and the geometric rows of Table 7 yield the remaining conclusions once their dynamical inputs are in hand.

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 G.4 and G.6, Lemma J.1, and Theorem G.7. The finite-n coding result in Proposition G.8 fixes a separate estimator scope. This appendix develops the supporting lemmas and scope, with longer proofs in Section O.

I Predictive Dimension Controls Direct-Head Approximation

Assumptions F.7 and F.8 provide the predictive manifold and its Lipschitz logit map. Assumption F.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 requires $\Theta(\varepsilon^{-d_{\text{pred}}})$ balls (Lemma I.2); a ReLU network with N parameters can approximate the logit map to sup-norm error $O_\eta(N^{-1/(d_{\text{pred}}+\eta)})$ for every $\eta > 0$ (Lemma I.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 I.4 and I.5).

Remark I.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 F.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 $O_\eta(N^{-\beta_{\text{app}}/(d_{\text{pred}}+\eta)})$ 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 I.2 (Kolmogorov-Tihomirov Metric Entropy)**.** *Let $\mathcal{M}$ be a compact d-dimensional C^1 embedded Riemannian manifold. Its minimum ε-cover size satisfies*

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

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

$$c_- r^d \leq \mathcal{H}^d\big(B(x, r) \cap \mathcal{M}\big) \leq c_+ r^d. \tag{18}$$

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 (18) and subadditivity of $\mathcal{H}^d$ give

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

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 $\{x_1, \ldots, x_J\} \subset \mathcal{M}$. Maximality makes the radius-ε balls about these points a cover, whereas the radius-$\varepsilon/3$ balls are pairwise disjoint. The lower bound in (18) therefore yields

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

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

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

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

thus N is the total, rather than per-coordinate, budget. Under Assumptions F.7 and F.8:

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

$$\sup_{s \in \mathcal{M}_{\text{pred}}} \|f_N(s) - \Lambda(s)\|_\infty = O_\eta\Big(N^{-1/(d_{\text{pred}}+\eta)}\Big)$$

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 $N = \Omega(\varepsilon^{-d_{\text{pred}}/(2\beta_{\text{app}})})$;*

(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* $N = \Omega(\varepsilon^{-d_{\text{pred}}/\beta_{\text{app}}} \log^{-2p_{\text{dep}}-1}(1/\varepsilon))$*. 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 I.4 originates. The full argument is in Section O.

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

Remark I.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 I.3(b1) (Yarotsky 2017, Thm 4a) gives $N = \Omega(\varepsilon^{-d_{\text{pred}}/2})$ 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:

$$\mathrm{KL}_2 \;\leq\; \frac{\varepsilon^2}{2\ln 2} \qquad \Longrightarrow \qquad \alpha_{\text{head}} \;\geq\; \frac{2}{d_{\text{pred}}} \quad (\textit{Proposition G.4}).$$

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 (β_{app}=2 vs. β_{app}=1, Remark I.1). The quadratic loss accounts for the common multiplier (Bahri et al., 2024).

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

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

Remark I.5 (Scope of the proportionality: a sharp-spectrum hypothesis beyond the exact-manifold model). The stronger claim that d_{pred} alone fixes a trained-run exponent through $\alpha_{\text{head}} \propto 1/d_{\text{pred}}$ is not a corollary (Remark I.4). On the exact compact $C^{k_{\text{pred}}}$ manifold of Assumption F.7, upper Minkowski, Hausdorff, and intrinsic dimension coincide at d_{pred} (proof of Lemma I.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 G.4. Order the local covariance or kernel eigenvalues as $\zeta_1 \geq \zeta_2 \geq \cdots$. 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 M illustrates only this last possibility: within the tested construction, two spectral shapes matched on d_{PR} produce a 47% spread in $\widehat{\alpha}_{1\mathrm{NN}}$. 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.

J Local Dimension of the Source Measure

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

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

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

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

Proof. Let $f = d\mu/d(\mathcal{H}^{d_{\mathrm{geo}}}|_{\mathcal{M}_{\mathrm{geo}}})$. By Assumption F.1, $c_1 \le f \le c_2$ almost everywhere. Combining this with the uniform small-ball estimate (18), for every $x \in \mathcal{M}_{\mathrm{geo}}$ and every sufficiently small r we have

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

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

K From Symbolic Recurrence to Geometric Dimension

Assumption F.3 supplies the symbolic entropy limit; Assumptions F.1, F.2 and F.4 supply the conformal geometric bridge. Results marked **[Cond. D]** require the smooth conformal-expanding structure of Assumption F.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 $d_{\mathrm{geo}} = h_X/\lambda$.

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 $\lim_{w\to\infty} h_w$; finite-n compression rates are treated separately in Remark K.5.

Lemma K.1 (Ornstein-Weiss Theorem (Ornstein & Weiss, 1993)). *Under Assumption F.3, the return-time complexity satisfies*

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

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 F.3. Moreover, if $M_n(x)$ is the LZ78 phrase count, then $M_n(x)\log_2 M_n(x)/n \to h_X$ a.s. (Ziv & Lempel, 1978).

Proof. Apply the Ornstein–Weiss recurrence theorem to the finite-alphabet stationary ergodic shift specified by Assumption F.3. Its cylinder return time is precisely the random variable R_ℓ in the statement, so the theorem gives $\ell^{-1}\log_2 R_\ell \to h_\mu(\sigma) = h_X$ 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. □

The next lemma records the equivalent match-length formulation of symbolic recurrence; the expected-rate limit and its order of limits are in Theorem G.1.

Lemma K.2 (Match-length/recurrence-time duality (Wyner & Ziv, 1989)). *Under Assumption F.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*

$$\frac{\log_2 n}{L_n(x)} \xrightarrow{\mathbb{P}} h_X, \qquad \frac{1}{\ell}\log_2 R_\ell(x) \xrightarrow{a.s.} h_X,$$

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 $\log_2 n/L_n \to h_X$ 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 $L_n/\log_2 n \to 1/h_X$ in probability. No reciprocal is taken in the zero-entropy case. □

Lemma K.3 (Recurrence-Dimension Identity (Barreira & Saussol, 2001) [**Cond. D**]). *Under Assumption F.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,*

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

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:

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

for μ-almost every x. Under Assumption F.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 J.1, $\underline{d}_\mu(x) = \overline{d}_\mu(x) = d_{\text{geo}}$ almost everywhere, so the displayed liminf and limsup coincide and equal d_{geo}. □

Lemma K.4 (Pesin Entropy Equality, Conformal ACIP Case (Rokhlin, 1967; Pesin, 1977; Young, 1982; Ledrappier & Young, 1985) [**Cond. D**]). *Under Assumption F.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:*

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

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 F.4.

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

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

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

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

The partition in Assumption F.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 J.1 yields $d_\mu(x) = h_X/\lambda = d_{\text{geo}}$ for μ-almost every x. □

Proposition G.6 ($d_{\text{geo}} = h_X/\lambda$, conformal case [**Cond. D**]) is proved in the body (Section G) 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 K.5 (What LZ Measures: The Entropy Side of the Bridge). Proposition G.6 uses the population identity $h_X = \lim_{w\to\infty} h_w$ 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 G.3. The non-identification result in Section L shows why a symbolic statistic alone cannot supply the missing geometric scale uniformly over its stated class.

Remark K.6 (General Gibbs / information-dimension extension). The absolute-continuity hypothesis of Assumption F.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 $h_X \le d_{\text{geo}}\lambda$. 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 still holds, but now gives the *measure / information dimension* rather than the manifold dimension:

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

by the Ledrappier–Young formula in the conformal case (Ledrappier & Young, 1985; Young, 1982). Here $\dim_{\text{H}}(\mu) = h_X/\lambda \le d_{\text{geo}}$, with equality iff μ is the ACIP of Assumption F.4. Within this conformal Gibbs class the population Rényi dimensions satisfy $D_2 \le D_1$, and the ratio equals the information dimension D_1, not an embedding dimension. A finite-sample empirical estimate $\widehat{D}_2$, 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 G.6 remains the absolutely continuous special case.

Remark K.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 $\lambda^{(1)} > \lambda^{(2)} > \cdots$ 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,

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

with $\lambda^{(i)}$ the distinct (unstable) exponents and $H_i - H_{i-1} \ge 0$ the entropy increments subordinate to the i-th invariant foliation (so $\sum_i (H_i - H_{i-1}) = h_X$, $0 \le \vartheta_i \le \dim E^{(i)}$). Conformality (Assumption F.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 G.6 is a genuinely *conformal* conversion of h_X into a dimension; non-conformal geometry contains strictly more information than this ratio captures.

L Why Symbols Do Not Identify Geometric Dimension

Theorem G.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 G.8 is a separate coding result; neither statement turns a symbolic residual into a geometric observable.

Theorem L.1 (Quantitative obstruction: an irreducible minimax error)**.** *Fix integers $q, k \ge 2$ and take the corresponding witnesses W_1, W_2 constructed in the proof of Theorem G.7, with dimensions*

1 *and* k*, respectively. For* every $n \geq 1$ *and every measurable estimator* $\widehat{d}_n : \{0, \ldots, q^k - 1\}^n \to \mathbb{R}$ *that reads the token stream alone,*

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

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

Proof. By Theorem G.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 $\widehat{d}_n(X_{1:n})$ has a single law $\mathbb{Q}_n$ on $\mathbb{R}$ and

$$\mathbb{E}_{W_i}|\widehat{d}_n - d_i| = \int |t - d_i|\,\mathrm{d}\mathbb{Q}_n(t), \tag{19}$$

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 (19) gives

$$\max_{i\in\{1,2\}} \mathbb{E}_{W_i}|\widehat{d}_n - d_i| \ \geq\ \tfrac{1}{2}\int \big(|t-1| + |t-k|\big)\,\mathrm{d}\mathbb{Q}_n(t) \ \geq\ (k-1)/2, \tag{20}$$

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. □

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 L.2 (Robust obstruction and a sample-size floor)**.** *Let* W_1, W_2 *satisfy Assumptions F.1, F.2 and F.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* $V_n = \mathrm{TV}(\mathbb{P}_1^{(n)}, \mathbb{P}_2^{(n)})$*. For every* $n \geq 1$ *and every (possibly randomized) estimator* $\widehat{d}_n$ *that reads the token stream alone,*

$$\max_{i\in\{1,2\}} \mathbb{E}_{W_i}\big|\widehat{d}_n - d_i\big| \ \geq\ \frac{\Delta}{2}\big(1 - V_n\big). \tag{21}$$

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

$$\max_{i\in\{1,2\}} \mathbb{E}_{W_i}\big|\widehat{d}_n - d_i\big| \ \geq\ \frac{\Delta}{2}\Big(1 - \sqrt{n\delta/2}\Big), \tag{22}$$

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*

$$n \ >\ \frac{1}{2\delta} \tag{23}$$

symbols. This necessary condition is independent of the estimator. Setting $\delta = 0$ *returns Theorem L.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 $t \in \mathbb{R}$,

$$\mathbb{E}_{W_1}\big|\widehat{d}_n - d_1\big| + \mathbb{E}_{W_2}\big|\widehat{d}_n - d_2\big| \ \geq\ \int \Big(\big|\widehat{d}_n - d_1\big| + \big|\widehat{d}_n - d_2\big|\Big)\,\mathrm{d}\xi_n \ \geq\ \Delta\,(1 - V_n),$$

and a maximum dominates half the sum, giving (21); for randomized $\widehat{d}_n$ apply the display conditionally on the randomization seed and average. Pinsker's inequality gives $V_n \leq \sqrt{\mathrm{KL}_{\mathrm{nat}}(\mathbb{P}_1^{(n)} \| \mathbb{P}_2^{(n)})/2} \leq \sqrt{n\delta/2}$, which is (22); and $\sqrt{n\delta/2} \leq \frac{1}{2}$ exactly when $n \leq 1/(2\delta)$. □

Two consequences follow. First, the witnesses of Theorem L.1 have $\delta = 0$, where (23) is vacuous and the floor $\Delta/2$ holds at every n: that is the impossibility. Second, (22) 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 consequence licenses a dimension estimate from symbols; together they show the boundary is a neighborhood rather than a single point.

Remark L.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 Q.4 gives their definitions, domain conditions, and full algebra.

The coding results do not validate this template for the deployed inputs C_i: Proposition G.8 is code- and source-class-specific. Thus $\widehat{\alpha}_{\mathrm{s}}^{(3)}$ 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 K.5.

Does a three-point floor lie above or below h_X? 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

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

and therefore

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

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 $\widehat{H}_0^{(3)} \geq h_X$; when $\Delta_a < 0$ the implication is reversed, and for residuals of mixed sign log-convexity is not defined. No theorem here establishes those hypotheses for the deployed curve. Whether the three-point floor lies above or below h_X therefore remains open. In a PPM screening sweep (Section T.1) the raw extrapolant for `opc_synth` is negative and the clamped $\widehat{h}$ is exactly 0. The clamp never fired on a reported number: no floor in any deployed panel is at or below 0, the smallest being 0.169.

M Controlled Checks of the Theory's Components

On problems with known answers the components behave as the theory predicts: switching long-range structure on in a source whose conditional-entropy tail exponent is known by construction raises $\widehat{\alpha}_{\mathrm{LZ}}$ by six to thirty times its noise floor, the finite-grid residual estimator tracks $1/d$ on four manifold families, 1-NN covering rates agree with $2/d$ to within 1%, 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 G.4: they test the covering-rate mechanism, not end-to-end Transformer scaling. Supporting figures and measurement protocols for the corpus statistics are in Section Q; the Case-1 training record is in Section B.

M.1 Controlled Measurement Checks with Known Targets

On four controlled manifold families the finite-grid residual estimator follows $\widehat{\alpha}_{\mathrm{s}}^{(3)} \approx 1/d$, and the floor-side estimates line up with the known entropy rates. That agreement is an empirical calibration on this panel; the coverage guarantee comes instead from the U_h procedure of Proposition G.3.

M.2 The Measurement on a Source With a Known Exponent

Switching long-range structure on, with the string held fixed, raises $\widehat{\alpha}_{\mathrm{LZ}}$ 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 $\gamma_{\mathrm{CE}} = 1 - \beta$, known by construction.

We pair the comparison. 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 $\gamma_{\mathrm{CE}} = 0.70, 0.50, 0.30, 0.10$ the measured $\widehat{\alpha}_{\mathrm{LZ}}$ 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 measurement at five start-index seeds. The sign is the same in all four, and the effect grows with the known exponent.

Long-range structure accounts for 4% to 19% of the value. Without reuse the same sources still return $\widehat{\alpha}_{\mathrm{LZ}}$ between 0.129 and 0.250, from the Zipf marginal and the rendering alone. An absolute $\widehat{\alpha}_{\mathrm{LZ}}$ 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. Section S gives the full seven-source sweep.

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

Noiseless 1-NN regression on $\mathrm{Unif}([0,1]^d)$ recovers the expected covering rate to within 1%: $\widehat{\alpha}_{1\mathrm{NN}} d = 2.013 \pm 0.018$ for $d = 1, \ldots, 8$. On the constant-scale conformal $\times 2$ torus we measure the two sides 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 $\widehat{\alpha}_{1\mathrm{NN}} \hat{d} = 2.03 \pm 0.02$. Away from sharp spectra the scalar summary fails: three shapes at matched $d_{\mathrm{PR}} = 4$ spread $\widehat{\alpha}_{1\mathrm{NN}}$ by 47%, ruling out that scalar as a sufficient description of the tested learner.

N Notation

Symbol	Where	Meaning
$\mathcal{M}_{\mathrm{geo}}$, d_{geo}, $d_{\mathrm{geo,amb}}$	Assumption F.1	geometric source manifold, Hausdorff dimension, and ambient dimension
$\mathcal{M}_{\mathrm{pred}}$, ν, d_{pred}, $d_{\mathrm{pred,amb}}$	Assumption F.7	predictive-state manifold, its law, representation-relative dimension, and ambient dimension
μ, T	Assumption F.2	invariant data measure; generating dynamics
$\mathbb{P}_X$	Assumption F.3	law of the stationary symbolic source
h_X, $h_\mu(T)$	Assumptions F.3 and F.4	symbolic entropy rate; measure-theoretic entropy of the geometric dynamics, equal under the generating code
h_m, p_m, q_m	Proposition G.2	optimal loss using the last m symbols; true conditional law given those symbols; a predictor using that same context
$G_{\mathrm{comp}}(n)$	Section B.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 (2)	functional notation for the finite-range curve template; not an unspecified asymptotic
diagonal $w = w(n)$		

Symbol	Where	Meaning
α_{s}, A_{H}	Equation (2)	structural exponent and amplitude of the Hilberg template; fit parameters, not source invariants (Remark L.3)
$\widehat{h}$, $\widehat{\alpha}_{\mathrm{LZ}}$	Section 2 and Remark G.9	empirical floor and rate-slope rankers; neither denotes a source invariant
$\widehat{H}_0^{(3)}$, $\widehat{\alpha}_{\mathrm{s}}^{(3)}$	Section Q.4	explanatory estimators from three rates on the grid n, gn, g^2n; $\widehat{h}$ is the first of these evaluated at $128, 2048, 32768$ characters, while the deployed slope $\widehat{\alpha}_{\mathrm{LZ}}$ is a separate eight-length regression (Remark L.3)
$\alpha_{1\mathrm{NN}}$	Section M	exponent of the nearest-neighbour learning curve in the fixed-participation-ratio construction
γ_{CE}	Section A	tail exponent of the conditional entropy, $H_n^{\mathrm{cond}} - h_X \asymp n^{-\gamma_{\mathrm{CE}}}$
ρ	Section M	Spearman rank correlation; the subscripted $\rho_{\mathrm{geo}}, \rho_{\mathrm{pred}}$ below are metrics, not correlations
U_h, κ, $\bar{Z}$	Equation (12)	one-sided finite-sample upper bound on h_X, its window-overhead factor $b/(b - s_{\max})$, and the mean per-symbol codelength it is built from
η	Proposition G.4	arbitrary positive slack in the approximation-rate exponent; the training learning rate is written lr
α_D, α_N, $\alpha_{D,N}$	Remark G.9	fitted data-axis exponent, fitted model-axis exponent, and data-axis exponent refit at one model size N
α_{head}	Proposition G.4	best-in-class cross-entropy exponent, $\sup\{a : \Delta\mathcal{L}(N) = O(N^{-a})\}$
λ	Assumption F.4	conformal Lyapunov exponent (geometric side-input)
d_{PR}	Section M	participation ratio $(\sum_i \zeta_i)^2 / \sum_i \zeta_i^2$ of a spectrum; a finite-resolution effective dimension, distinct from the exact $\dim_{\mathrm{H}}(\mathcal{M}_{\mathrm{pred}})$ of Assumption F.7
β_{app}, β_{dyn}, γ	Remark I.1, Assumption F.4, and Lemma F.10	target smoothness, dynamical smoothness margin, and Hölder exponent of the source-to-state factor
N, D, n, n_{tot}	Section F.1	parameter count; training-token budget; text length; total certificate-corpus length

Symbol	Where	Meaning
w, m, ℓ, b	Section F.1	LZ window; predictive context length; recurrence block length; certificate window length in source symbols
$\ell_{\mathrm{utf}}, \bar{\ell}_{\mathrm{utf}}$	Lemma G.11	UTF-8 codeword length in bytes and its mean; the bytes-per-character factor relating the byte and character clocks
$\rho_{\mathrm{geo}}, \rho_{\mathrm{pred}}, r, g$	Section F.1 and Lemma F.10	source and predictive metrics; metric radius; three-point grid ratio n_{i+1}/n_i

O 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.

O.1 Proofs Deferred from the Main Results

Proof of Lemma F.10. Set $s = \dim_{\mathrm{H}} \mathcal{M}_{\mathrm{geo}}$ and fix $q > s/\gamma$. Then $\gamma q > s$, so $\mathcal{H}^{\gamma q}(\mathcal{M}_{\mathrm{geo}}) = 0$. For every $\delta > 0$, let $\{U_j\}_{j\ge1}$ be a countable relative cover of $\mathcal{M}_{\mathrm{geo}}$, with $U_j \subseteq \mathcal{M}_{\mathrm{geo}}$ and $\operatorname{diam}(U_j) \le \delta$. The Hölder condition implies

$$\operatorname{diam}\big(\Phi(U_j)\big) \le K \operatorname{diam}(U_j)^{\gamma},$$

and $\{\Phi(U_j)\}_{j\ge1}$ covers $\Phi(\mathcal{M}_{\mathrm{geo}})$. Consequently,

$$\mathcal{H}^{q}_{K\delta^{\gamma}}\big(\Phi(\mathcal{M}_{\mathrm{geo}})\big) \le K^{q}\mathcal{H}^{\gamma q}_{\delta}(\mathcal{M}_{\mathrm{geo}}).$$

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

Proof of Theorem G.1. Let $\mathsf{L}^{\mathrm{WZ}}_{n,w} = nC_{\mathrm{WZ}}(n;w)$ be the codelength. If the initial window is sent verbatim, Kraft's inequality gives $\mathbb{E}\mathsf{L}^{\mathrm{WZ}}_{n,w} \ge H(X_1^n)$. If it is instead treated as an initial database available to both encoder and decoder, the conditional form of the same inequality gives

$$\mathbb{E}\mathsf{L}^{\mathrm{WZ}}_{n,w} \ge H(X_{w+1}^n \mid X_1^w) = H(X_1^n) - H(X_1^w).$$

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

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

$$\limsup_{w\to\infty} \limsup_{n\to\infty} \mathbb{E}C_{\mathrm{WZ}}(n;w) \le h_X.$$

Together with $h_w \ge h_X$, this implies $h_w \to h_X$. The argument concerns the expected iterated limit; it neither asserts that a fixed-w sample-path limit exists under Assumption F.3 nor controls a diagonal $w = w(n)$.

When $h_X > 0$, the match-length assertion is the theorem quoted as Lemma K.2: under the same stationary ergodic hypothesis, $L_n/\log_2 n \to 1/h_X$ in probability. The waiting-time formulation in the same lemma gives $\ell^{-1}\log_2 R_\ell \to h_X$ 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. □

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

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

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

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

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

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

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

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

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. □

O.1.1 Exact Wrapper for the Finite-Sample Certificate

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

1. If $|\mathsf{Defl}(e(y))| < 8L_{\max}b$, emit a one-bit compressed flag, the q_b-bit binary representation of the compressed length, and then $\mathsf{Defl}(e(y))$.
2. Otherwise emit a one-bit literal flag followed by an $8L_{\max}$-bit index for each of $y_1, \ldots, y_b$, for $8L_{\max}b$ 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 G.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 $a_1, \ldots, a_J$ taken from distinct documents. For a stationary source,

$$H(X_1^a) = \sum_{i=1}^{a} H(X_i \mid X_1^{i-1}) \geq a h_X \qquad (a \geq 1).$$

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 $\mathsf{Defl}(e(y))$. 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

$$\frac{h_X}{\kappa} \leq \mathbb{E}_{\mathrm{docs},U}[Z_U]. \tag{24}$$

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 $\bar{z}(\mathcal{D}) = \mathbb{E}[Z_U \mid \mathcal{D}]$. 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 $|\mathsf{Defl}(e(y))| < 8L_{\max}b$, 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

$$\bar{z}(\mathcal{D}) \le \bar{Z} + \varepsilon_M. \tag{25}$$

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

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

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

$$\mathbb{E}_{\mathcal{D}}\bar{z}(\mathcal{D}) \le \bar{z}(\mathcal{D}) + \sqrt{\frac{1}{2}\ln(2/\delta)\sum_k c_k^2} = \bar{z}(\mathcal{D}) + \varepsilon_{\text{doc}}. \tag{26}$$

On the intersection of the events in (25) and (26), (24) gives

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

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. □

Proof of Proposition G.4. The canonical map $\Lambda\colon \mathcal{M}_{\text{pred}} \to \mathbb{R}^{|\Sigma|}$ is Lipschitz by Assumption F.8. The density and support hypotheses of Lemma I.3(a) are supplied by Assumption F.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 $f_N \in \mathcal{F}_N$ such that

$$\varepsilon_N := \sup_{s \in \mathcal{M}_{\text{pred}}} \|f_N(s) - \Lambda(s)\|_\infty = O_\eta\Big(N^{-1/(d_{\text{pred}}+\eta)}\Big). \tag{27}$$

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

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

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

$$\mathrm{KL}_2(p_s\|q) \le \frac{\|\delta\|_\infty^2}{2\ln 2} \le \frac{\varepsilon_N^2}{2\ln 2}. \tag{28}$$

This estimate is uniform over the simplex and therefore introduces no lower bound on the coordinates of p_s. Integrating (28) with respect to ν and using the infimum in (13) yields

$$\Delta\mathcal{L}(N) \le \mathbb{E}_{s\sim\nu}\mathrm{KL}_2(p(\cdot \mid s) \,\|\, \operatorname{softmax}(f_N(s))) = O_\eta\Big(N^{-2/(d_{\text{pred}}+\eta)}\Big).$$

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

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

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

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

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

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

$$H(X_{t+1} \mid S_t) = H(X_{t+1} \mid X_{\le t}).$$

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 $m \to \infty$ identity proved in Proposition G.2. □

Proof of Theorem G.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

$$T_2(y_1, \ldots, y_k) = (qy_1, \ldots, qy_k) \pmod 1$$

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 F.1, F.2 and F.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,

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

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

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

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

$$\mathbb{P}(|\widehat{d}_n - 1| < \varepsilon) \longrightarrow 1, \qquad \mathbb{P}(|\widehat{d}_n - k| < \varepsilon) \longrightarrow 1.$$

The two events are disjoint, so their probabilities cannot both tend to one. This proves the obstruction. Theorem L.1 strengthens the same construction to the finite-n minimax lower bound $(k-1)/2$. □

Proof of Proposition G.8. Equation (14) 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

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

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 G.3, whose proof does not use a redundancy expansion. □

Proof of Lemma G.11. (i) $\ell_{\mathrm{utf}}(X_t)$ is a bounded measurable function of a stationary ergodic process, so Birkhoff's theorem gives $n^{-1}\sum_{t\leq n} \ell_{\mathrm{utf}}(X_t) \to \bar{\ell}_{\mathrm{utf}}$ almost surely, and the uniform bound $\ell_{\mathrm{utf}} \leq 4$ 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 $x_1^n \mapsto \mathsf{L}(\phi(x_1^n))$ an injective map with prefix-free range on the block alphabet Σ_{ch}^n, so Kraft's inequality gives $\mathbb{E}|\mathsf{L}(\phi(X_1^n))| \geq H(X_1^n)$. Stationarity gives $H(X_1^n) = \sum_{i\leq n} H(X_i \mid X_1^{i-1}) \geq nh_X$, 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 $-n^{-1}\log_2 \mathbb{P}(X_1^n) \to h_X$ 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 $\bar{\ell}_{\mathrm{utf}} \geq 1$, gives the stated limit. □

O.2 Proofs of Supporting and Appendix-Only Results

Proof of Lemma I.3. Part (a): Nakada & Imaizumi (2020) prove that, for every

$$\overline{\dim}_{\mathrm{M}} \operatorname{supp} \nu < d' < d_{\mathrm{pred,amb}},$$

deep ReLU networks achieve

$$\inf_{\hat{f} \in \mathcal{R}_N} \|\hat{f} - f\|_{L^\infty(\nu)} \;\le\; C_{f,d'} N^{-\beta_{\mathrm{app}}/d'}, \tag{29}$$

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 $d' = \overline{\dim}_{\mathrm{M}} \operatorname{supp} \nu$.

For the compact $C^{k_{\mathrm{pred}}}$ manifold $\mathcal{M}_{\mathrm{pred}}$ of Assumption F.7, the upper Minkowski, Hausdorff, and intrinsic dimensions all equal d_{pred}. The upper bound $d' < d_{\mathrm{pred,amb}}$ is part of Assumption F.7. Given $\eta > 0$, choose $0 < \eta' < \min\{\eta, d_{\mathrm{pred,amb}} - d_{\mathrm{pred}}\}$ and set $d' = d_{\mathrm{pred}} + \eta'$ in (29). Each scalar coordinate of the Lipschitz logit map extends from $\mathcal{M}_{\mathrm{pred}}$ to the ambient unit cube by McShane's theorem. Moreover, the strictly positive density in Assumption F.7 implies $\operatorname{supp} \nu = \mathcal{M}_{\mathrm{pred}}$. If u is continuous on $\mathcal{M}_{\mathrm{pred}}$, every nonempty relatively open subset of $\mathcal{M}_{\mathrm{pred}}$ has positive ν-measure, and hence

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

Thus (29) is a uniform bound on $\mathcal{M}_{\mathrm{pred}}$. Since $N \ge 1$ and $\eta' < \eta$, its Lipschitz specialization is the stronger estimate $O_{\eta'}(N^{-1/(d_{\mathrm{pred}}+\eta')}) = O_{\eta}(N^{-1/(d_{\mathrm{pred}}+\eta)})$. When $\beta_{\mathrm{app}} > 1$, this application requires the additional ambient or Whitney extension hypothesis stated in Remark I.1; Lipschitz regularity alone does not give that hypothesis.

Part (b): Throughout this part $\beta_{\mathrm{app}} \in \mathbb{N}$ 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.

The counting reduction is as follows. Place $q \asymp \varepsilon^{-d_{\mathrm{pred}}/\beta_{\mathrm{app}}}$ 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 $W^{\beta_{\mathrm{app}},\infty}$ 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 $N = \Omega(\varepsilon^{-d_{\mathrm{pred}}/(2\beta_{\mathrm{app}})})$ without a depth restriction and $N = \Omega(\varepsilon^{-d_{\mathrm{pred}}/\beta_{\mathrm{app}}} \log^{-2p_{\mathrm{dep}}-1}(1/\varepsilon))$ when $L_{\mathrm{net}} \le C_{\mathrm{depth}} \log^{p_{\mathrm{dep}}}(1/\varepsilon)$. These are minimax statements over the target ball; they do not give a lower bound for each individual logit map.

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

Oracle composition and finite-sample propagation.

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

$$d_{\mathrm{geo}} = \frac{h_X}{\lambda}. \tag{30}$$

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

$$d_{\mathrm{pred}} = \dim_{\mathrm{H}} \Phi(\mathcal{M}_{\mathrm{geo}}) \le \frac{d_{\mathrm{geo}}}{\gamma}. \tag{31}$$

Because $d_{\mathrm{pred}} > 0$ and $\gamma > 0$, (31) implies $d_{\mathrm{geo}} > 0$; then (30) and $\lambda > 0$ imply $h_X > 0$. Applying Proposition G.4 and taking reciprocals in (31) yields

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

It remains to identify the reference loss. The state S_t is past measurable and predictively sufficient, so Lemma G.5 gives $\mathbb{E}H(p(\cdot \mid S_t)) = h_X$: 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 G.7 have identical symbolic data but different d_{geo}, so a symbolic-only rule cannot recover (30) uniformly over the stated class. Without a sufficient factor, Assumption F.7 supplies no comparison between $\mathcal{M}_{\mathrm{geo}}$ and $\mathcal{M}_{\mathrm{pred}}$, and hence no replacement for (31). A separately supplied direct predictive state still supports Proposition G.4, and the symbolic entropy-rate result remains available in either case. □

Proof of Corollary H.3. Let

$$E_h = \{h_X \le U_h\}, \qquad E_\lambda = \{\lambda \ge \lambda_{\mathrm{L}}\}.$$

No independence is needed: the union bound gives

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

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

$$d_{\mathrm{geo}} = \frac{h_X}{\lambda} \le \frac{U_h}{\lambda} \le \frac{U_h}{\lambda_{\mathrm{L}}}.$$

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

$$d_{\mathrm{pred}} \le \frac{d_{\mathrm{geo}}}{\gamma} \le \frac{U_h}{\gamma \lambda_{\mathrm{L}}}.$$

If the direct-head assumptions Assumptions F.7 and F.8 additionally hold, taking reciprocals and applying Proposition G.4 gives

$$\alpha_{\mathrm{head}} \ge \frac{2}{d_{\mathrm{pred}}} \ge \frac{2\gamma\lambda_{\mathrm{L}}}{U_h}.$$

Every applicable conclusion holds on the same event. If λ is known exactly, set $\lambda_{\mathrm{L}} = \lambda$ and let E_λ be the sure event. □

P Case-Study Details and Overlap Checks

P.1 The Document-Level Quadrant Scorer

The corpus-level rate-slope ranker $\widehat{\alpha}_{\mathrm{LZ}}$ 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 $k_{\max} \le 32$), reported as

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

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 $k_{\max} = 64$, correlates with $\widehat{\alpha}_{\mathrm{LZ}}$ at Spearman $\rho = 0.76$ on 13 corpora. Aggregating the document scores gives $\rho = 0.58$ in the 18-corpus comparison of Section P.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 $\rho = 0.97$–0.99, $n \ge 173$k each), and their median splits agree on 93–97% of documents.

The four cells of the within-pool median split on $(c_i^{\mathrm{doc}}, r_{\mathrm{Hig}})$ 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 $\widehat{h}$ and r_{Hig} next to the corpus-level rate slope $\widehat{\alpha}_{\mathrm{LZ}}$, and they estimate neither of those corpus-level quantities.

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 contain 413/497/961/129 documents for Formal/Rich/Rote/Colloquial, so this sample's split is unbalanced on both coordinates. Formal contains 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 contains 65 of the 2,000 sampled documents: concise, well-structured problems with clean chains of thought. RICH contains harder and longer problems with denser reasoning, ROTE repetitive or trivially simple ones, and COLLOQUIAL rare, structurally malformed examples. These descriptions come from a few documents per cell. The selection experiments test the split by its training outcome.

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 measurement (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 describe agreement between the reported corpus rankings.

comparison	raw ρ	reliability	model-based reference	corrected ρ (95% CI)
$\widehat{h}$ vs. aggregate c^{doc}	$+0.84$	$0.99/0.99$	0.99	$+0.85$ $[0.53, 0.98]$
$\widehat{\alpha}_{\mathrm{LZ}}$ 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 raises the correlation 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 in 99% of them the corrected value is below 0.90.

P.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 $(c_i^{\mathrm{doc}}, r_{\mathrm{Hig}})$ score quadrants (FORMAL/RICH/ROTE/COLLOQUIAL, Section P.1) and pretrain transformers of four reported sizes (150M, 300M, 600M, 2B) on each quadrant at the Chinchilla budget $D = 20N$ tokens. Realized tokens, read out of each cell's own training record, are 3.00B at 150M and 6.00B at 300M, exactly $20N$. 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 $D/N = 19.2$ 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.

The 1.2B row is excluded. At 1.2B the four quadrants did not train on the same data. The grid restarted many times, some restarts switched pretokenized pools, and the four cells drew from two pools in different proportions, against sixteen reported cells that each trained on one pool throughout. Two of the four were also scored from checkpoints stopped mid-schedule rather than at the annealed endpoint, where every other reported cell ends within 2% of its cosine schedule. The four reported sizes still span 150M to 2B because 1.2B was interior to that range; results in Section D.1.

Shared selection and uncertainty. The same scored partition is used at every model size. The four-of-four 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 measures 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 includes the seed noise twice, so the margin at 2B (0.003) falls below the noise floor and we do not interpret it.

The mini-grid also has 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 measure 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 6 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 D.1.

P.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.

Q DATA AND MEASUREMENT DETAILS

DMO-Corpus implementation details. For a requested document budget the loader keeps strings longer than 100 characters, reads the declared text field with fallbacks, and joins documents with a blank line, repeating the corpus if it is shorter than twice the largest requested window. In the three-point floor calculation the implementation returns C_{32768} when the denominator $|C_{128}+C_{32768}-2C_{2048}|$ falls below 10^{-9}, and otherwise clips only negative extrapolates to zero. Window start indices are drawn from one fixed seed, and the reported $\widehat{h}$ values are the realized Monte Carlo outputs.

Q.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 9, 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 9: 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)
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)

continued on the next page

Table 9, continued

Corpus	Hugging Face dataset	Subset	Source
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	–	(Azerbayev 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	(Azerbayev et al., 2024)
Proof Pile 2 (arxiv)	aklein4/proof-pile-2-fixed	arxiv	(Azerbayev et al., 2024)
Proof Pile 2 (open-web-math)	aklein4/proof-pile-2-fixed	open-web-math	(Azerbayev et al., 2024)
PubMed (full text)	armanc/scientific_papers	pubmed	(Cohan et al., 2018)
SlimPajama-6B	DKYoon/SlimPajama-6B	default	(Soboleva et al., 2023)
SmolLM cosmopedia-v2	HuggingFaceTB/smollm-corpus	cosmopedia-v2	(Ben Allal et al., 2024)
Stack Edu Python	zaydzuhri/stack-edu-python	–	(Ben Allal et al., 2025)
StackExchange	HuggingFaceH4/stack-exchange-preferences	–	–
StackMathQA	math-ai/StackMathQA	stackmathqafull-1q1a	(Zhang, 2024a)

continued on the next page

Table 9, continued

Corpus	Hugging Face dataset	Subset	Source
StarCoderData	`bigcode/starcoderdata`	–	(Li et al., 2023)
TemplateGSM	`math-ai/TemplateGSM`	`templategsm-7473-1k`	(Zhang, 2024b)
The Pile Deduplicated	`EleutherAI/the_pile_deduplicated`	–	(Gao et al., 2020)
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`	–

Q.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.

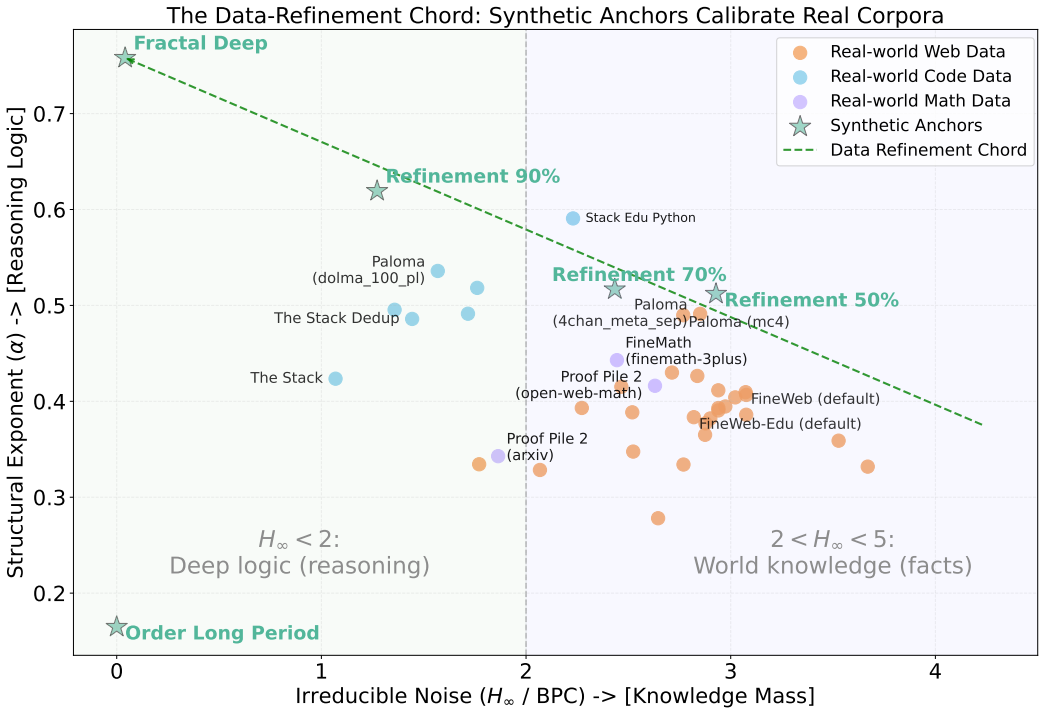


Figure 7: **Synthetic controls in the $(\widehat{h}, \widehat{\alpha}_{\mathrm{s}}^{(3)})$ plane.** Stars mark five 6-symbol sources evaluated by the same three-point routine: a deterministic L-system, three nested corruption levels ($p = 50/70/90\%$) generated with one fixed mask, and a repeated 1,000-symbol 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 coordinates in Figure 3. Generator details are in Section Q.3.

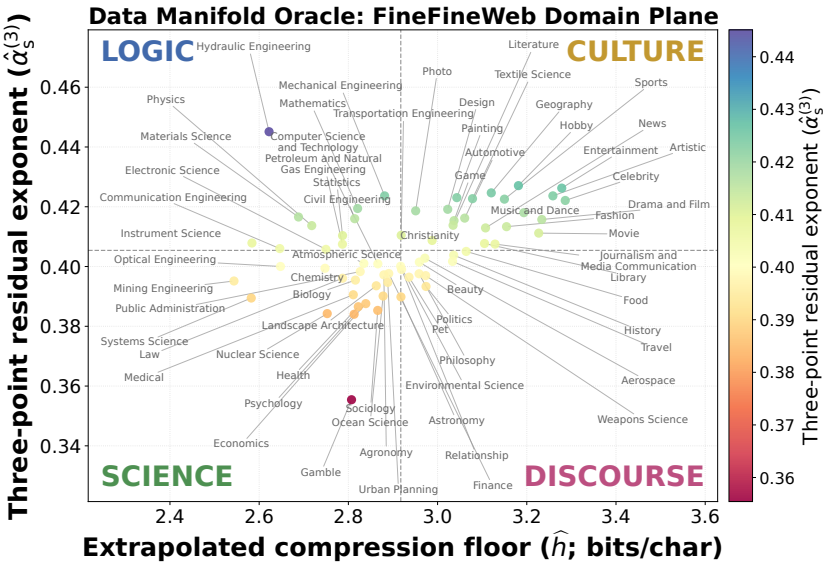


Figure 8: **Within-corpus domain comparison on FineFineWeb.** We apply the three-point statistic of Figure 1 once per domain; dashed lines mark the medians. Technical domains tend toward the low-$\widehat{h}$ side and cultural domains toward the high-$\widehat{h}$ side, so the same exploratory $(\widehat{h}, \widehat{\alpha}_{\mathrm{s}}^{(3)})$ 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.

Q.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 $A \to AB{+}C$, $B \to CA{-}B$, $C \to B/AC$, 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).

Q.4 Derivation of the Explanatory Three-Point Estimators

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

$$\widehat{H}_0^{(3)} = \frac{C_1 C_3 - C_2^2}{C_1 + C_3 - 2C_2}, \qquad \widehat{\alpha}_{\mathrm{s}}^{(3)} = \frac{\left|\log\big((C_1 - C_2)/(C_2 - C_3)\big)\right|}{\log g}.$$

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

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

Substitution in the slope definition gives $\widehat{\alpha}_{\mathrm{s}}^{(3)} = \alpha_{\mathrm{s}}$. For the floor estimator,

$$C_1 C_3 - C_2^2 = aH_0(x-1)^2,$$
$$C_1 + C_3 - 2C_2 = a(x-1)^2,$$

so $\widehat{H}_0^{(3)} = H_0$. Under the stated template, the denominator is nonzero and the adjacent-difference ratio is positive, so both estimators are defined. Outside that template, 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 G.8). We still compute the Case 1 floor by a high-sample three-point extrapolation. The deployed $\widehat{\alpha}_{\mathrm{LZ}}$, however, is the separate eight-point slope of the compression rate itself. It is neither $\widehat{\alpha}_{\mathrm{s}}^{(3)}$, the Hilberg parameter, nor $1/d$.

R The Entropy-Matched Parity Family

Proposition G.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 $n = 1, \ldots, 6$ and every support order $1, \ldots, 4$, 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.

Rendering. Each process step is written as one of four space-prefixed digits, each of which the GPT-NeoX tokenizer maps to exactly one token. One token is therefore one process step and the parity window is a token window, with no byte-pair grouping in between. The generator carries k symbols across chunk boundaries so the parity is continuous; a break there would insert an unpredictable step per chunk and would not affect the arms equally.

The measured statistics on the five arms.

$\lvert a\rvert$	$\widehat{\alpha}_{\mathrm{LZ}}$	start-index SD	$\widehat{h}$	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 $\widehat{\alpha}_{\mathrm{LZ}}$ is 0.0005 against a largest start-index SD of 0.0009. Both statistics fall mid-range for natural corpora, $\widehat{\alpha}_{\mathrm{LZ}}$ near the panel median of 0.180 and $\widehat{h}$ inside the code and math band, so these are not pathological inputs to the instrument. We froze the protocol, the locked per-arm predictions and the decision rule before training any cell; 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.

S Calibration Certificate on a Source With a Known Exponent

Section M.2 gives the paired result. This section records the full sweep and the reach of the deployed grid. All measurements use the deployed grid of Table 1 unless stated, on sources with 10^6 facts.

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

β	$\gamma_{\mathrm{CE}} = 1 - \beta$	$\widehat{\alpha}_{\mathrm{LZ}}$	$\widehat{\alpha}_{\mathrm{s}}^{(3)}$	$\widehat{\alpha}_{\mathrm{s}}^{(3)}$ (floor-subtracted)	$\widehat{h}$
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

$\widehat{\alpha}_{\mathrm{LZ}}$ 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 $\widehat{h}$ is rank-collinear with β and the floor-attenuation route and the memory route change $\widehat{\alpha}_{\mathrm{LZ}}$ in the same direction. The paired arm in Section M.2 is what separates them.

The three-point column 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.

Reach of the deployed grid. At 8192 characters a Santa Fe source has revealed only tens of distinct facts, so none of these measurements falls 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.

Table 10: **Estimator × corpus-set sensitivity of the exponent association.** Each cell reports Spearman $\rho(\widehat{\alpha}_{\mathrm{LZ}}, \alpha_D)$ / Pearson R^2 / leave-one-out Q^2 against *identical* fitted α_D; only the compressor behind $\widehat{\alpha}_{\mathrm{LZ}}$ and the corpus set vary. Sets: the selected 38-corpus set; that set with the six corpora excluded at the panel expansion added back; 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.

	$\rho(\widehat{\alpha}_{\mathrm{LZ}}, \alpha_D)$ / R^2 / Q^2 by corpus set		
$\widehat{\alpha}_{\mathrm{LZ}}$ estimator	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 closed form	.32 / .19 / .12	.33 / .18 / .11	.24 / .00 / −1.37

T Estimator Sensitivity

T.1 Compressors Ablation

This ablation compares how several compression algorithms behave as the available context length increases, using the rate-slope ranker $\widehat{\alpha}_{\mathrm{LZ}}$ and floor ranker $\widehat{h}$. We compare standard dictionary methods against more elaborate predictive models (figure in the supplementary note).

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* is a different question: we choose the compressor by how well $\widehat{\alpha}_{\mathrm{LZ}}$ ranks the fitted data-scaling exponent α_D, 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 $\widehat{\alpha}_{\mathrm{LZ}}$ varies), and repeat it on two wider corpus sets. Six of the ten exclusions were introduced when the panel expanded by eighteen corpora (Table 10).

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 $\widehat{h}$), 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 (2). 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 raw-rate slope $\widehat{\alpha}_{\mathrm{LZ}}$ and the three-point residual exponent $\widehat{\alpha}_{\mathrm{s}}^{(3)}$ agree only moderately on corpus rankings ($\rho \approx 0.52$), so the choice genuinely matters. 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 $R^2_{\mathrm{out}} < 0$; the gap remains after we partial out the overlap with their superseded exponents (details with the Case 1 statistics).