AUDITING CAPACITY TRENDS IN TRANSFORMER RESIDUAL-STREAM GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Interpreting scaling relationships in transformer residual-stream geometry requires separating measurement conventions, estimation precision, and predictive performance. We audit these factors using released model checkpoints, organizing the analysis around identifiability, reproducibility, trend diagnostics, and internal consistency. Our primary analysis measures late-layer cosine displacement over self-generated TriviaQA answers from six Pythia checkpoints under a prespecified reanalysis protocol. Among constant, capacity-only, and capacity-plus-shape models, the capacity-only model achieves the lowest mean leave-one-model-out absolute log prediction error. With capacity represented by $N_{\mathrm{proxy}} = 12d^2L$, its fitted exponent is -0.744, with a conditional ordinary-least-squares standard error of 0.093. Predictive errors remain substantial: the geometric-mean symmetric multiplicative error factor is 1.63, reaching 2.69 in the worst fold. Across three specified analysis windows, the capacity model retains the lowest mean leave-one-out error, although fold-level rankings differ and fitted exponents range from approximately -0.58 to -0.75. Adding a width-to-depth shape term does not improve the mean prediction metric in these comparisons; this finding does not establish the absence of shape effects. Historical extrapolation to two larger Pythia checkpoints further illustrates the limits of predictive accuracy. These results support a family- and window-dependent capacity trend and demonstrate how measurement choices affect its interpretation. Because models generate different text and we do not include fixed-text controls, the reported relationships characterize task-conditioned, self-generated trajectories rather than isolated causal effects of architecture.

1 INTRODUCTION

The residual stream of a decoder-only transformer is rewritten block by block. We measure the directional change between consecutive residual-stream states with a cosine displacement $R = 1 - \cos(h, h')$, and ask a question logically prior to fitting it: given a public set of released checkpoints, what about such a scaling relation can be *reliably estimated*? We treat this as a measurement and estimability audit, separating the measurement definition, the conditioning of the released design, the precision of the estimated coefficients, and their out-of-sample predictive performance before reading any of them as a law.

Relation to prior residual-stream geometry. A related observation already exists: Lawson et al. (2025) report that dataset-mean-centered adjacent-layer cosine similarity rises with model size across the Pythia suite (Biderman et al., 2023), measured on test text. The tuned lens documents cross-layer representational drift and uses learned affine transformations to decode intermediate predictions (Belrose et al., 2023). Our measurement differs in convention — an internal late-layer cosine displacement over self-generated answers, not mean-centered — so it is not the same quantity as the prior observation, and what we add is not a new observation but a joint check of what a fit to it supports: the measurement convention, the rank and conditioning of the released design, the uncertainty of the fitted coefficients, and out-of-sample prediction error, reported together rather than as a single exponent.

What we measure, and how we compare models. On six released Pythia checkpoints we measure the internal late-layer cosine displacement of the residual stream over self-generated Trivi-

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.

aQA answers, under a window and aggregation specified before the present reanalysis. We compare three models prespecified for this reanalysis — a constant, a one-variable capacity model in $N_{\text{proxy}} = 12d^2L$, and a model that adds $\log(d/L)$ — by their mean absolute log error under leave-one-model-out evaluation, and report that error rather than only the training fit.

What the comparison supports. Among the three candidate models the capacity model attains the lowest mean leave-one-model-out absolute log error and is our main empirical description; its error is not small and its fitted exponent varies across the windows examined, and on two held-out extrapolation checkpoints the capacity-plus-shape model is slightly better, so the ranking is specific to this evaluation. Adding $\log(d/L)$ does not reduce the mean leave-one-model-out absolute log error in the windows examined, which we report as an absence of incremental evidence, not an absence of shape effects. We use identifiability, reproducibility, trend diagnostics and internal consistency as complementary diagnostics, rather than as a universal set of necessary conditions for a scaling law. Because different models generate different text and we do not include fixed-text controls, the relations characterise self-generated trajectories.

Contributions.

- **A measurement and estimability audit.** We separate and report the measurement definition, the rank and conditioning of the released design, the precision of the estimated coefficients, and out-of-sample predictive performance — quantities not captured by fitted coefficients and in-sample R^2 alone.
- **A limited empirical capacity trend.** On six Pythia checkpoints a one-variable capacity model gives the lowest mean absolute log error under leave-one-model-out evaluation among three prespecified models; we report that error (geometric-mean factor 1.63, worst fold 2.69), the conditional exponent (-0.744, OLS SE 0.093), and a two-checkpoint historical extrapolation, none supporting a precise or universal exponent.
- **Window sensitivity and scope limitations.** The capacity choice is lowest in *mean* across three analysis windows while the fitted exponent and error level vary across them; adding $\log(d/L)$ does not reduce the mean leave-one-model-out absolute log error in the three analysis windows examined, without establishing that shape has no effect; and, absent fixed-text controls and with audited cross-seed evidence at a single model size, the relations characterise self-generated trajectories rather than architectural causes.

Supporting analyses — the depth/time axis separation and a pooling artefact near 0.25, a preregistered detection null, and cross-family comparisons — are reported under their own conventions in later sections. Related work is discussed in Appendix A.

2 The measurement and its estimability

This section defines the quantity we measure, fixes the window and aggregation used for the capacity re-analysis, states what is fitted and how it is scored out of sample, and closes with four diagnostics that bound what the fit can be asked to mean. The claims that follow stop at the level these diagnostics support.

2.1 The measured quantity, window, and aggregation

For a prompt processed by an L-block transformer, write h_0 for the embedding output and h_ℓ for the residual stream after block ℓ, so the zero-based transition j carries $h_j \to h_{j+1}$ and $j = L-1$ is the final block. The depth displacement at a transition is the cosine drift

$$R_{\text{depth},\,j} \;=\; 1 - \cos\big(h_j,\, h_{j+1}\big), \tag{1}$$

one scalar per transition, per token, per answer. The capacity re-analysis in this paper uses one frozen aggregation of equation 1, fixed before the re-analysis (Section 2.2); the supporting analyses of Section 6 state their own. The *re-analysis window* for a model of depth L is

$$W_L \;=\; \{\, \lfloor 0.6\,L \rfloor,\; \ldots,\; L-2 \,\}, \tag{2}$$

the late block transitions with the final transition $j = L-1$ excluded. Each answer is weighted equally — not each token, so long answers do not dominate. For model m, answer i with T_{mi} saved

forward steps,

$$r_{mi} \;=\; \frac{1}{T_{mi}\,|W_{L_m}|}\sum_{t=1}^{T_{mi}}\sum_{j\in W_{L_m}} R_{\mathrm{depth},\,mitj}, \qquad \bar{R}_m \;=\; \mathrm{median}_i\; r_{mi}, \qquad y_m \;=\; \ln \bar{R}_m. \tag{3}$$

Here T_{mi} counts stored forward-step rows, the first row predicting the first generated token; stored end-of-sequence prediction steps are kept, and a short answer is included whenever it has at least one valid row. Three checks apply at different stages and are not the same guarantee. On the stored displacement arrays the re-analysis substitutes no fallback for a non-finite or non-positive summary: it fails rather than smooth or drop, and the paired acceptance report released with the code is run at this stage. The extraction that produced the arrays used a denominator numerical guard in the cosine against division by a zero norm; this is a numerical safeguard at extraction time, not a zero-norm detection or a semantic acceptance check, and it is not re-verified here. The raw pre-storage state conditions were not audited directly in this re-analysis, and we do not claim they were.

2.2 Models, data, and the pre-specified re-analysis

The fit uses six Pythia models (Biderman et al., 2023) (70M–2.8B), which share one training recipe — one corpus of $\sim$300B Pile tokens (Gao et al., 2020), one architecture family, one optimizer schedule — so that d and L are the design axes that move. Two larger Pythia models, 6.9B and 12B, are held out of every fit and used only as historical extrapolation points; they are not re-fit and not treated as a blind test. All models saw the same 200 TriviaQA (Joshi et al., 2017) items under 5-shot plain prompting. The window equation 2 and aggregation equation 3 were frozen before this re-analysis. Trajectories are stored in float16 and upcast to float64 for analysis, which does not restore precision lost at storage. A paired acceptance check, released with the code, establishes only that, on the samples it pairs, the window arrays of the raw and corrected exports agree to float16; it does *not* establish that the correction changes solely the final state — that would need a code-path argument — and the revision and precision metadata remain unaudited and are marked unknown. Excluding $j = L-1$ is therefore a choice of window, not a removal of contamination: the corrected final-block update is a valid measurement, and dropping it defines a different window rather than a cleaner one.

2.3 What is fitted, and how it is scored

Three nested mean models for y_m are compared: a constant, a capacity model $y = a + b\,\ln N_{\mathrm{proxy}}$ with $N_{\mathrm{proxy}} = 12\,d^2L$, and a capacity-plus-shape model adding $c\,\log(d/L)$. Each is fitted by ordinary least squares, minimising the squared log residual on its training points. Models are *scored* by leave-one-model-out mean absolute log error,

$$\mathrm{MALE}_{\mathrm{LOO}} \;=\; \frac{1}{6}\sum_{m=1}^{6} \left|\, y_m - \hat{y}_m^{(-m)} \,\right|, \tag{4}$$

each fold re-fitting on the other five points. The statistical unit is the six model points; the folds share training data and are not six independent replications, and the same LOO used to choose among candidates cannot also serve as an independent confirmation of the chosen model. Table 1 reports the comparison. In the frozen main window the capacity model has the lowest mean out-of-sample error (0.490), a constant baseline is far worse (1.322), and the shape term raises mean LOO error to 0.650 while in-sample R^2 is 0.942 for both models to three decimals — no incremental out-of-sample evidence for shape in this six-point design, which is not evidence that shape has no effect. The choice is not an artefact of the one window: re-scored under the include-final window $\{\lfloor 0.6L \rfloor, \ldots, L-1\}$ and the fixed last-two interior window $\{L-3, L-2\}$, capacity keeps the lower *mean* LOO error, but this is a shared-data robustness check, not three independent replications, and the ranking is not uniform — in the last-two window the mean favours capacity while it wins only 2 of 6 held-out folds. Held out entirely, Pythia-6.9B and 12B sit at absolute log errors of 0.448 and 0.239 under capacity; these are the two points on which capacity+shape is marginally better (0.443, 0.233), a small difference on two historical points that we report rather than resolve the models on.

Table 1: Model comparison across three prespecified reanalysis windows, scored by leave-one-model-out mean absolute log error equation 4 over the six Pythia fit points; the frozen main window is not re-selected. "cap wins" counts held-out folds where capacity beats capacity+shape. A constant baseline scores 1.322 in the main window. Per-fold errors and coefficients are archived with the released analysis scripts.

window	cap MALE	cap+shape MALE	cap wins	capacity exponent (SE)
main $\{\lfloor 0.6L \rfloor..L-2\}$	**0.490**	0.650	6/6	-0.744 (0.093)
include-final $\{..L-1\}$	0.208	0.286	5/6	-0.576 (0.048)
last-two $\{L-3, L-2\}$	0.512	0.586	2/6	-0.748 (0.112)

2.4 Uncertainty of the fitted coefficient

Under the frozen window the capacity model gives an exponent of -0.744 with a conditional standard error of 0.093 — the classical fixed-design OLS value $\mathrm{SE} = \sqrt{[\hat{\sigma}^2 (X^\top X)^{-1}]_{kk}}$, $\hat{\sigma}^2 = \mathrm{RSS}/(n-p)$, with $n = 6$, $p = 2$. It assumes a correct linear mean and uncorrelated homoscedastic errors (normality only for an exact finite-sample t interval). Crucially the six points share a model family and a prompt set, so the independence and uncorrelated-error assumptions are not verified by the design: 0.093 is a standard error *conditional on the model's assumptions*, not a verified population quantity, and it excludes the uncertainty of choosing the model or the window and the cross-seed dispersion that Section 6 audits at a single size. Appendix C discusses which coefficient combination this variance leaves well determined.

2.5 Four estimability diagnostics

Four diagnostics bound what the fit can mean. We state them as checks on measurement consistency, estimation precision, and scope — not as universal necessary conditions for a law.

D1. Identifiability and precision. The centred design over $(\ln d, \ln L)$ for the six sizes has cross-product eigenvalues with ratio $\lambda_1/\lambda_2 = 10.95$ — its Gram-matrix condition number (the design-matrix 2-norm condition number is $\sqrt{\lambda_1/\lambda_2}$). Full column rank makes both coefficients identifiable; the condition number is a statement about estimation *precision*, not a certificate that any particular combination is well estimated. Under this design the capacity combination is estimated more precisely than the shape combination — it aligns with the strong eigen-direction, the shape combination with the weak one (Appendix C, Figure 1). This is a relative statement within the two-variable fit (there the capacity combination carries a conditional SE of 0.109), distinct from the 0.093 headline SE of the one-variable capacity slope.

D2. Seed-dispersion reference. The dispersion of $\ln \bar{R}$ across independently seeded runs describes how much of a response could be seed-driven variation. We measured it at a single size (410M, Appendix D) and treat it as a descriptive reference for that scale, not a universal detection threshold below which every effect or model is ruled out; for families where it was not measured it is marked unknown rather than passed or failed. Dispersion is compared as a standard deviation, never a max-to-min ratio, whose expectation grows with sample size.

D3. Monotonicity as a trend description. Along the released sequence, R_{depth} falls monotonically across approximately $133\times$ in the architectural capacity proxy $N_{\mathrm{proxy}} = 12d^2L$. Non-monotonicity along a sequence where d and L co-vary would *not* by itself refute a multivariate power law — for instance $\ln \bar{R} = -\ln N + \ln \eta$ with $\eta = d/L$ is an exact law whose response is non-monotone along a release sequence in which N and η move together; we therefore read monotonicity as an ordering description of the observed sequence, not as a test for whether a trend exists.

D4. Internal consistency. The residual stream is additive, so a block's write decomposes as $\Delta_l = \Delta_{\mathrm{attn}} + \Delta_{\mathrm{mlp}}$, and the rotation and norm growth read from the hidden states can be checked against the two submodule writes measured separately (Appendix B). Because both sides are computed from the same forward pass, this check covers only the evaluated layers and samples.

Table 2: Data provenance of the four families, which is not uniform. A cross-family comparison therefore varies these factors together; only the Pythia grid, Qwen3 and SmolLM2 share base weights, plain few-shot prompting and the corrected convention. (This table describes the data; the analysis window used for the cross-family numbers is stated in Table 3.)

family	sizes	weights	prompting	convention
Pythia	70M–2.8B	base	plain few-shot	corrected
Qwen2.5	0.5B–7B	instruct	chat template	raw
Qwen3	0.6B–8B	base	plain few-shot	corrected
SmolLM2	135M–1.7B	base	plain few-shot	corrected

Table 3: The same probe across four families, a descriptive comparison and not one controlled experiment (provenance in Table 2). **All window-dependent columns — response, the σ ratio, and the slope $\hat{b}_N^{(1)}$ — use the historical include-final window**, not the frozen-window main analysis; under the frozen window Pythia's capacity exponent is -0.744 (Table 1), against the -0.576 here. $\hat{b}_N^{(1)}$ is the *single-variable* capacity fit, a regression on $\ln N$ alone, so the last column's $\ln N$–$\ln \eta$ correlation is *not* its collinearity; that correlation is a separate diagnostic of the conditioning a *two*-variable capacity+shape fit would face. The three non-Pythia slopes are near zero and inconsistent in sign; we report them as what the single-variable fit returns and attach no precision rating here. The σ ratio is family dispersion over a Pythia-410M seed dispersion, and its two endpoints are two *denominators*, not a confidence interval: the all-nine-seed spread ($s = 0.356$) and a sensitivity subset dropping the two lowest-level seeds ($s = 0.222$, a trim rather than a cleaning — the seven are not more valid than the nine). The other families have no audited cross-seed spread of their own.

family	n	N range	response	$\sigma_{\ln \bar{R}}/s_{\text{seed}}$	monotone	$\hat{b}_N^{(1)}$	$\text{corr}(\ln N, \ln \eta)$
Pythia	6	133×	14.4×	2.99–4.81	yes	-0.576	$+0.203$
Qwen2.5	4	19×	1.39×	0.41–0.65	no	$+0.009$	$+0.919$
Qwen3	4	21×	1.35×	0.37–0.60	no	-0.034	$+0.968$
SmolLM2	3	10×	1.64×	0.69–1.11	no	$+0.117$	$+0.982$

It corroborates that the reported series and the submodule hooks decompose the same write; it is not an independent test that would localise an instrument fault on its own. Appendix Figure 2 shows, under the earlier include-final two-variable fit, the same qualitative cross-family contrast: one suite falls monotonically while three others stay flat and non-monotone.

3 Results

The one-variable capacity model of Section 2 is identifiable wherever parameter count varies, which is every family; what distinguishes the Pythia suite is that a monotone capacity trend is actually present there and its design is better conditioned than the others'. The rest of this section asks what the same probe returns in families that lack one or both, and what the limited capacity trend this paper reports can and cannot be used for.

3.1 The same probe on three further families

We ran the same probe on Qwen2.5 (Qwen Team, 2024), Qwen3 (Qwen Team, 2025) (0.6B–8B) and SmolLM2 (Allal et al., 2025) (135M–1.7B), chosen by screening the open zoo on design conditioning *before* any measurement. Their provenance differs from the Pythia grid and from one another (Table 2), so the comparison is descriptive: a cross-family difference is confounded by architecture, training recipe, adaptation state, prompting format and measurement convention at once (Section 6), and we do not attribute it to any single factor.

Table 3 compares the four families descriptively. Pythia's median R_{depth} falls monotonically across approximately $133\times$ in the capacity proxy $N_{\text{proxy}} = 12d^2L$; the other three are flat and non-monotone over their released sizes. We read this as a description of these sequences, not as proof that no multivariate law exists for them: as D3 in Section 2.5 notes, non-monotonicity along a se-

quence where d and L co-vary does not by itself refute a law. Pythia's seed-referenced dispersion is larger than the other families', but the other families have no audited cross-seed spread of their own, so this is a descriptive comparison to the single Pythia-410M reference, not a threshold any family passes or fails.

Two limited checks bound, but do not eliminate, range and window explanations. Normalising per decade of N, the responses are $3.51\times$ for Pythia against 1.26–$1.63\times$ for the others, so the contrast is not merely a wider range; and dropping the last transition leaves all three non-monotone with Pythia still monotone (response rising to $33\times$), so it is not created by that window choice. Neither check rules out every range or window effect — they show only that the contrast persists under these two. It is also not incapacity: task accuracy rises 2.1–$3.0\times$ across the families whose geometry barely moves (Qwen3, $0.20 \to 0.60$; SmolLM2, $0.16 \to 0.48$).

The two-variable capacity+shape design is a separate matter, and it is not one of identifiability: for each family the design matrix $[\,1,\ \ln N,\ \ln \eta\,]$ is full rank (rank 3), so its coefficients are unique. What the near-collinear design ($\mathrm{corr}(\ln N, \ln \eta)$ of 0.92–0.98) and the few released sizes cost is estimation *precision*, not identifiability; and SmolLM2, with three sizes, saturates the three-parameter model exactly, leaving zero residual degrees of freedom and no usual standard error at all. Pythia's 0.203 — a 1B model at $L{=}16$ beside a 1.4B at $L{=}24$ — is better conditioned, a property of which checkpoints exist rather than of the architectures.

3.2 A WIDTH- AND DEPTH-MATCHED COMPARISON

An earlier include-final *two-variable* fit was also used as a reference frame, placing other families against the Pythia surface by their offset Δ at each model's own (d, L) rather than fitting them; because that is the historical two-variable model, not the frozen-window capacity model of Section 2, we keep it only as a design-sensitivity analysis in Appendix C and do not read the offsets as a recipe's position. A fit-free alternative matches width and depth directly — though matched (d, L) is not matched architecture. The archive's only exact (d, L) match among nineteen models is Pythia-1.4B and SmolLM2-1.7B (both $d = 2048$, $L = 24$), which also share base weights, plain few-shot prompting and the corrected convention (Table 2). Under the historical include-final window their median R_{depth} differs by $1.60\times$ (SmolLM2 over Pythia), larger than the $1.32\times$ between two Pythia models at the same width but different depths. We report 1.60 as a descriptive ratio only: it is one pair, with the rest of the architecture and any unlogged pipeline differences uncontrolled, so it is not evidence that recipe causes the gap, and the archived bootstrap resamples generations rather than shared items, so we quote no interval for it. Nor is the offset a fixed family constant: across four width-matched but depth-mismatched pairs the Qwen3-to-Pythia level ratios are 0.55, 1.48, 1.81 and 2.22. These vary across the examined pairs and do not establish a family divergence under controlled architecture (Appendix C).

4 WHAT THE DEPTH TREND IS NOT

It does not extend to the time axis. The same functional form fitted to temporal displacement gives $R^2 = 0.62$, and an intercept-only model wins the nested leave-one-out comparison: the depth trend does not carry to the time axis. We report this as an empirical result under its own protocol (Section 6), not as a prediction the functional form was constructed to make.

Pooling the two axes manufactures an apparent constant. In the saved samples of these five models the depth and time axes occupy non-overlapping empirical ranges, so any fixed threshold placed between them is degenerate on both in this data — it never fires on depth and always fires on time. On this data and the specified kernel-density settings, pooling the two axes produces a separated density structure, and the density minimum between the two axis medians falls at 0.235 (Qwen2.5-0.5B), 0.208 (1.5B), 0.195 (3B), 0.212 (SmolLM2-1.7B) and 0.307 (TinyLlama-1.1B). An analyst who pools the axes and reports that minimum would therefore "discover" a critical value near 0.25 on 0.5B-class models and a different one elsewhere, with no dynamical content required — a measurement-axis mixture suffices (Figure 5). Within the Qwen family the reported location declines monotonically with scale, tracking the decline of both axis levels. We report this as a general caution: thresholds on this quantity, if used at all, must be stated as (axis, model, task, target-FPR) quadruples.

Limited incremental detection value. The quantity was motivated by failure detection, and the pre-registered test of that motivation returns a null. Across the confirmatory cells, depth-axis features separate correct from incorrect generations on their own, but adding the eleven-feature geometry set to single-pass probability baselines computed from the *same* forward pass changes held-out AUROC by about -0.005 on average over the nine confirmatory cells; only two cells (Qwen2.5-TriviaQA at 0.5B and 1.5B) have a positive point estimate, the 1.5B interval spanning zero, and a pre-registered branch predicting geometrically *quiet* confabulations was not supported. This characterizes the observed increment of that feature set under this protocol and these baselines; it is not a claim that R_{depth} alone carries no downstream signal. The question we ask is whether the variation in R_{depth} can be described at all, not whether it is a detector.

5 Discussion

The positive result is narrow and stated at the level its evidence supports: in one suite whose design identifies it, a one-variable capacity model in N_{proxy} gives the lowest mean leave-one-model-out error among the three candidates in the frozen window (Table 1). Its out-of-sample error is not small, and the six-point fit and the two entirely held-out historical extrapolations (Pythia-6.9B and 12B) are separate evaluations that should not be run together. The model *choice* is stable across the three windows examined; what moves with the window is the exponent, the error level and the per-fold ordering. We therefore report a limited, window-sensitive capacity trend, not an architectural law.

Shape, and what the data do and do not say. Adding a width/depth shape term does not lower the mean leave-one-out error in this six-point design (Table 1), though capacity+shape is marginally better on the two historical extrapolation points. This is an absence of incremental mean-LOO evidence for shape, not evidence that shape has no effect: the shape coefficient's interval spans zero, so we neither confirm nor exclude a shape dependence, and the estimate is too imprecise to contrast with reports that loss is insensitive to width/depth shape (Kaplan et al., 2020).

The cross-family comparison is a bounded observation. Across three further families the same probe returns a flat, non-monotone response, but the comparison is descriptive: it varies architecture, recipe, adaptation state, prompting format and measurement convention together (Section 6). The other families' measured sequences do not show the same monotone decline, and their *two*-variable capacity+shape designs are near-collinear with few sizes, so those coefficients are estimated imprecisely; neither fact proves a trend is absent or prevents the *one*-variable capacity model from having a unique coefficient estimate in every family. What we can point to is a property of the released checkpoints, not of the models: the three families correlate $\ln N$ with $\ln \eta$ at 0.92–0.98 against Pythia's 0.203 — a fact about which (d, L) pairs a laboratory published, unchanged if the internal geometry were identical.

What transfers. The method transfers; the content does not, and we do not claim it. The centring, the design-spectrum screen, and the seed-dispersion reference all ran unchanged on four suites. But four families are not a sample from which the population of transformers can be inferred, and three share a post-2023 lineage. We present Pythia's limited capacity trend as the main empirical result and the other families as a descriptive comparison; the framework is *applicable* across four families and *universal* in neither exponent nor level.

Limitations. The largest is measured rather than argued. Each fitted model is a single pre-training run, and at 410M the seed dispersion is a large fraction of the reported cross-family deviations; this study does not include or audit independent-seed data above 410M, so this reference is unmeasured at the three sizes carrying most of the fit's range and we do not extrapolate it there, so every statement about the size of a cross-family deviation should be read with that in mind. Detection is correlational and supports no intervention claim. Pythia is measured under plain few-shot base prompting while some validation models use instruction-tuned weights and chat templates — a confound we disclose rather than remove. In our only ablation (Qwen2.5-0.5B, historical include-final convention), the depth-axis level was about 1.5% higher with plain prompting than with the chat template — a local sensitivity for that model and protocol, not an upper bound on cross-family, cross-scale or adaptation-state confounds. Mixture-of-experts architectures have no single (d, L),

so they exit the parameterisation rather than stressing it. Establishing broader generality requires additional evidence beyond the released suites examined here; controlled comparisons across aspect ratios would be one useful next step.

6 METHODS

Metrics. For a decoder-only transformer with L blocks, let $h_l(t) \in \mathbb{R}^d$ be the residual-stream state after block l at generated position t, with $l = 0$ the embedding output. During greedy generation we record, in float32,

$$R_{\text{depth}}(t,l) = 1 - \cos\big(h_l(t), h_{l-1}(t)\big), \quad R_{\text{time}}(t,l) = 1 - \cos\big(h_l(t), h_l(t-1)\big),$$
$$\Delta\ln\|h\|(t,l) = \ln \tfrac{\|h_l(t)\|}{\|h_{l-1}(t)\|}, \tag{5}$$

together with token-level baselines from the same forward pass (next-token entropy, maximum probability, chosen-token log-probability, generation length). Every quantity is a cosine displacement $1 - \cos$ or a log-norm ratio between two specified vectors, with no stronger structure assumed. The capacity re-analysis uses the frozen window W_L of Section 2.1, which excludes the final transition; the detection study instead aggregates its features over a late-depth window (Appendix F). Per-sample $[S \times L]$ trajectories are archived, so every profile analysis runs offline without a GPU.

Why the log–log–linear form. The candidate models compared in Section 2.3 are all log–log–linear in counts: a parsimonious descriptive family, not a proven global approximation, chosen so that capacity and shape enter as separate terms and simple hypotheses appear as parameter restrictions (width independence is a zero width coefficient). We claim no theoretically derived exponent, and the time-axis result (Section 4) is an empirical finding under its own protocol, not a prediction of this form.

Models, tasks and prompting. Scaling grid: Pythia 70M, 160M, 410M, 1B, 1.4B, 2.8B (base weights, 5-shot plain prompting; one recipe of $\sim$300B Pile tokens (Gao et al., 2020), corrected convention). In-family extrapolation: Pythia-6.9B and 12B, held out entirely from the fit (base, corrected). The out-of-family runs are not a uniform base-weight set, and we list their actual state because it bears on what the cross-family contrast can be attributed to: Qwen2.5 (0.5B–7B) uses *instruction-tuned* weights with *chat-template* prompting on the raw convention; Qwen3 (0.6B–8B) and SmolLM2 (135M–1.7B) use *base* weights with the same plain few-shot prompting as Pythia on the corrected convention. So a cross-family difference confounds architecture, training recipe, adaptation state (base vs instruction-tuned), prompting format (plain few-shot vs chat template), and measurement convention at once, and cannot be read as an architectural or recipe effect. Tasks with objective labels are GSM8K and TriviaQA rc.nocontext, $N = 200$ items per model$\times$task with identical item sets across models (the detection study's five instruction-tuned models are listed in Appendix F). Three precisions are distinct and not interchangeable: inference ran in fp16 where the run logs record it (batch size 1, greedy decoding, single T4 unless weights required sharding), the metrics are computed in float32 (Eq. 5), and trajectories are stored in float16 (Section 2.2). A few runs' inference precision and revision metadata are not recorded and are left unverified rather than asserted.

Statistics. Uncertainty on the main capacity coefficient is the conditional standard error of Section 2.4 and the model comparison uses the leave-one-model-out metric of Section 2.3; cross-family comparisons bootstrap over models, the level that actually varies. Detection significance uses Holm-corrected Monte-Carlo permutation tests, detailed with the per-cell results in Appendix F. Prediction error is reported as a signed log residual $\delta = \ln \bar{R} - \ln \hat{R}$ and its multiplicative ratio e^{δ}.

Conventions and protocol. Runs are measured under two normalisation conventions and every quantity states which (Appendix B): the capacity fit and its extrapolation are on the corrected convention, but the cross-family comparison sets Qwen2.5 (raw) against Qwen3 and SmolLM2 (corrected) — a cross-convention comparison, audited and flagged, not silent. The detection protocol was frozen before the confirmatory runs and returned the pre-registered YELLOW outcome (Appendix F; deviations logged in the released protocol document (Hofman et al., 2023; Pineau et al., 2020)).

REPRODUCIBILITY STATEMENT

We distinguish results by how far the shipped archive verifies them. *(a) Automatically checked.* A one-command checker (`reproduce.py`) re-derives five headline values — the frozen-window capacity exponent, its conditional standard error, the fit R^2, and the two signed historical-extrapolation residuals — from the shipped raw per-sample $[S \times L]$ trajectories of the six-model fit grid and, for the two extrapolation residuals, of the two held-out models (Pythia-6.9B and 12B); it compares each, at the paper's printed precision, against the value stated in the paper, refusing if the sample set does not match a frozen manifest. It does not re-check every entry of Table 1 and does not run the detection study. *(b) Re-runnable from shipped trajectories, by their own commands.* The item-level bootstrap — the β, γ coefficient intervals of the historical corrected, include-final *two-variable* fit, not the single-variable headline coefficient — runs on the same fit-grid trajectories, and the pre-registered detection null is recomputed by a separate command that re-extracts features from the shipped detection-shard trajectories. *(c) Report only.* The matched-scale control, the from-scratch family comparison, and the joint architecture–data conditioning check rest on trajectories too large to ship; for these we include the anonymised report and run log so their outputs and verdicts are inspectable, but they cannot be re-run end to end from this package alone. The archive also contains the extraction probe, the frozen pre-registration protocol with its deviation log, the analysis and quality-assurance scripts, and the derived per-run tables for every deposited run. The complete trajectory archive of the curated runs will be deposited with a persistent identifier and released publicly upon publication; the identifier and repository links are withheld during review only because they would break double-blind anonymity. All datasets used are public: GSM8K and TriviaQA rc.nocontext. Section 6 states the measurement convention each quantity rests on, and Appendix B records the audit that enforces it.

REFERENCES

Yedidia Agnimo, Anna Korba, Annabelle Blangero, Nicolas Chesneau, and Karteek Alahari. Evaluating the relevance of uncertainty estimators for llm hallucination. *arXiv preprint arXiv:2605.27016*, 2026.

Aisha Alansari, Malak Alkhorasani, and Hamzah Luqman. Crosshallu: Do hallucination signals generalize across languages and domains in large language model's internals? *arXiv preprint arXiv:2607.04029*, 2026.

Loubna Ben Allal et al. Smollm2: When smol goes big — data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Kristijan Armeni, Marko Pranjić, and Senja Pollak. Transformer verbatim in-context retrieval across time and scale. *arXiv preprint arXiv:2411.07075*, 2024.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Maissam Barkeshli, Alberto Alfarano, and Andrey Gromov. On the origin of neural scaling laws: from random graphs to natural language. *arXiv preprint arXiv:2601.10684*, 2026.

Nora Belrose, Quintin Pope, Lucia Quirke, Alex Mallen, and Xiaoli Fern. Neural networks learn statistics of increasing complexity. *arXiv preprint arXiv:2402.04362*, 2024.

Nora Belrose et al. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Stella Biderman et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2304.01373.

Chao Chen et al. INSIDE: LLMs' internal states retain the power of hallucination detection. In *International Conference on Learning Representations (ICLR)*, 2024.

Nolan Dey et al. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster. *arXiv preprint arXiv:2304.03208*, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Jesseba Fernando and Grigori Guitchounts. Dynamics of the transformer residual stream: Coupling spectral geometry to network topology. *arXiv preprint arXiv:2605.14258*, 2026.

Leo Gao et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Arthur E. Hoerl, Robert W. Kennard, and Kent F. Baldwin. Ridge regression: Some simulations. *Communications in Statistics*, 4(2):105–123, 1975.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Jake M. Hofman, Angelos Chatzimparmpas, Amit Sharma, Duncan J. Watts, and Jessica Hullman. Pre-registration for predictive modeling. *arXiv preprint arXiv:2311.18807*, 2023.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

Saurav Kadavath et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

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.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.

J. F. Lawless and P. Wang. A simulation study of ridge and other regression estimators. *Communications in Statistics — Theory and Methods*, 5(4):307–323, 1976.

Tim Lawson, Lucy Farnik, Conor Houghton, and Laurence Aitchison. Residual stream analysis with multi-layer SAEs. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2409.04185.

Quan Liu, Han Zhou, Wenquan Wu, Hua Wu, and Sen Su. Residual stream analysis of overfitting and structural disruptions. *arXiv preprint arXiv:2603.13318*, 2026.

Huan Ma, Jingdong Chen, Joey Tianyi Zhou, Guangyu Wang, and Changqing Zhang. Estimating llm uncertainty with evidence. *arXiv preprint arXiv:2502.00290*, 2025a.

Huan Ma, Jiadong Pan, Jing Liu, Yan Chen, Joey Tianyi Zhou, Guangyu Wang, Qinghua Hu, Hua Wu, Changqing Zhang, and Haifeng Wang. Semantic energy: Detecting llm hallucination beyond entropy. *arXiv preprint arXiv:2508.14496*, 2025b.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Rohan Pandey. gzip predicts data-dependent scaling laws. *arXiv preprint arXiv:2405.16684*, 2024.

Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010. doi: 10.2202/1544-6115.1585.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *arXiv preprint arXiv:2003.12206*, 2020.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Andres Saurez, Yousung Lee, and Dongsoo Har. Why linear interpretability works: Invariant subspaces as a result of architectural constraints. *arXiv preprint arXiv:2602.09783*, 2026.

Harald Semmelrock, Simone Kopeinik, Dieter Theiler, Tony Ross-Hellauer, and Dominik Kowald. Reproducibility in machine learning-driven research. *arXiv preprint arXiv:2307.10320*, 2023.

Adam S. Shai, Sarah E. Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream. *arXiv preprint arXiv:2405.15943*, 2024.

Wenjie Sun, Jinning Yang, Shuai Zhang, and Mengnan Du. Law of neural interaction: Depth-width shape, interaction efficiency, and generalization. *arXiv preprint arXiv:2605.27989*, 2026.

Team OLMo et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. LLM circuit analyses are consistent across training and scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2407.10827.

Oskar van der Wal, Pietro Lesci, Max Muller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. Polypythias: Stability and outliers across fifty language model pre-training runs. *arXiv preprint arXiv:2503.09543*, 2025.

Gijs van Dijk. Detecting hallucinations in large language models via internal attention divergence signals. *arXiv preprint arXiv:2605.05025*, 2026.

Zhipeng Yang, Junzhuo Li, Siyu Xia, and Xuming Hu. Internal chain-of-thought: Empirical evidence for layer-wise subtask scheduling in llms. *arXiv preprint arXiv:2505.14530*, 2025.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

Susan Zhang et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Related work

Residual-stream geometry and adjacent-layer diagnostics. The quantity we measure is not new as a diagnostic. Lawson et al. (2025) report that cosine similarity between adjacent-layer residual activations rises with model size across the Pythia suite. That observation is qualitative, it conflates width with depth because the suite scales both together, and it removes norm growth by mean-centring; we measure the same quantity on a grid where width and depth move independently, retain norm growth as a separate axis, and fit the resulting surface. The tuned lens (Belrose et al., 2023) documents rotation and norm growth with depth without quantifying their dependence on architecture. Spectral analyses of the layer-to-layer Jacobian reach a compatible picture from a different operator (Fernando & Guitchounts, 2026; Liu et al., 2026): rotation-dominated early depth

giving way to near-symmetry, which is consistent with our geometric trend though the operator being measured is not the same. A larger body of work asks *what* directions encode rather than *how much* the stream turns — truth-valued directions with consistent geometry (Marks & Tegmark, 2023), simplex-like belief structure (Shai et al., 2024), layer-wise subtask stages during reasoning (Yang et al., 2025), with a recent argument that such linear structure is itself a consequence of architectural constraints (Saurez et al., 2026). All take the coordinate system as given; the magnitude of per-block movement fixes the scale on which any such direction is measured and has not been given a law.

Scaling laws, and why our target is a different kind of quantity. Neural scaling laws relate loss to parameters, data and compute (Kaplan et al., 2020; Hoffmann et al., 2022), and a line of work asks where that power-law structure originates, from the statistics of the data (Pandey, 2024) to the graph structure of the generative process (Barkeshli et al., 2026). Our target differs in kind: not an end-to-end performance quantity but an internal magnitude. The most direct dialogue is with Tigges et al. (2024), who find circuit-level algorithms consistent across training and scale and state as an explicit limitation that covering only the Pythia family cannot distinguish a property of scale from a property of that family's recipe; related work follows retrieval behaviour and learned statistics along the same lineage (Armeni et al., 2024; Belrose et al., 2024). Our cross-family transport measurements speak to that open question at the level of geometric magnitudes: within-family and cross-family consistency behave differently in our data, with within-family alignment holding and cross-family alignment not. Because that cross-family contrast confounds architecture, training recipe, adaptation state, prompting format and measurement convention at once, we report this as an observation, not as a resolution of whether the effect is one of scale or of recipe. Sun et al. (2026) report from gradient space that depth-to-width allocation affects generalisation at small budgets; on our data b_η has an interval spanning zero, so this is a shape effect we can neither confirm nor exclude, and we say so rather than claiming agreement.

Failure detection from internal state. Semantic-entropy methods detect confabulation by sampling a model repeatedly (Farquhar et al., 2024; Kuhn et al., 2023); activation probes read internal state directly (Azaria & Mitchell, 2023; Chen et al., 2024). A parallel line reads uncertainty from a single forward pass rather than from repeated sampling — from attention-head divergence (van Dijk, 2026), from logit-level evidence (Ma et al., 2025a), from a model's own estimate that a claim is true (Kadavath et al., 2022), or from an energy over semantic clusters (Ma et al., 2025b). Our comparison class is deliberately narrower than any of these: the single-pass probability baselines computed from the *same* forward pass as the geometry, so what is measured is the marginal value of geometric information rather than the value of a second inference pass. This is why our null is a statement about increment and not about detectability, and it is consistent with a systematic finding that the association between uncertainty estimators and hallucination is itself variable and often weak (Agnimo et al., 2026). Alansari et al. (2026) report that internal-state hallucination signals transfer across languages and domains for most models, but conditionally — cross-lingual performance depends on class separability and feature-space alignment, and cross-domain transfer varies with the datasets used. It is that conditional, rather than universal, transfer that touches our finding that discriminative signatures align within a family and task; we do not read their result as evidence that such signals fail to transfer.

Pre-registration and reproducibility. We adopt a protocol frozen before the confirmatory runs so that a null is evidence rather than a disappointment, following arguments for pre-registration in empirical machine learning (Hofman et al., 2023), and against the background of the field's reproducibility efforts (Pineau et al., 2020; Semmelrock et al., 2023). The frozen protocol, the deviations from it, and the adjudication of each pre-registered prediction are reported in full rather than summarised.

B Extended methods

Labeling rules GSM8K: numeric match at 10^{-4} relative tolerance, taking the first number after the last occurrence of "answer is" and falling back to the last number in the completion. TriviaQA: normalized alias containment, lower-casing, stripping punctuation and removing articles. For base models, which continue with fabricated further question–answer pairs, only the answer to the asked question is scored. Both rules were fixed before the confirmatory runs.

What the last recorded transition is HuggingFace appends to `hidden_states` before applying each decoder layer and applies the final normalisation after the loop, so the tuple is $(\text{embeddings}, h_1, \ldots, h_{L-1}, \text{norm}(h_L))$ and the last adjacent pair spans a decoder block *and* the final LayerNorm/RMSNorm. Earlier runs treated it as an ordinary block. A forward pre-hook on the normalisation module recovers the state after the last decoder layer and before normalisation, which makes that transition a genuine block and isolates the normalisation as a separate quantity.

The normalisation turns out to rotate very little — median 0.003 to 0.061 across the grid, against 0.20 to 0.79 for the last block itself — while changing the norm a great deal, which is what a normalisation should do. The correction is correspondingly modest: window levels fall by 2–23% and the fit moves from $\beta = -0.963$, $\gamma = -0.674$, $R^2 = 0.968$ to $\beta = -1.206$, $\gamma = -0.515$, $R^2 = 0.973$. The shift in β is 0.24, smaller than the corrected fit's leave-one-model-out β spread of 0.46, and R^2 improves slightly.

Which figures rest on the corrected measurement, and which do not, is stated per analysis rather than inferred from which models an analysis uses, because the same Pythia grid and the two extrapolation models appear in both corrected and raw analyses. The main-text capacity analysis — the headline exponents, the nested model comparison, the identifiability analysis, and the in-family extrapolation as reported there — uses the corrected measurement, obtained by re-running the probe on the fit grid and the two extrapolation models. Several historical analyses in this appendix are raw and are labelled as such where they appear: the matched-scale control uses the raw exponents and the raw extrapolation runs (Supplementary Note 4), and the pre-training seed floor (Supplementary Note 7) and the training-intensity coefficient predate the correction and are raw. Among the out-of-family runs the convention is likewise not uniform: as listed with the model set in the main text, Qwen3 and SmolLM2 are on the corrected convention while Qwen2.5 is raw. Convention is therefore reported for each analysis rather than for a model set, and we do not claim a blanket immunity to the correction. The two places where a corrected quantity is set against a raw one are the cross-recipe coefficient comparison of Supplementary Note 5 and the joint architecture-and-training fit; both are flagged where they appear, and in both the conclusion is unchanged when the comparison is repeated with every arm measured raw.

We note the failure mode rather than only its fix. The contaminated index passed every check that had been applied: rotation stayed inside $[0, 2]$, nothing was non-finite, no shape was wrong, and the distribution was unremarkable. What was wrong was semantic, one index denoting a different object from its neighbours, and only inspecting the raw per-layer profile reveals it — the first and last entries are visibly discontinuous with the interior in every model measured. Summarising a profile before plotting it is what allowed this to persist.

An internal consistency check on the measurement chain Every quantity this paper reports is derived from hidden states, so an independent check on those derivations is worth more than a restatement of them. One is available in closed form, because the residual stream is additive and both block topologies in our archive write the same total. Writing $\Delta_l = h_l - h_{l-1}$, the parallel form of GPT-NeoX computes $h_l = h_{l-1} + \Delta_{\text{attn}} + \Delta_{\text{mlp}}$ from a common normalised input, while the sequential Llama form computes $h' = h_{l-1} + \Delta_{\text{attn}}$ and then $h_l = h' + \Delta_{\text{mlp}}$; in both cases $\Delta_l = \Delta_{\text{attn}} + \Delta_{\text{mlp}}$ exactly.

Define the relative write magnitudes

$$\rho = \frac{\|\Delta_l\|}{\|h_{l-1}\|}, \qquad \rho_a = \frac{\|\Delta_{\text{attn}}\|}{\|h_{l-1}\|}, \qquad \rho_m = \frac{\|\Delta_{\text{mlp}}\|}{\|h_{l-1}\|}, \qquad r = \frac{\|h_l\|}{\|h_{l-1}\|} = e^{\Delta \ln \|h\|}. \tag{6}$$

Expanding $\|\Delta_l\|^2$ two different ways gives two expressions for the same number. Using only the states, $\|h_l - h_{l-1}\|^2 = \|h_l\|^2 + \|h_{l-1}\|^2 - 2\langle h_l, h_{l-1}\rangle$, and substituting $\cos(h_l, h_{l-1}) = 1 - R_{\text{depth}}$ yields $\rho^2 = (r-1)^2 + 2rR_{\text{depth}}$. Using only the two submodule outputs gives the law of cosines. Equating them,

$$\underbrace{(r-1)^2 + 2r\,R_{\text{depth}}}_{\text{states only}} \;=\; \rho^2 \;=\; \underbrace{\rho_a^2 + \rho_m^2 + 2\rho_a\rho_m \cos(\Delta_{\text{attn}}, \Delta_{\text{mlp}})}_{\text{submodule hooks only}}. \tag{7}$$

equation 7 must hold at every layer of every sample of every model. Its two sides are read from different recorded series — the left from the rotation and norm-growth series this paper reports, the right from forward hooks on the attention and feed-forward modules — but both are computed from

the same forward pass through the same implementation, so they are not independent measurement chains and their agreement does not cross-validate one against the other. It is a consistency check with no free parameters: over the layers and samples where it is evaluated it confirms that the reported series and the submodule hooks decompose the same residual write, and a residual would flag a discrepancy between them without, on its own, localizing which of a shared state, a shared implementation, or a genuine instrument fault produced it.

Measured over interior layers — indices 1 through $L - 2$, dropping index 0 and the last recorded transition for the reasons given in Sec. B — the identity holds to the limit of float32 accumulation. Across seven models and 359,346 layer–sample observations the median relative residual on ρ lies between -1.5×10^{-5} and $+1.8 \times 10^{-6}$, with 5th and 95th percentiles inside $\pm 0.05\%$ for every model. For the two Pythia models whose per-layer trajectories were retrieved in full, 70M and 1.4B, this rests on 18,944 observations; the remaining four Pythia trajectory archives were lost to a Kaggle versioning error described in Sec. 6, and only their per-sample summaries survive, which the identity cannot use.

The same identity fails, and fails only, at the last recorded transition. The branch hooks observe the attention and feed-forward submodules and nothing else, so if that transition bundles a block with the final normalisation, the right side of equation 7 is missing a term the left side contains. The measured residuals there are $+491.8\%$ (Pythia-70M), $+287.7\%$ (Qwen2.5-0.5B), $+278.2\%$ (TinyLlama-1.1B) and $+10.0\%$ (Pythia-1.4B), between 3.4×10^3 and 2.0×10^7 times each model's interior residual. This agrees with the contamination that Sec. B identifies by inspecting per-layer profiles: a second route, algebraic rather than visual, points to the same final-transition anomaly. Because both routes read the same stored trajectories, the agreement corroborates the correction there rather than establishing it independently.

Metric variant: mean-centring Adjacent-layer cosine has been measured elsewhere after mean-centring (Lawson et al., 2025). Cosine is already invariant to positive rescaling, so centring does not remove a norm effect from it; subtracting a per-layer mean moves the reference point and so changes the directions being compared. We do not centre, and treat norm growth as a separate axis. Because the archived trajectories store only derived scalars, the variant cannot be recomputed offline, so the Pythia grid was regenerated with the raw per-layer states cached, and both variants were computed from the same forward passes. The per-layer mean is taken over every cached generation step of a model, and the aggregation order matches the main fit — one window mean per sample, then median across samples — while the window and the (raw) convention remain those of this historical comparison, not the main analysis. The comparison is gated on the raw arm reproducing the fit obtained under the same measurement convention, which it does ($\alpha = 6.919$, $\beta = -0.963$, $\gamma = -0.674$, $R^2 = 0.968$). That run predates the pre-norm correction of Sec. B, so both of its arms carry the old convention; the centred-versus-raw contrast below is therefore made entirely within that old convention, on identical models and items, and its magnitude is a property of that convention rather than a claim about how either arm would move under the correction. The absolute exponents below are the uncorrected ones and should not be read against the corrected fit.

Centring shifts the exponents by a statistically clear amount. The centred fit gives $\beta = -0.886$ ($[-0.891, -0.882]$) against -0.963 ($[-0.970, -0.957]$) raw, and $\gamma = -0.718$ ($[-0.724, -0.713]$) against -0.674 ($[-0.682, -0.665]$), with R^2 falling from 0.968 to 0.937. Neither interval overlaps its counterpart, so the fit is formally conditional on the choice of whether to centre, and both variants are reported. For reference we record two perturbation magnitudes under this same historical, raw convention: swapping one model out of the *raw* fit set moves β by 0.35 (the raw leave-one-out spread), while centring moves it by 0.077. These are descriptive magnitudes obtained under different perturbations, and their ratio is reported as such, not as a measure of scientific importance. In both variants the width coefficient (the coefficient on $\ln d$) is negative — $\beta = -0.963$ raw, -0.886 centred — an observation we report under the old convention without restating it as a test of width-independence or extending it to the corrected measurement.

Level-versus-shape test for the within-family reversal To separate a change in the level of the depth profile from a change in its shape, per-sample R_{depth} matrices are averaged over generated steps, then over samples, giving one profile per model, which is resampled by linear interpolation onto a common 32-point fractional-depth grid. Shape is defined as that profile divided by its own mean, which removes the level exactly. The decision rule was fixed before the analysis: if the 7B

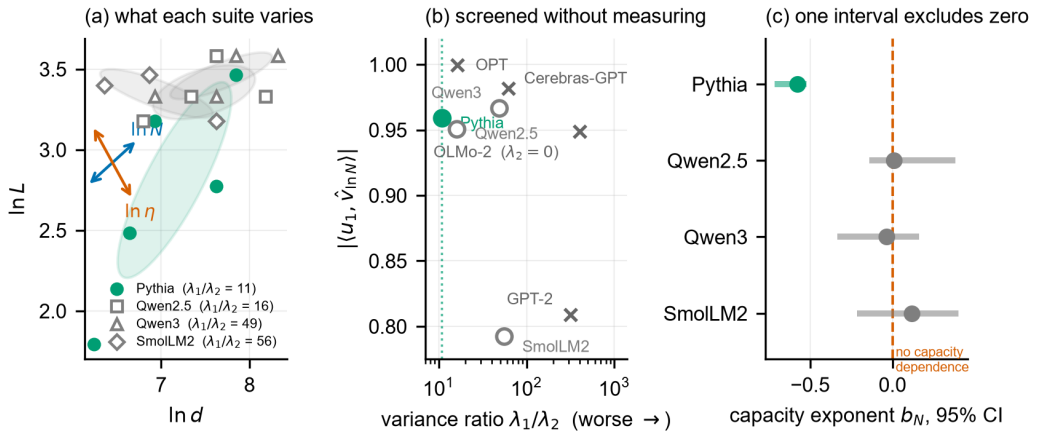

Figure 1: A suite's design decides what can be measured in it. **(a)** Released sizes in $(\ln d, \ln L)$; each ellipse is the centred cross-product $S = V\Lambda V^\top$ at one standard deviation per eigenvector, so its axis-length ratio is $\sqrt{\lambda_1/\lambda_2}$ (the square root of the variance ratio λ_1/λ_2) and its orientation the determined direction. Pythia alone is tilted, because a 1B model at L=16 was released alongside a 1.4B at L=24; the others are thin slabs that vary width at nearly fixed depth. The architectural coordinates are fixed directions in the same plane. **(b)** Every open multi-size suite in the same coordinates, including four never measured here, since the screen requires no response variable. Filled markers are suites measured in this paper; crosses are design only. Pythia is the best conditioned available; the suites built expressly for scaling work lie furthest right, and OLMo-2 has $\lambda_2 = 0$ exactly. **(c)** Under the historical include-final, single-variable, model-point bootstrap, only Pythia's capacity-slope interval does not cross zero. The coefficient is identifiable in every family, and whether an interval crosses zero depends on both effect size and uncertainty, not on a precision rating alone; this panel's intervals are not the conditional OLS intervals of the main capacity model. Levels, by contrast, are estimated in all four families, because a mean needs no leverage.

shape lies within the spread the family already shows among its smaller members, the reversal is a level effect; if it lies outside, a shape change is implicated. Uncertainty is a nonparametric bootstrap over samples within each model (B=2000), recomputing the mean profile and every pairwise comparison on each replicate. The analysis uses only archived trajectories; no additional generation was performed.

Each comparison statistic is accompanied by a descriptive separation check. The check compares two reference classes: pairs of Qwen models below 7B and Qwen–Pythia pairs. Recipe labels do not supply ground-truth profile classes, so overlap cannot be interpreted as failure to detect a known geometric difference. All three statistics tried here — cosine between mean-normalized profiles, cosine after removing the constant component, and scale-free distances — show overlap between these reference groups. They therefore provide no calibrated family separation for deciding the 7B case, so the level-versus-shape question is reported as unresolved rather than answered. An earlier version of this analysis reported a cosine comparison without a control and concluded that the reversal was a level effect; that conclusion is withdrawn.

Level-based comparisons do not have this problem, and two are used. The (d, L) baseline takes every pair of Pythia models as an instance of ordinary architectural movement under a fixed recipe and records $\mathrm{observed}/\mathrm{predicted}$ level ratio for each, giving the spread against which a Qwen step is read. The depth-localization analysis computes that same excess ratio per fractional-depth point between 7B and 3B, with bootstrap intervals over samples; uniformity is judged by whether per-point intervals are mutually disjoint rather than against an arbitrary threshold on a coefficient of variation, which is what the earlier version used.

C IDENTIFIABILITY AND CHOICE OF BASIS

Choice of basis and what it does and does not decide The transformation between $(\ln d, \ln L)$ and $(\ln N, \ln \eta)$ is exact only if $\ln N$ is exactly $2 \ln d + \ln L$ plus a constant, which holds for the capacity proxy $N_{\text{proxy}} = 12d^2L$ (the standard non-embedding-parameter approximation) and not for the published totals. The embedding matrix dominates at the small end — N_{proxy} is 18.9M for Pythia-70M against a published 70.4M, a factor of 3.7, falling to 1.10 at 2.8B — so the two conventions are different models, not different coordinates, and they give different exponents ($b_N = -0.57$ for the proxy against -0.77 for the published total, both on the corrected measurement under the *include-final* window used by the historical two-variable analysis of Appendix C). The frozen-window one-variable capacity model that is the main text's headline gives -0.744 (conditional SE 0.093, Section 2.4). Any statement about a parameter-count exponent has to say which convention *and window* it uses; all are reported, and the sign agreement across recipes holds under either convention.

The identifiability comparison in Supplementary Note 5 uses the perturbations described elsewhere in these Methods rather than an analytic condition number, because the relevant question is empirical: how far does a coefficient move when the fit set changes in a way it plausibly could have. We note explicitly that the greater *precision* of b_N is a property of the model suites that exist, which vary parameter count far more than aspect ratio, and not a claim that capacity is more physically fundamental than shape. Under full column rank both coefficients are identifiable; they differ in estimation precision, not in identifiability (Supplementary Note 9). A suite built to vary aspect ratio at fixed N_{proxy} would instead hold $\ln N$ constant, so the capacity slope would confound with the intercept — a rank failure that leaves b_N unestimable while b_η becomes the estimable slope, not a symmetric swap of precision. What follows from the present measurement is narrower and sufficient: β and γ reported separately are two imprecisely estimated numbers where one combination is well determined, and the sign difference in β between two such fits alone does not establish an architectural or training-recipe effect.

Finally, on the temptation of a round number: there is no sufficient basis to fix the exponent at $-1/2$, but the conditional intervals do not exclude it either. The 95% conditional t-intervals (six points, 4 degrees of freedom) are $[-1.00, -0.49]$ under the frozen window (one-variable slope $\hat{b} = -0.744$) and $[-0.71, -0.44]$ under include-final (-0.576), and both contain -0.5. The fit-set perturbation range $[-0.70, -0.54]$ is a stability check, not a confidence interval, and a point estimate lying away from -0.5 does not make -0.5 statistically excluded. We therefore report that no inverse-square-root reading is *established*, not that it is ruled out.

Design-matrix spectrum and the identified direction For the six-model grid the centred cross-product has eigenvalues 3.5895 and 0.3278, a variance ratio of 10.95, and, for the coefficient combinations (Supplementary Note 9), the capacity combination lies 2.0° from the best-determined direction and the shape combination 16.4° from the worst. Screening every open multi-size suite the same way, without any measurement, returns ratios of 16 to infinity and places Pythia first. Full treatment in Supplementary Note 9.

What the spectrum does and does not license for prediction Rotating into the eigen-coordinates cannot change a prediction, verified at 2×10^{-15}, because least squares is equivariant under invertible reparameterisation. Truncation and shrinkage are different models and do change predictions. In the historical corrected, include-final analysis, their apparent held-out gains are sensitive to the selected penalty and the evaluation summary and do not identify a mechanism. Full treatment in Supplementary Note 10.

A closed-form description of a family The form $\ln \bar{R} = \alpha + \beta \ln d + \gamma \ln L$ is the right one for adjudicating the architectural hypotheses, but a poor one for describing a family, because two of its three parameters mix a precisely with an imprecisely estimated direction and the equation does not show it. Rewriting the same fit in the family's own centred, identified coordinates repairs this without changing anything fitted. With $N = 12d^2L$, $\eta = d/L$, and N_0, η_0 the geometric means over the family's own models,

$$\ln \bar{R}(N, \eta) \;=\; a_0 \;+\; b_N \ln \frac{N}{N_0} \;+\; b_\eta \ln \frac{\eta}{\eta_0}, \qquad a_0 = \frac{1}{n} \sum_i \ln \bar{R}_i. \tag{8}$$

Every quantity is a closed-form least-squares functional of the family's own measurements; nothing is searched or tuned.

Two properties make equation 8 the appropriate description. Centring makes the level parameter exactly orthogonal to both slopes — we verify $|\mathrm{cov}(a_0, b)|/\mathrm{var}(a_0) < 10^{-13}$ against 0.13 and 0.15 in the uncentred form — so a_0 is determined independently of how badly b_η is determined, whereas the uncentred intercept inherits every slope's instability through the lever arm out to $N = \eta = 1$. And a_0 is the level at a point inside the family's measured range rather than far outside it. The gain is not cosmetic: under leave-one-model-out the level parameter moves by 0.19 in the centred form against 2.69 uncentred for Pythia, and 0.10 against 3.95 for Qwen2.5, factors of 14 and 38.

The family signature is then the pair (a_0, b_N) attached to its reference point, with b_η reported and flagged: The width of Pythia's b_η interval, spanning 2.2 in a coefficient of order 0.1, is the *precision* (interval-width) result of Supplementary Note 5 stated as a number — b_η is identifiable, only imprecise. That Qwen2.5's is narrower is not a contradiction but the expected consequence of a different design, and should be read with its one residual degree of freedom in mind.

Comparing two families requires evaluating both at a common (N, η), and this is admissible only where both measured ranges contain that point — otherwise the comparison moves along a direction neither design determines. We therefore report the overlap explicitly rather than assuming it: the families span $N \in [1.9\times10^7, 2.5\times10^9]$ and $[2.3\times10^8, 4.3\times10^9]$, and $\eta \in [42.7, 128]$ and $[37.3, 128]$, so a reference point at $N^* = 7.6\times10^8$, $\eta^* = 73.9$ is interpolation for each. The reference-point comparisons of this section's historical centred fit are made there; the direct matched-pair ratio ($1.60\times$) and the width-matched anchors of the matched comparison in Section 3 are separate, fit-free comparisons not evaluated at this point.

Computing the penalty instead of searching for it In eigen-coordinates the ridge estimator is $\hat{b}_j^{\mathrm{ridge}} = [\lambda_j/(\lambda_j + \alpha)]\hat{b}_j$. Under an isotropic Gaussian coefficient prior the Bayes relation is $\alpha = \sigma^2/\tau^2$; Hoerl–Kennard–Baldwin and empirical-Bayes estimates are reported under their stated assumptions. Both are equivariant under rescaling of the log predictors; Lawless–Wang is not in the present raw-coordinate implementation and is withdrawn. Full treatment in Supplementary Note 11.

A historical two-variable reference frame This analysis uses the *historical include-final two-variable fit* $\ln \bar{R} = 8.107 - 1.206 \ln d - 0.515 \ln L$, not the frozen-window one-variable capacity model of the main text; its offsets, interval widths and the figure below are properties of that earlier fit. The fitted surface serves as a stated baseline against which other families are placed by their offset,

$$\Delta(\mathrm{model}) = \ln \bar{R}_{\mathrm{measured}} - \ln \bar{R}_{\mathrm{Pythia}}(d, L), \tag{9}$$

read at that model's own (d, L); we do not read it as the position of a "recipe" or as evidence that architecture has been controlled, since (d, L) fixes only width and depth.

These intervals come from a bootstrap over the six Pythia grid points (4000 resample attempts with replacement, 3915 retained after discarding draws with fewer than three distinct design points; on this fixed grid every three-point subset is full rank, so the screen removes only degenerate resamples, and each retained draw refits the two-variable surface); a target's interval is the 2.5–97.5 percentile of $\ln \bar{R}_{\mathrm{measured}} - \hat{\beta}^{(b)} \cdot x$ with the target's own measured level held fixed. They therefore reflect only sampling of the fit grid — not the target's measurement uncertainty, cross-seed dispersion, or model selection — and are not the conditional OLS standard error of the main capacity model (Section 2.4). So computed, the interval widths are 0.32–0.51 for models inside the calibrated range, 0.72–1.06 for those beyond it in capacity (the direction the design determines well, a cost of $1.9\times$), and 1.25–2.74 for those beyond it in aspect ratio (the direction it determines worst, $4.2\times$). Across the models checked here, ordering by displacement in $\ln \eta$ recovers the widths monotonically while displacement in $\ln N$ does not — at these points it is distance in the poorly determined direction, not distance from the grid, that widens the interval; we state this for the points examined and do not extend it to arbitrary displacements. This is D1 made operational; the offsets and the spectrum share the same data, so it is a consistency check rather than a fully independent one, but the spectrum did name which direction would be expensive from the design alone, before any of these offsets existed.

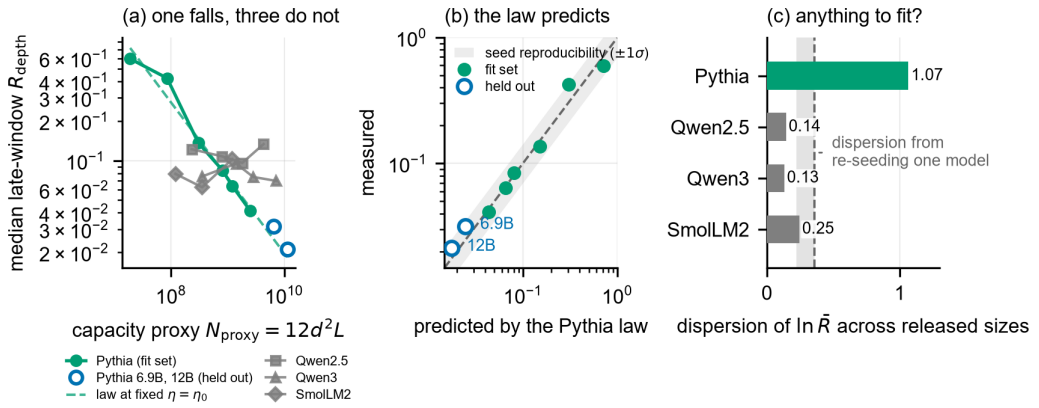

Figure 2: **Historical include-final, two-variable analysis, retained as supporting context**; the headline capacity result uses the frozen window and the one-variable model of Table 1 in the main text, not this figure. Levels here use the include-final window $\{\lfloor 0.6L \rfloor, \ldots, L-1\}$ and the dashed line is the earlier two-variable fit $\ln \bar{R} = \alpha + \beta \ln d + \gamma \ln L$ drawn along the grid's geometric-mean aspect ratio. **(a)** Median late-window R_{depth} against the capacity proxy $N_{\text{proxy}} = 12d^2L$. Pythia (green) falls monotonically across $133\times$ in N; the held-out 6.9B and 12B (open circles) sit above this two-variable line — under the frozen window their residuals are instead negative (main text Section 2.3). Qwen2.5, Qwen3 and SmolLM2 (grey) are flat and non-monotone. **(b)** Measured against predicted for Pythia under this two-variable fit; the band is the $\pm 1\sigma$ re-seed dispersion of one 410M model. **(c)** Dispersion of $\ln \bar{R}$ across each family's released sizes against that same seed dispersion (dashed line; the band spans the two dispersion conventions of D2). Only Pythia's dispersion exceeds this single-size reference.

D NOISE FLOORS AND ROBUSTNESS

Pre-training seed floor PolyPythias provides nine independently seeded pre-training runs at each of 14M, 31M, 70M, 160M and 410M, identical in architecture and token budget (van der Wal et al., 2025). We measure the three sizes that overlap the fit grid under the protocol used everywhere else, and summarise each size by the deviation of its per-seed median level from the mean across seeds. The bootstrap resamples *seeds*, not items, because the spread across seeds is the quantity of interest; resampling items would measure how precisely an individual seed's level was pinned down, which is a different question. Nine seeds give a wide interval, and that is the honest width. The impact on the fit is assessed by substituting each seed's level into the six-model fit in turn and recording the resulting spread in the exponents.

Three properties of the resulting number are reported rather than smoothed over. First, the summary statistic does not drive it: at 410M the floor is 31.5% by median absolute deviation and 27.3% by mean, against a depth-axis transport residual of 39.7% and 37.2% under the same two conventions, giving ratios of 0.79 and 0.73. Comparisons in the main text use the median throughout, matching the convention the transport residual is reported under. Second, the 410M floor is driven substantially by two of the nine runs, which reach TriviaQA accuracies of 0.015 and 0.035 against 0.08–0.12 for the others; per-seed level and per-seed accuracy correlate at $+0.69$, and excluding those two takes the median floor to 17.3%. Since PolyPythias documents outlier runs, a case could be made for excluding them, but the criterion would have been chosen after seeing the data, so we report both figures and exclude neither. Third, a two-parameter line fitted to three sizes leaves only one residual degree of freedom, so uncertainty inference is extremely fragile. A two-point extrapolation made before the 410M run predicted 12.7% where the measurement gave 31.5%; the measured values are reported and the floor is not extrapolated above 410M.

Training-intensity test Within-lineage training dependence is estimated by regressing $\ln \bar{R}_{\text{depth}}$ on $\ln D$ across public Pythia checkpoints at fixed architecture (410M and 1B, at 29/75/151/300B tokens). Two lineages give $\delta = \mathrm{d} \ln \bar{R} / \mathrm{d} \ln D = +0.558$ and $+0.322$: within a fixed architecture,

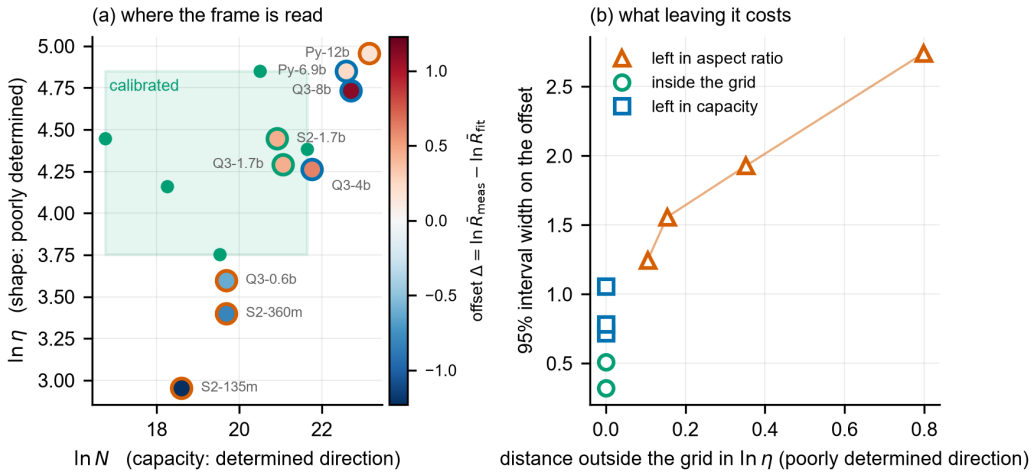

Figure 3: **Historical two-variable design-sensitivity analysis** (include-final window, fit $\ln \bar{R} = \alpha + \beta \ln d + \gamma \ln L$), not the frozen-window one-variable capacity model. A reference frame is only as good as where it is read. **(a)** The design plane. The shaded rectangle is the range this fit was calibrated on; every other model is a point in the same coordinates, with fill colour giving its measured offset Δ from equation 9 and edge colour showing which boundary, if any, it crosses. Leaving rightwards crosses the capacity edge, which the design determines well; leaving vertically crosses the aspect-ratio edge, which it does not. **(b)** What each crossing costs. Interval width against displacement in $\ln \eta$: the four models that leave in aspect ratio (triangles) widen monotonically with it, while those that leave only in capacity (squares) stay near the calibrated width across the displacements shown here. D1 anticipated this asymmetry from the design alone.

more training raises late-layer rotation. The fit grid holds D constant, so $\ln \mu = \ln D - \ln N$ is perfectly anti-correlated with $\ln N$ inside it (raw-scale Pearson -0.66); with published parameter counts μ carries little independent variation there, so the grid gives δ little leverage. Adding the checkpoints, which vary D at fixed architecture, the joint fit of $\ln \bar{R} = a + b_0 \ln d + c_0 \ln L + \delta \ln \mu$ returns $\delta = +0.423$ $[+0.230, +0.651]$, consistent with both direct estimates and robust to the measurement convention (fitting the grid raw gives $+0.478$).

The remaining coefficients are not reportable as architecture-only exponents, and we say so because a positive reading is tempting: $b_0 = +0.05$ with an interval spanning zero invites the conclusion that the width dependence is mostly training intensity in disguise. We decline that reading, but not on an algebraic-inconsistency argument. The grid and the joint fit use different observation sets and different parameterisations — the joint fit is on published parameter counts, for which $N \propto d^2 L$ holds only approximately — so their coefficients need not coincide, and the would-be identity $b = b_0 - 2\delta$, $c = c_0 - \delta$, which assumes the exact relation, does not apply here. We report δ and decline to read b_0 and c_0 as architecture-only exponents.

Prompt-format and post-training ablation The scaling grid uses base models with plain prompts; the Qwen2.5 out-of-family runs use instruction-tuned weights with chat templates (Qwen3 and SmolLM2 are base weights with plain prompting, like the grid). On Qwen2.5-0.5B two confounds are therefore separable and were measured under the historical include-final convention: switching only the prompt format at fixed weights and items isolates the format effect ($\approx 1.5\%$ of the depth-axis level), and switching instruction-tuned for base weights of the same architecture at fixed format isolates the post-training effect ($\approx 0.9\%$). Both are small against the 22–42% deviations they might otherwise have confounded, but each is a single local measurement, not a bound.

E Supporting Analyses

Depth dependence: consistent in direction, unresolved in size Fitting the relation separately at each fractional depth gives different precision for different coefficient combinations: the deep-

minus-shallow contrast in b_N is -0.61 with a bootstrap interval of $[-1.27, -0.24]$, while the corresponding contrast in the raw width coefficient β has an interval spanning zero. Thus the capacity combination shows a resolved depth contrast in this analysis, whereas the width-coefficient contrast remains uncertain. Full treatment in Supplementary Note 1.

The time axis refuses the same form The temporal displacement axis does not admit the form that fits the depth axis: R^2 falls to 0.62 and an intercept-only model wins the nested leave-one-out comparison. We report this as an empirical result under its own protocol, not as a prediction the functional form was constructed to make. Full treatment in Supplementary Note 2.

At matched scale, two family steps move in different directions Comparing single steps rather than fitted trends (a *raw*, historical analysis), Qwen2.5's observed/predicted step from 3B to 7B is $2.01\times$ — about $1.21\times$ the within-recipe baseline's maximum deviation of $1.66\times$ — while the matched Pythia step falls and sits inside that baseline. The two steps move in opposite directions, so this reversal is not a scale trend shared by these models; adaptation and prompting also differ, so the comparison does not localize a cause. Full treatment in Supplementary Note 4.

Raw width and capacity-combination coefficients answer different questions The width coefficient changes sign between recipes in (d, L), whereas the capacity combination $b_N = (\beta + \gamma)/3$ remains negative in the fits examined. These are different estimands under an exact change of basis: sign agreement for b_N neither erases nor validates the sign contrast in β. Full treatment in Supplementary Note 5.

The transport residual: two analyses give different associations Carrying the fit across recipes leaves a residual that correlates with training-data volume, but a within-lineage checkpoint analysis gives the opposite sign. The two analyses give different associations in the settings measured, and a single cross-family-transportable training-data coefficient is not supported by current evidence; we do not identify the source. Full treatment in Supplementary Note 6.

How much of the transport residual could be pre-training seed noise Nine independently seeded runs per size at 70M, 160M and 410M give median deviations of 2.6%, 5.5% and 31.5%. At 410M the floor is 0.79 of the cross-family transport residual, so magnitude alone is not diagnostic. This study does not include or audit independent seeds above 410M, so the floor is not extrapolated there. Full treatment in Supplementary Note 7.

Discriminative signatures align within family and task, not universally Feature-importance profiles correlate within a family and task and not across them, and a leave-one-model-out regression from architecture meta-features to profile coefficients beats its nulls only where the family already aligns. We claim the conditional alignment, not a predictive architecture formula. Full treatment in Supplementary Note 8.

F SUPPLEMENTARY NOTES

Supplementary Note 1: Depth dependence of coefficient combinations. Those coefficients summarise a window, so it is natural to ask what the fit looks like at each depth separately. Refitting the same two-variable form at each fractional depth fits well almost everywhere ($R^2 = 0.84$–0.99 across $z \in [0.10, 0.95]$), and the width exponent's point estimates run from about -0.4 at mid-network to about -2.1 in the last quarter.

In $(\ln d, \ln L)$ the point estimates are too imprecise for a resolved trend. Each depth is a six-point fit of three parameters, leaving one model out moves the width exponent by up to 2.16 at a single depth, and the deep-minus-shallow difference of -1.37 carries a bootstrap interval over the model set of $[-4.24, +1.62]$, which contains zero. Neither a reliable depth-resolved width trend nor a critical layer is supported in that basis.

The design is full rank, so the coefficients are identifiable, but different coefficient combinations are estimated with different precision, as Sec. F diagnoses at the window level.

Which direction a grid determines follows from its design. For a least-squares fit the best-determined coefficient combination is the leading eigenvector of the predictor cross-product, and for this grid the two eigenvalues are 3.59 and 0.33: a coefficient along the weak direction carries eleven times the variance of one along the strong direction. The capacity direction sits 2.002° from the strong direction and the shape direction 16.433° from the weak direction. Adding the two extrapolation models raises the condition number from 10.9508 to 15.9497, so conditioning worsens. Under a common homoscedastic, uncorrelated noise model, however, the design-variance factors fall for both combinations: from 0.063 to 0.045 for b_N and from 1.57 to 1.28 for b_η. Adding observations can therefore improve both variances without improving their ratio. β and γ each mix the two axes and therefore inherit uncertainty from both design directions. In the observed perturbations, the per-depth leave-one-model-out width falls from a median of 0.55 for β to 0.16 for b_N, and the deep-minus-shallow contrast in b_N is -0.61 with a bootstrap interval of $[-1.27, -0.24]$, excluding zero, whereas the interval for the corresponding contrast in β contains zero. With relative write magnitude ρ in place of rotation, the b_N contrast is -0.28 with an interval of $[-0.68, -0.11]$, while the sign of the $(\ln d, \ln L)$ version does not survive leave-one-out.

Thus, under this fitted specification, the capacity combination shows a resolved depth contrast: per-block rotation falls more steeply with N_{proxy} deeper in the network. The corresponding contrast in the raw width coefficient remains uncertain. These are different estimands with different precision, not the same quantity varying in one basis and not in another. Results for other model sets are reported separately: predictor spans alone do not establish that an analogous contrast must succeed or fail. Four Qwen2.5 models fitted with three parameters leave little residual information, so that result is weak, but detection is not mathematically impossible.

Three perturbation schemes give different spans for the fitted coefficients, and we report them as descriptive spans, not as a variance decomposition or a ranking of uncertainty sources. Resampling items within each model and refitting propagates item-sampling noise to the coefficients under a fixed model set, giving $\beta \in [-1.213, -1.199]$ and $\gamma \in [-0.523, -0.506]$ (item bootstrap resampling each model's answers independently with replacement, not paired by item, so it does not preserve cross-model correlation from the shared prompt set; $B = 2000$, seed 0, corrected include-final window, 95% percentile intervals; archived). Refitting with one model left out at a time gives $\beta \in [-1.51, -1.05]$ and $\gamma \in [-0.88, -0.18]$; this is a leave-one-out max-to-min span, not a bootstrap percentile interval, so the two are not on a common scale and we do not read their ratio as a magnitude comparison. Substituting each raw 410M seed level in turn into the raw include-final six-point grid (Sec. F) spans β across 0.86 and γ across 0.96. Across all examined refits and seed substitutions β stays negative; we report this as a sensitivity description, not a formal test, and do not generalise it beyond the cases examined. The numerical values should be read as order-of-magnitude statements about the exponents rather than as measurements at the precision a single fit reports.

The following figures are historical: this is an earlier candidate-model comparison under the *corrected* measurement and the *include-final* window, using a median relative-error metric $\mathrm{median}_m\,|\hat{R}_m/R_m - 1|$ (the out-of-range figures below are signed relative errors with the same measured denominator), and it is superseded by the frozen-window comparison of the main text (Sections 2.3 and 2.4), which we do not restate here. Under that historical setting, leave-one-out gave the two-variable model 23% median error against $\ln d$-only (24%), $\ln L$-only (52%) and intercept-only (83%); a one-variable capacity fit $\ln \bar{R}_{\mathrm{depth}} = 9.33 - 0.576 \ln N$ with $N = 12d^2L$ ($R^2 = 0.973$, exponent range $[-0.71, -0.52]$ over fit-set perturbations) predicted held-out grid members at 8.5% median error against the two-variable model's 23.1% (capacity proxy; 13.5% on published totals), and out of range at 6.9B and 12B underpredicted by 20.3% and 14.2% against 22.1% and 16.5%.

We do *not* carry the earlier note's conclusion that "the simpler law is at least as good on both counts" or its decision to "report the two-variable form throughout". The current main analysis reports the one-variable capacity model as the headline, with the two-variable form kept only as a nested comparison and sensitivity analysis; on the frozen window the capacity model has the lower mean leave-one-out error, adding the shape term gives no incremental mean-LOO evidence (which is not evidence that shape has no effect), and capacity+shape is in fact marginally better on the two historical extrapolation points, with the shape coefficient's interval spanning zero. The withdrawn rotation-budget reading of $\gamma \approx -2$ is no longer used to justify the functional form.

An earlier version of this comparison reported the reverse conclusion, on figures of 35% and 39% against 17% and 18%. Those two pairs were measured under different conventions: the first against

held-out levels from the pre-norm-corrected probe, the second against uncorrected ones. Re-running 6.9B and 12B under the corrected probe, so that both arms share the measurement of Methods, 'What the last recorded transition is', removes the apparent trade-off. We record this because the corrected numbers are individually less favourable to the two-variable form that the earlier analysis used, and a reader is entitled to know that that form was not chosen on predictive grounds.

The fit is not an artifact of a single run. The grid has been regenerated from cold twice — fresh forward passes over the same 200 items, 1,200 generations each, zero extraction errors. The first regeneration, using the original measurement, returned per-model levels agreeing to four decimal places and the exponents to three. The second added the pre-norm correction and is the fit quoted above.

Supplementary Note 2: The time axis refuses the same functional form. Applying the identical procedure to temporal displacement fails. Its depth exponent is *positive* ($\gamma \approx +0.34$ — deeper models move more per token at fixed width), R^2 falls to 0.62, and intercept-only *wins* the nested leave-one-out comparison (18% against 27%). Within every model the two axes also occupy disjoint ranges: late-window R_{depth} concentrates at median 0.090–0.148 while R_{time} concentrates at 0.293–0.481, with per-sample Spearman correlation near zero. The axes are different quantities, and the fit is a property of one of them.

Supplementary Note 3: In-family extrapolation and the out-of-family offset. Applied blind to five instruction-tuned models from other families (Qwen2.5-0.5B/1.5B/3B (Qwen Team, 2024), TinyLlama-1.1B (Zhang et al., 2024), SmolLM2-1.7B (Allal et al., 2025); different architectures, tokenizers, and 10–60$\times$ more training data), the Pythia-fitted trends yield median absolute deviations of 40% (R_{depth}), 22% (R_{time}) and 42% ($\Delta \ln \|h\|$), with a systematic sign: these models drift *more* than the Pile-trained fit predicts.

These historical figures evaluate a Pythia-fitted predictor on the five stated instruction-tuned target models under that measurement pipeline; no second target-family fit is required to define the prediction errors. The $40\%/22\%/42\%$ deviations and the target response span do not by themselves separate a level offset from a slope difference or from pipeline differences. Results, 'The same probe on three further families', uses different model sets and is reported separately. We retain the figures as biases of this historical predictor on the specified targets, not as a general estimate of how a scaling law transports across recipes.

In-family the historical include-final two-variable fit extrapolates. Pythia-6.9B and 12B — $2.5\times$ and $4.3\times$ beyond the largest fit point, measured under the same corrected probe as the fit grid — land, under that historical fit, at log residuals of $+0.249$ and $+0.180$ ($1.28\times$ and $1.20\times$ the predicted level), both inside that fit's own leave-one-out band, with levels still declining monotonically. These are the include-final historical numbers, in which the held-out points sit *above* the line; under the frozen-window one-variable capacity model of the main text the same two points instead have negative residuals (Section 2.3), so the sign of the extrapolation error depends on which model and window are used.

Supplementary Note 4: Matched-scale steps into the 7B class. This note is a *raw*, historical analysis: it uses the raw exponents $\beta = -0.963$, $\gamma = -0.674$ and the raw extrapolation runs, and should not be read against the corrected extrapolation above. It compares not two fitted trends but two *steps*, which needs no trend on either side. Qwen2.5's level rises from 3B to 7B where the raw Pythia fit predicts a fall; the corresponding Pythia step (6.9B, same parameter class, a different recipe) falls as predicted and sits inside the within-recipe baseline — the spread of observed-over-predicted ratios across Pythia pairs, reaching $1.66\times$ — while the Qwen step is outside that baseline and moves the other way.

We read this narrowly. The two steps move in different directions, so the reversal is not a scale trend shared by the models examined; it does *not* localize a cause. Both sides here are measured raw, but Qwen2.5 and Pythia still differ in adaptation state and prompting as well as recipe (Section 6), and matched parameter class is not a controlled experiment, so we do not attribute the difference to the training recipe. As a raw-fit reference the predicted 6.9B$\rightarrow$12B step misses by 2%; under the corrected include-final two-variable fit the same step instead misses by about $+7\%$, so this agreement, too, depends on the convention.

Fitting the same functional form across recipes rather than within one shows both what is shared and what is lost. On the six instruction-tuned models the fit is not worse by the usual criteria — $R^2 = 0.958$, and its leave-one-out predictive R^2 of 0.866 is in fact slightly better than the Pythia grid's 0.823. Those models span a factor of 1.5 in R_{depth} against the Pythia grid's 13.4; this range difference complicates comparison of the two R^2 values but does not explain their ordering by itself.

A prediction made and half overturned. Between 0.5B and 3B, detection AUROC declines in step with the drift levels. We initially invoked an angle law proportional to $d^{-1}L^{-0.7}$ to predict continued attenuation at 7B and beyond; that interpretation is withdrawn because the law was not established and the closure run contradicted it. At 7B, R_{depth} AUROC *recovers* to 0.749 on TriviaQA and drift levels *rise* above the 3B values (median R_{depth} 0.133 against 0.096), while $\Delta\ln\|h\|$ continues to collapse and flips orientation and entropy strengthens monotonically. The durable lesson is sharper than monotone attenuation: feature identity, orientation and level are all scale-dependent in these runs, so a monitor's applicability should be validated on its intended model and task before transfer is assumed.

Local comparisons bound the 7B reversal but do not identify its cause. Three candidate factor families are within-family reversal: an unusual path through (d, L) space, a change in training intensity, or a redistribution of computation across depth. Each is testable on archived trajectories, and the three tests give different verdicts.

Comparison with the within-Pythia fit-error scale. Every pair of Pythia models is an instance of ordinary architectural movement under a fixed recipe, so the spread of $\mathrm{observed}/\mathrm{predicted}$ level ratios over those pairs is the natural baseline, and it is wide: pairwise ratios already deviate from exact by up to $1.66\times$ within the family it was fitted on. The Qwen 3B→7B step deviates by $2.01\times$ ($[1.98, 2.04]$) — outside that baseline, but only by a factor of 1.2, with the earlier Qwen steps inside it. The Qwen step is therefore modestly beyond this descriptive reference scale. This comparison does not decompose the reversal or show how much is caused by the (d, L) path. Its direction differs from the matched Pythia step, and Sec. F isolates that: the corresponding Pythia step moves the level down, as the fit expects, while the Qwen step moves it up.

Training intensity predicts the wrong sign. Within a Pythia lineage more training raises late-layer rotation, $\delta_{\mathrm{within}} = +0.558$ (410M) and $+0.322$ (1B). Qwen2.5 trains every size on one corpus, so tokens per parameter falls monotonically ($36{,}700 \rightarrow 2{,}400$), and carrying the Pythia sign across, 7B's lower intensity predicts *lower* rotation than 3B. The measurement is the opposite with disjoint intervals (0.0959 against 0.1334). The intensity story explains the decline from 0.5B to 3B and fails exactly at the reversal it was invoked for — conditionally, since transporting a coefficient from Pythia to Qwen presumes the cross-family transport this paper shows to be unreliable.

The available profile statistics do not provide calibrated family separation. Cosine between mean-normalized profiles, the same cosine with the constant component removed, and scale-free distances all place at least one Qwen–Pythia pair inside the within-Qwen range. Recipe labels do not establish a ground-truth geometric profile difference, so this overlap neither confirms nor rules out redistribution across depth. What can be localized is where the excess sits: not uniform across depth (416 of 496 depth-point pairs have disjoint bootstrap intervals), running at $2.8\times$ in the first eighth, dipping to $1.6\times$ between a quarter and a half, and settling near $2.1\times$ thereafter — so within the late-depth window our features use, it is close to flat.

Supplementary Note 5: Raw width and capacity-combination coefficients are distinct estimands. In (d, L) the two fits look qualitatively unlike each other: the width exponent changes sign, to $\beta = +0.19$ across recipes and $+0.18$ within Qwen2.5 alone, against -1.21 within Pythia. With the capacity proxy $N = 12d^2L$, $(\ln d, \ln L)$ and $(\ln N, \ln \eta)$ with aspect ratio $\eta = d/L$ are related by an exact linear transformation, under which $b_N = (\beta + \gamma)/3$ and $b_\eta = (\beta - 2\gamma)/3$; the factor $1/3$ is the magnitude of the determinant of the inverse change of basis (the map $[[2, 1], [1, -1]]$ has determinant -3) and carries no empirical content. The fit is unchanged and only the reported coefficients move. The coefficients nevertheless represent different directions in predictor space, so agreement in the sign of b_N cannot negate the observed sign difference in β. In the transformed basis the two fits agree in sign on the capacity combination: per-block rotation, $b_N = -0.57$ within Pythia against -0.26 across recipes.

We do not claim the same for b_η, and the reason is the argument of this section applied to itself. Pythia gives $b_\eta = -0.06$ and the cross-recipe fit $+0.71$, which are opposite in sign; but b_η lies within 16.433° of the direction the Pythia grid determines worst (Methods, 'Design-matrix spectrum and the identified direction'), so its sign there is estimated much less precisely than b_N. The change of basis separates the well-aligned capacity combination from the less precise shape combination; it does not adjudicate the sign of the distinct raw-width estimand.

Two caveats belong with these numbers. The Pythia figures are on the corrected measurement of Methods, 'What the last recorded transition is' while the cross-recipe figures are not, so this particular comparison spans two measurement conventions; the sign of b_N is unchanged under either, but the magnitudes should not be differenced. And under the correction b_η moves from $+0.13$ to -0.06 while b_N moves only from -0.55 to -0.57. This contrast is consistent with the design estimating the shape combination less precisely; it is not a variance decomposition of the measurement correction.

We did not choose that basis in advance. The sign reversal is what prompted us to check whether (d, L) was well conditioned at all, and the check is what follows; the diagnostic value lies in its being available at all, not in having anticipated the answer. Sign agreement in a rewritten basis is not by itself evidence of anything either, so the relevant question is how precisely and stably each compared coefficient combination is estimated. The two perturbations above provide sensitivity checks. Substituting a different 410M pre-training seed moves β by 0.857 and b_N by 0.033; leaving one model out moves them by 0.346 and 0.171. The parameter-count exponent is $26\times$ more stable than the width exponent under the first perturbation and $2\times$ under the second. The reason is visible in the grid: it spans $\ln N$ over 4.9 but $\ln \eta$ over only 1.1, so b_N is estimated more precisely than b_η, and β and γ each mix directions with unequal precision. These stability results favor b_N for the capacity question, but do not turn the sign of a different coefficient into a coordinate artifact.

What survives the change of basis is a disagreement in b_N that is real but narrower than the point estimates suggest, and that is not an artifact of pooling recipes. Bootstrapping over models rather than generations gives -0.574, $[-0.761, -0.501]$ within Pythia. Fitting Qwen2.5 alone — a single recipe, four models — gives -0.252, $[-0.482, -0.146]$, and the three-recipe set gives -0.260, $[-0.441, -0.105]$: the pooled and the pure estimates agree to 0.008, so mixing recipes is not what produces the difference from Pythia. The intervals against Pythia do not overlap, but they miss by 0.02 for the pure family and 0.06 for the pooled one, which is a marginal separation rather than a comfortable one, and the pure comparison rests on four models fitting three parameters.

The families also differ in level, and saying by how much requires more care than comparing fitted intercepts. Those intercepts — $+9.53$ for Pythia against $+0.25$ for Qwen2.5, a factor of 10^4 — are levels at $N = \eta = 1$, a point many orders of magnitude outside every model measured, and they are the least stable parameter in the fit: their leave-one-out spread is $38\times$ that of the level at a family's own centroid, and they are correlated with both slopes. Evaluated instead at a reference point interior to both designs (Methods, 'A closed-form description of a family'), the level difference is a factor of 1.63, with a model-level interval of $[1.08, 1.95]$ that excludes unity. This $1.63\times$ comparison is far smaller than the inappropriate intercept comparison.

Nor do the families separate more clearly in level than in scaling, once the two are put on one footing. A slope difference is not comparable to a level difference until it is integrated over the range where both families exist: across the overlapping span of 2.39 in $\ln N$, the difference in b_N of 0.322 produces a swing of ± 0.384 in $\ln \bar{R}$ about the centre, against a level offset of 0.486. These are the same order. What can be said is narrower than a ranking: the families differ in both, comparably, and the direction of the capacity effect is shared across every recipe measured. This appendix uses the (d, L) form where the architectural hypotheses of Methods, 'Why the log–log–linear functional form' are stated, and the identified-coordinate form of Methods, 'A closed-form description of a family' wherever two fits are compared; the main text's headline is the one-variable capacity model, and the two-variable (d, L) form is a comparison and sensitivity analysis, not the reported result.

Supplementary Note 6: What the transport residual tracks. Out-of-family residuals correlate with training-data volume ($r = -0.49$ to -0.86 across axes), which invites a missing-$\ln D$ reading. Fitting that extension with β, γ frozen at their Pythia values gives strikingly consistent estimates across all three axes, $\hat{\delta} \in [-0.32, -0.28]$ (bootstrap CIs of width ≤ 0.02), with positive family offsets ($c = 0.28$–0.56).

We then estimated δ within lineage, using public Pythia intermediate checkpoints (410M and 1B at 29/75/151/300B tokens; same weights lineage, same d and L; N=1600 new generations, zero errors). This within-lineage estimate has the *opposite sign* and is axis-heterogeneous: per-block rotation *grows* with training ($\delta_{\text{within}} = +0.56$ at 410M, $+0.32$ at 1B), temporal displacement is flat ($+0.01/-0.09$), and norm growth is mixed ($+0.14/-0.20$). The cross-family and within-lineage estimates therefore disagree in sign; the uniform negative cross-family value is consistent with absorbing a family-level offset rather than a data-volume dependence, but these data do not identify which.

Two narrower statements are what the data support. Within a lineage at fixed architecture, more training is associated with more late-layer rotation in these checkpoints. And the within-lineage and cross-family analyses do not give a consistent, transportable δ estimate.

A joint fit over both arms — which reuses the checkpoint data, so it is not an independent replication — points the same way. On the grid, training data is held constant, so $\ln \mu = \ln D - \ln N$ is a constant minus $\ln N$: on the log scale $\ln \mu$ and $\ln N$ are perfectly anti-correlated there (raw-scale Pearson -0.66), and μ carries almost no independent variation. Using published parameter counts the grid design is nevertheless full rank (rank 4), because published N departs slightly from $12d^2L$; but that fourth direction is tiny and we do not read it as independently controlling training intensity. The checkpoint set provides training-data variation at fixed architecture; the joint fit over both arms returns $\delta = +0.42$, $[+0.23, +0.65]$, consistent in sign and range with the directly measured $+0.32$ and $+0.56$. Because it reuses the checkpoint data, this is a consistency check, not an independent confirmation.

The joint fit does not, however, license reading its architectural coefficients as data-free versions of the published ones. The two fits use different observation sets and parameterisations — the joint fit is on published counts, for which $N \propto d^2L$ holds only approximately — so their coefficients need not coincide; in particular the joint fit's $b_0 = +0.05$ spans zero. We do not read b_0 and c_0 as a data-free decomposition of the grid fit's exponents. A separate ablation (Methods) measured prompt format and instruction tuning at $\approx 1.5\%$ and $\approx 0.9\%$ of the depth-axis level for the one model where it could be run, which does not bound cross-family or cross-scale effects.

Supplementary Note 7: The pre-training seed floor. Attributing that residual to family identity assumes it exceeds what a different random seed would produce, and each model in the fit is a single pre-training run. PolyPythias releases nine independent seeds per size up to 410M (van der Wal et al., 2025), which makes the floor measurable. It is not constant, and it is not small:

Read against the like-for-like quantity — the depth-axis transport residual of 39.7%, not the 22–42% range that spans three axis statistics — the 410M floor of 31.5% is 0.79 of it, and 0.73 if both are summarised by the mean instead. At 410M, the descriptive seed dispersion is large relative to the corresponding cross-family deviation; this is a magnitude comparison, not a decomposition of that deviation. Two caveats belong with this and are set out in Methods: the floor is driven substantially by two of the nine runs, anomalous in task accuracy as well as geometry, and it is measured at three sizes only, so we do not extrapolate it above 410M.

Seed choice also moves the exponents. Substituting each seed's 410M level in turn into the raw grid moves β across $[-1.49, -0.63]$ and γ across $[-1.04, -0.09]$; we report this as another descriptive span under a distinct perturbation, not as a magnitude comparison against the item-bootstrap or leave-one-out spans, which are different statistics. What survives is qualitative and offered as a sensitivity description rather than a formal test: β stays negative across the examined substitutions. We do not generalise this beyond the cases examined or read it as a test of width-independence. The numerical values are not determined to the precision a single fit suggests.

The floor also puts the fit's own residuals on a scale for the first time (Table 4). At 70M and 160M the fit misses by $7.05\times$ and $6.54\times$ the local descriptive seed dispersion, while at 410M the residual is $0.21\times$ it. These ratios compare magnitudes; the seed floor is not a calibrated rejection threshold and does not classify residuals as structural or random. Above 410M no floor is measured in this study, so the five largest models — including both extrapolation points, which overshoot consistently — cannot be placed on this scale here.

Two observations do not depend on the floor's magnitude: the within-lineage checkpoint estimate and the cross-family association have opposite signs, and at 7B one family's level moves opposite

Table 4: Within-family residuals against the pre-training seed floor. Residual is measured minus predicted, relative to predicted, using the archived historical raw two-variable fit $(\alpha, \beta, \gamma) = (6.9190, -0.96283, -0.67403)$; these are neither the frozen primary model nor the corrected include-final fit. The two extrapolation points (starred) are not in that fit. This study includes independent seeds only up to 410M, so for the five larger models no noise level is available here; the shown ratios are descriptive magnitude comparisons, not calibrated tests.

Model	d/L	residual	vs. seed floor
Pythia-70M	85.3	-18.3%	$7.05\times$
Pythia-160M	64.0	$+35.9\%$	$6.54\times$
Pythia-410M	42.7	-6.6%	$0.21\times$
Pythia-1B	128.0	$+8.6\%$	—
Pythia-1.4B	85.3	$+0.1\%$	—
Pythia-2.8B	80.0	-11.3%	—
Pythia-6.9B*	128.0	$+19.7\%$	—
Pythia-12B*	142.2	$+22.4\%$	—

to prediction while the other's moves with it (Sec. F). We report both as directional observations; as noted there, neither localizes a cause, and we draw no recipe conclusion from them.

Supplementary Note 8: Discriminative signatures align within family and task. We tested whether a shared discriminative signature exists across models using a common representation adopted in this study: per-sample layer profiles interpolated to a 32-point fractional grid, with a cell's discriminative direction defined as its per-point Cohen's d (incorrect against correct).

Across all models, no. Mean pairwise cosine between cells' discriminative directions is 0.21 (depth profiles), 0.06 (time-layer) and 0.17 (time-course); same-task means do not exceed 0.29. Pooled-PCA subspaces show that profiles themselves are strongly low-dimensional (leading component: 67% of variance for time-layer profiles), but held-out projections onto shared components carry only weak signal (AUROC 0.44–0.63). Thus the observed shared low-dimensional structure does not yield consistent held-out discrimination in these measurements.

This connects, on the geometric side, to what has been reported for detector transfer: internal-state hallucination signals transfer across languages and domains for most models, but conditionally, with cross-lingual performance depending on class separability and feature-space alignment (Alansari et al., 2026). Theirs concerns the transfer of a trained detector and ours the transfer of the underlying discriminative direction; both indicate that transfer is condition-dependent rather than automatic in the settings examined, and we do not read their result as universal non-transfer.

Within a family, conditionally yes. Within Qwen2.5 (0.5B$\to 1.5$B$\to 3$B) on TriviaQA, temporal signatures align at pairwise cosine $0.86/0.76/0.81$ (time-layer) and $0.74/0.78/0.86$ (time-course), stable across a $6\times$ parameter range, while depth-profile alignment is moderate ($0.36/0.06/0.54$). On GSM8K nothing aligns, within family or across ($|\cos| \leq 0.36$, often negative). A leave-one-model-out ridge regression from architecture meta-features $(\ln d, \ln L, d_{\text{ffn}}/d, \ln D)$ to profile-PCA coefficients beats both pre-registered nulls only for time-layer profiles on TriviaQA (cosine 0.740 against $0.565/0.203$) — exactly where the family already aligns — and adds no held-out detection AUROC (0.595 against 0.597). We therefore claim the conditional alignment itself, not a predictive architecture formula.

Supplementary Note 9: Design-matrix spectrum and the identified direction. The conditioning argument of Supplementary Note 5 is stated informally there in terms of the spans of $\ln N$ and $\ln \eta$. Its exact form is a spectral statement about the design and is given here, because it fixes the relative information the grid carries on each linear combination of β and γ — which combinations it constrains more tightly and which less — independently of any measured response. It does not by itself fix their absolute precision, which also depends on the noise level and the error assumptions.

Let $X \in \mathbb{R}^{n\times 2}$ have rows $(\ln d_i, \ln L_i)$ and let $\tilde{X} = X - \mathbf{1}\bar{x}^\top$ be its column-centred form. Write the eigendecomposition of the centred cross-product as

$$S \;=\; \tilde{X}^\top \tilde{X} \;=\; V \Lambda V^\top, \qquad \Lambda = \operatorname{diag}(\lambda_1, \lambda_2), \quad \lambda_1 \geq \lambda_2 > 0, \tag{10}$$

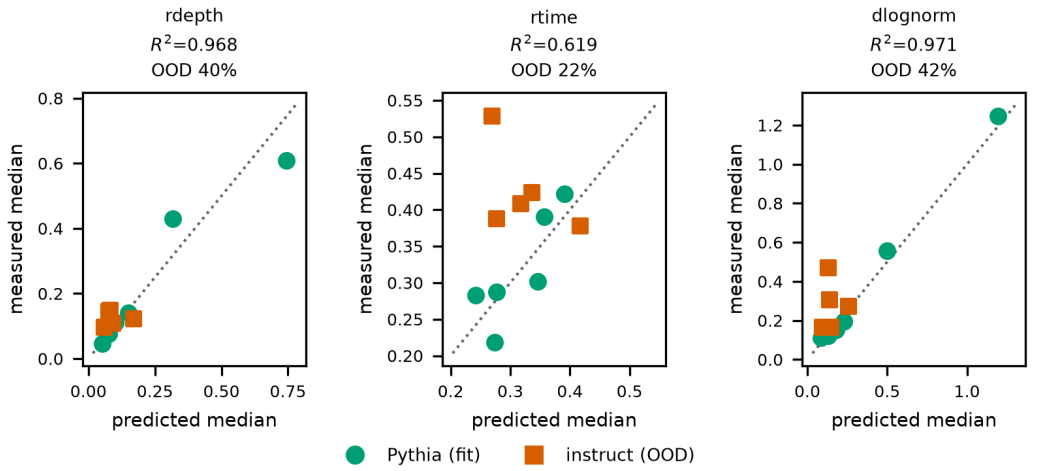

Figure 4: **Historical raw, include-final, two-variable fit and its boundary.** Predicted against measured median axis levels: Pythia fit set (circles) and out-of-family instruction-tuned models (squares; Q = Qwen2.5, TL = TinyLlama, S2 = SmolLM2). This figure is not the frozen main capacity fit. For the left depth-axis panel, four of five out-of-family levels are underestimated and one is overestimated; the time axis (centre) fits poorly even in-family. Points and fits are retained as the historical analysis.

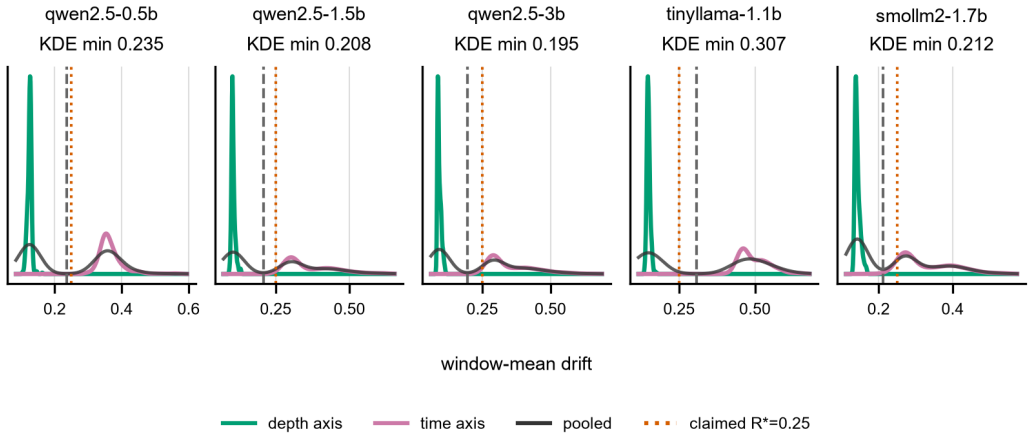

Figure 5: Pooling the two axes manufactures an apparent universal constant. Kernel density of the pooled statistic for each model, with the depth and time axes overlaid. In these samples the two axes occupy non-overlapping empirical ranges; on this data and the specified KDE settings, pooling them produces a separated density structure, and the reported location (dashed) is the density minimum between the two axis medians. It falls at 0.235, 0.208, 0.195, 0.212 and 0.307 across these five models — near 0.25 for the 0.5B-class models and elsewhere for the others — and it moves with the axis levels. No dynamical critical point need be assumed to produce this density structure; a measurement-axis mixture suffices.

with $V = [u_1\ u_2]$ orthonormal. Under $\mathrm{Cov}(\varepsilon \mid X) = \sigma^2 I$ (errors homoscedastic *and* uncorrelated — homoscedasticity alone would not suffice) the ordinary least squares covariance is $\sigma^2 S^{-1}$, so the variance of the coefficient along direction w with $\|w\| = 1$ is $\sigma^2 w^\top S^{-1} w$, which equals σ^2/λ_j for $w = u_j$. The ratio λ_1/λ_2 is therefore exactly the factor by which the design inflates the uncertainty of the worst-determined coefficient direction relative to the best-determined one, and it is a property of the chosen model sizes alone.

For the six-model fit grid,

$$S = \begin{bmatrix} 2.0725 & 1.6269 \\ 1.6269 & 1.8448 \end{bmatrix}, \quad \lambda_1 = 3.5895, \ \lambda_2 = 0.3278, \quad \frac{\lambda_1}{\lambda_2} = 10.95, \tag{11}$$

so $\kappa(\tilde{X}) = \sqrt{\lambda_1/\lambda_2} = 3.31$ and the variance inflation factor of the two predictors is 3.25 at $\mathrm{corr}(\ln d, \ln L) = +0.832$. The eigenvectors are $u_1 = (+0.7314, +0.6820)$ and $u_2 = (+0.6820, -0.7314)$.

The two architectural coefficients are combinations of the $(\ln d, \ln L)$ slopes $\beta = (b_d, b_L)$: from $\ln N = 2\ln d + \ln L + \mathrm{const}$ for $N = 12d^2L$ and $\ln \eta = \ln d - \ln L$, the capacity coefficient is $b_N = a_N^\top \beta$ with $a_N = (1,1)/3$ and the shape coefficient is $b_\eta = a_\eta^\top \beta$ with $a_\eta = (1,-2)/3$. What equation 10 controls is the variance of these *coefficient combinations*, through the alignment of their directions with the eigenvectors:

$$\left|\langle \hat{a}_N, u_1 \rangle\right| = 0.9994 \ \ (2.0^\circ), \qquad \left|\langle \hat{a}_\eta, u_2 \rangle\right| = 0.9592 \ \ (16.4^\circ), \tag{12}$$

with $\hat{a}$ the unit direction. The capacity combination lies almost along the best-determined direction u_1 and the shape combination almost along the worst-determined u_2, so under these released sizes b_η carries about $25\times$ the variance — roughly $5\times$ the standard error — of b_N (the design's condition number $\lambda_1/\lambda_2 = 10.95$ sets the worst-to-best inflation between unit directions; the two combinations differ in norm as well). Both coefficients are identifiable: the design has full column rank, so neither b_N nor b_η is unestimable. This is a statement about estimation *precision*, not about identifiability — b_N is simply estimated far more precisely than b_η here — and it matches the main text's D1, where full rank makes both coefficients identifiable and the condition number is a precision statement rather than a certificate that any one combination is well estimated. Adding the two held-out models raises the condition number λ_1/λ_2 to 15.95 — worse conditioning, not better — yet under the common-noise assumption lowers the design-variance factor $\mathrm{Var}(\hat{b})/\sigma^2$ of *both* coefficients (from 0.063 to 0.045 for b_N and from 1.57 to 1.28 for b_η); this is a fall in the design factor from adding data, not a claim that the realised standard errors drop or that the design becomes better conditioned.

Two consequences are worth stating because they bound the generality of the result rather than supporting it. First, equation 12 is an accident of which models were released, not a fact about transformers: the coefficient directions a_N, a_η are fixed, while the design's principal axes u_1, u_2 — and hence the alignment — depend on the released sizes. A suite with more ample, independent variation in capacity and shape could improve the information available on the shape coefficient; one would then re-check full rank, the condition number and the variances for that design rather than assume the present ordering persists. Holding N_{proxy} strictly fixed instead would make $2\ln d + \ln L$ constant, so the intercept-plus-two-slope design would have rank at most two and b_N would confound with the intercept — a rank failure, not a re-ordering of precision.

The same calibration, response-free because equation 10 involves no measured response, compares design *conditioning* across the released suites — not absolute coefficient precision, and not whether any released suite could improve the shape estimate. The centred-Gram variance ratios are 16.1 (Qwen2.5), 16.4 for OPT (Zhang et al., 2022) including its 350M size and 61.6 without, 62.6 (Dey et al., 2023), 317 for GPT-2 (Radford et al., 2019) and, for OLMo-2 (Team OLMo et al., 2024), infinity — every OLMo-2 size has η exactly 128, so that design has rank one and b_η is not poorly determined but absent. The Pythia grid at 10.95 is the best-conditioned of them. OPT is instructive twice over: the only size supplying aspect-ratio variation, 350M, is also the only one using post-LayerNorm rather than pre-LayerNorm, so including it buys conditioning at the cost of the recipe homogeneity a residual-stream measurement depends on.

The near-collinearity is therefore a property of model-release practice rather than of our choice of grid, and the suites built expressly for scaling work are the most collinear: compute-optimal sizing prescribes a relation between width and depth. Across the audited designs the spectrum reveals a directional information imbalance — the shape combination sits near the worst-determined axis — but this describes *relative* information, not absolute coefficient precision (which also depends on the noise level and the error assumptions), and it does not establish that the shape coefficient cannot be estimated more precisely from released models, nor that retraining is the only route: more data along the same axes lowers the variance at fixed conditioning — scaling S to $2S$ leaves λ_1/λ_2 unchanged yet halves both coefficient variances (the shape factor falls $1.57 \to 0.79$). OLMo-2, with

η fixed exactly, is the one genuinely rank-one case, where b_η is absent rather than merely imprecise. The practical consequence extends beyond this paper — a width or depth exponent fitted on any of these suites inherits the same imbalance, whether or not a conditioning diagnostic accompanies it. Second, because equation 10 involves no response variable, the calibration can and should be run on any candidate family *before* its levels are measured; we do so in Supplementary Note 5 and report the per-family result rather than importing the Pythia axes. The factor $1/3$ appearing in $b_N = (\beta+\gamma)/3$ and $b_\eta = (\beta - 2\gamma)/3$ is the magnitude of the determinant of the *inverse* of the change of basis $[[2, 1], [1, -1]]$ (whose own determinant is -3) and carries no family-specific information.

Supplementary Note 10: What the spectrum licenses for prediction. A natural next step is to fit in the eigen-coordinates themselves, on the expectation that the intrinsic axes would predict better. They cannot, and the reason is worth recording because the expectation is a common one. Ordinary least squares is equivariant under any invertible reparameterisation of the design: replacing X by XA for invertible A leaves the column space, the projection onto it, and therefore every fitted value and every out-of-sample prediction unchanged. The eigenvector matrix V in equation 10 is orthogonal, so a fit in (z_1, z_2) and a fit in $(\ln d, \ln L)$ are one model in two alphabets. We verified this rather than assuming it: the two agree at every point to 2×10^{-15}, and $V(b_{z_1}, b_{z_2})^\top$ recovers (β, γ) to six decimals. The corollary is that no held-out model can ever adjudicate between $(\ln d, \ln L)$ and $(\ln N, \ln \eta)$, and any claim that a change of coordinates improved a prediction is a claim about something else.

What the spectrum does license is discarding or shrinking the weak direction, which changes the model. Two forms were tried. Hard truncation keeps only u_1. Ridge on the eigen-components has, since $z^\top z = \Lambda$ is diagonal, the exact per-component solution $b_j^{\text{ridge}} = [\lambda_j/(\lambda_j + \alpha)]\, b_j^{\text{ols}}$, so a single penalty shrinks u_2 more than u_1 automatically. The components must not be standardised, which would make the shrinkage isotropic and remove the mechanism, and ridge does not nest truncation, since it always shrinks u_1 as well.

This historical comparison uses the corrected, include-final measurement and the two-variable fit; it is not the frozen main window. The penalty was selected by leave-one-out on the six-model fit grid alone, with the eigenvectors recomputed inside each fold so that the held-out model's design row does not influence the direction under test; 6.9B and 12B were never consulted in selection. Against 23.1% leave-one-out for the two-variable fit, one-component PCR gives 13.3% and directed ridge at the selected $\alpha = 0.16$ gives 14.5%. Isotropic shrinkage of the whole coefficient vector, tuned identically, gives 23.1%. PCR is lowest in this comparison and directed ridge remains below isotropic shrinkage; these are descriptive model comparisons and do not identify why their errors differ.

We nonetheless decline to promote either form. On the two held-out points, mean error is monotone in α, running -19.3% at $\alpha = 0$ to $+450\%$ at the top of the sweep and crossing zero at $\alpha = 0.316$. The leave-one-out optimum sits a factor of 2.0 from that crossing. Two extrapolation points do not identify whether curvature, a level offset or another mismatch produces this path. Summarising the six fold errors by their median puts the optimum at $\alpha = 0.16$ and by their mean at 0.45, a factor of 2.8 apart. This is sensitivity to the evaluation summary in a six-model fit, not evidence that the algorithm failed to converge.

Supplementary Note 11: Closed-form ridge penalties and their equivariance. One proposed extension is to fit $\alpha(N) = \alpha_0 N^\gamma$ and give each scale its own. That relation is not estimable here. A penalty is one number per *fit set*, not per model: estimating a two-parameter scaling relation requires several independent (N, α^*) pairs, each from its own fit set, and this archive contains one fit set per family. The leave-one-out folds supply six more α^* values, but they share four of five models with one another and a fit set has no scale to regress against — it is a collection of models, not a point on an axis. The Hoerl–Kennard–Baldwin rule $\alpha_{\text{HKB}} = p\,\hat\sigma^2/\|\hat b\|^2$ is a fit-derived estimator under its assumptions, not a universal MSE optimum. It therefore does not prove that a penalty cannot vary with study scale or noise structure.

An alternative removes the search under explicit assumptions. In eigen-coordinates $\hat b_j \sim \mathcal{N}(b_j, \sigma^2/\lambda_j)$, and under a prior $b_j \sim \mathcal{N}(0, \tau^2)$ the posterior mean is exactly the ridge estimator

with

$$\alpha = \sigma^2/\tau^2, \qquad \hat{b}_j^{\mathrm{ridge}} = \frac{\lambda_j}{\lambda_j + \alpha}\,\hat{b}_j, \tag{13}$$

so the Bayes penalty is a noise-to-prior-variance ratio rather than a value set by generalized cross-validation (Golub et al., 1979) or a similar held-out search. This relation is distinct from HKB. Two fit-derived rules follow, neither of which consumes validation data: Hoerl–Kennard–Baldwin (Hoerl & Kennard, 1970; Hoerl et al., 1975), $\alpha = p\hat{\sigma}^2/\hat{b}^\top\hat{b}$, and empirical Bayes, with τ^2 from the marginal moments $\mathbb{E}[\hat{b}_j^2] = \tau^2 + \sigma^2/\lambda_j$. Because $\hat{\sigma}^2$, $\hat{b}$ and λ_j are properties of the fit in hand, their values are fit-specific. The present archive does not establish that either rule transports across families.

We carry two variance inputs through as a sensitivity analysis. The first is the residual variance from the six-point, three-parameter architecture fit (three residual degrees of freedom), $\hat{\sigma}^2 = 0.051$; the second is the across-seed variance from nine 410M runs at fixed architecture and training budget (Methods, 'Pre-training seed floor'), $\hat{\sigma}^2 = 0.127$. These have different sampling units and are not interchangeable estimates of one population quantity. Their numerical ordering does not establish a usual size relation, a measurement-noise upper bound or absence of residual signal. The two documented seed-level performance differences likewise do not by themselves invalidate the nine-run dispersion.

A third estimator, Lawless–Wang (Lawless & Wang, 1976), $\alpha = p\hat{\sigma}^2/\sum_j \lambda_j\hat{b}_j^2$, was implemented and withdrawn, for a reason that applies to any use of these formulae outside their original setting. Its denominator is $\|X\hat{b}\|^2$ and therefore carries the units of the response, the same as $\hat{\sigma}^2$, so its α is dimensionless — while λ_j, the quantity α is added to in equation 13, has the units of the squared log predictor. Both estimators are defined in the literature for standardised predictors, where λ is dimensionless and the mismatch cannot arise; ours are raw centred logarithms. The consequence is not academic: changing from natural to base-2 logarithms, which is not a modelling decision, moves its shrinkage factors from $(0.995, 0.946)$ to $(0.998, 0.974)$ while leaving Hoerl–Kennard–Baldwin and empirical Bayes unchanged to four decimals. We verify equivariance numerically for every estimator we report, by rescaling the predictors and confirming that α scales as c^2 and the shrinkage factors do not move.

On the corrected measurement the four reported HKB and empirical-Bayes combinations (residual- or seed-based noise) give α between 0.060 and 0.196, and each reduces the two held-out errors relative to unpenalised least squares (-22.1% and -16.5%). The largest candidate is empirical Bayes with seed noise, $\alpha = 0.196$, giving -10.7% and -1.8%; HKB with seed noise gives $\alpha = 0.148$ and errors of -13.2% and -5.1%. These candidates use no held-out data, but two extrapolation points neither rank the estimators nor identify why shrinkage helps in this comparison.

Qwen2.5 contributes three models to a three-parameter fit. A full-rank saturated fit can have RSS zero, but the usual residual-variance estimator $\mathrm{RSS}/(n-p)$ is undefined because it has no residual degrees of freedom; it is not an estimate of zero variance. External noise information could support some rules under additional assumptions, but this protocol does not evaluate or validate that substitution. We therefore exclude the saturated fit from this analysis as a protocol decision, not as a claim of mathematical impossibility.

Supplementary Note 12: Detection results per confirmatory cell. Under a protocol frozen before the confirmatory runs, we evaluated 11 geometric features against 7 single-pass probability baselines on five instruction-tuned models and two labeled task families (Table 5).

In 7 of 9 evaluable cells the best geometric feature attains $|\mathrm{AUROC} - 0.5| \geq 0.15$ with Holm-corrected permutation p between 0.0022 and 0.0066 (six of the seven at the Holm-corrected minimum $0.00220 = 11/(B{+}1)$ of a Monte-Carlo $B{=}5000$ permutation test — a computed value at the procedure's resolution, not an upper bound), and in *every* cell that feature is a depth-axis quantity — norm growth $\Delta\ln\|h\|$ on GSM8K, rotation R_{depth} or $\Delta\ln\|h\|$ on TriviaQA. Temporal features never win a cell.

Adding all 11 geometry features to a cross-validated logistic model over the 7 baselines changes AUROC by -0.005 on average: the pre-registered YELLOW outcome. The exception is a coherent pocket — Qwen×TriviaQA at 0.5B ($0.792 \to 0.850$, bootstrap CI of the difference $[+0.029, +0.157]$) and 1.5B ($0.827 \to 0.849$) — localized on the depth axis (base+depth 0.840

Table 5: **Detection results per confirmatory cell.** Task accuracy, best geometry feature with AUROC (incorrect=positive), 95% bootstrap CI and Holm-corrected permutation p; single-feature baselines; 5-fold cross-validated logistic AUROC for baselines-only against baselines+geometry. Best-geometry AUROCs below 0.5 indicate inverted orientation. TinyLlama–GSM8K (accuracy 0.02) is excluded by the pre-registered minimum-class rule. Holm correction is over the 11 geometry features within each cell (per cell, not over all tests in the paper). A Monte-Carlo permutation test with B=5000 and the (exceedances+1)/(B+1) estimator (Phipson & Smyth, 2010) cannot resolve a raw p below $1/(B{+}1) = 2.0 \times 10^{-4}$, so the smallest corrected value it can express is $11/(B{+}1) = 0.00220$; the six cells at that value carry that computed p (the procedure's minimum), not an upper bound.

Model	Task	Acc	Best geo. feature	AUROC [CI]	p_{Holm}	H_{tok}	len	$\mathrm{CV}_{\mathrm{base}}$	$\mathrm{CV}_{+\mathrm{geo}}$
Qwen2.5-0.5B	GSM8K	.41	$\Delta\ln\|h\|$ max	.727 [.65,.80]	.00220	.714	.759	.803	.778
Qwen2.5-0.5B	TriviaQA	.25	$\Delta\ln\|h\|$ mean	.793 [.72,.86]	.00220	.749	.609	.792	**.850**
Qwen2.5-1.5B	GSM8K	.59	$\Delta\ln\|h\|$ max	.651 [.57,.73]	.00220	.697	.787	.812	.802
Qwen2.5-1.5B	TriviaQA	.35	R_{depth} mean	.662 [.58,.74]	.00220	.820	.469	.827	**.849**
Qwen2.5-3B	GSM8K	.77	$\Delta\ln\|h\|$ mean	.720 [.64,.80]	.00220	.571	.859	.836	.824
Qwen2.5-3B	TriviaQA	.40	R_{depth} mean	.663 [.59,.74]	.0044	.840	.523	.854	.835
SmolLM2-1.7B	GSM8K	.48	$\Delta\ln\|h\|$ max	.641 [.56,.72]	.0066	.662	.713	.767	.735
SmolLM2-1.7B	TriviaQA	.35	R_{depth} mean	.682 [.60,.76]	.00220	.792	.563	.801	.779
TinyLlama-1.1B	TriviaQA	.31	R_{depth} max	.636 [.55,.72]	.0198	.690	.632	.824	.820

against base+time 0.796 at 0.5B) and gone by 3B ($0.854 \to 0.835$). Within that pocket geometry alone matches the full baseline set (geometry-only CV AUROC 0.851).

We pre-registered the hypothesis that $\Delta\ln\|h\|$ would act as a simple, stable, monotonic reparameterization of the tested output-confidence proxies. The $\Delta\ln\|h\|$–entropy Spearman correlation ranges from -0.34 to $+0.51$ with unstable sign (partial correlation given log-probability: -0.24 to $+0.35$), while $\Delta\ln\|h\|$ is the best single geometry feature in 5 of 9 cells. These results do not support that tested proxy explanation; they do not establish statistical independence or exclude nonlinear confidence relationships.

Supplementary Note 13: Block topology and branch alignment. One architectural distinction separates the fit family from every validation family. Pythia computes its attention and feed-forward branches in parallel from a common normalised input; Qwen2.5, Qwen3, TinyLlama and SmolLM2 are Llama-style and compute them sequentially, the feed-forward branch reading the attention branch's output. Per-block rotation is a statement about the sum of a block's writes, so the angle between those two writes is the quantity through which a topology difference would act, and the composition identity of the main text exposes it directly: with $\cos$ at zero the two branches contribute in quadrature, and with $\cos$ negative they partially cancel, so the same pair of branch magnitudes yields a smaller total write.

The following summaries pool finite $\cos(\Delta_{\mathrm{attn}}, \Delta_{\mathrm{mlp}})$ values over tokens and interior layers within each model, using 8 trajectories for Pythia-70M, 12 for Pythia-1.4B and 200 for each sequential model. Longer answers therefore contribute more pooled cells; the cells are not independent samples, and the IQR is a descriptive distribution rather than an uncertainty interval.

The two parallel models bracket zero and the five sequential ones are all negative and mutually close, which is the pattern the topology argument predicts. We do not report it as a test of that argument. Every parallel model here is Pythia and every sequential one is not, so topology is perfectly confounded with tokenizer, corpus, optimiser schedule and model generation, and a table in which the grouping variable coincides exactly with the family variable cannot separate them. The measurement is also asymmetric — two models against five, with the two differing from each other by more than the five differ among themselves.

Identifying a topology effect requires an evidence design that separates it from the tokenizer, corpus, optimiser and model-generation differences bundled with family in this table. A controlled topology ablation is one direct option; a second parallel-residual suite could test recurrence of the level pattern, but would not by itself remove every recipe confound. The available GPT-J and GPT-NeoX-20B

Table 6: Branch alignment by block topology, pooled over token×interior-layer cells within each model. The two parallel-residual models bracket zero and the five sequential ones are all negative, which is the pattern the topology argument predicts — but topology is perfectly confounded with family here, so this is reported as an observation and not as a test. Trajectory coverage is 8, 12 and 200 per row as specified in the text; pooled cells are not independent observations.

model	topology	median cos	IQR	fraction < 0
Pythia-70M	parallel	$+0.020$	$[-0.061, +0.094]$	0.43
Pythia-1.4B	parallel	-0.023	$[-0.157, +0.087]$	0.56
Qwen2.5-0.5B	sequential	-0.114	$[-0.205, -0.007]$	0.76
Qwen2.5-1.5B	sequential	-0.133	$[-0.214, -0.048]$	0.85
Qwen2.5-3B	sequential	-0.117	$[-0.198, -0.039]$	0.85
SmolLM2-1.7B	sequential	-0.137	$[-0.284, -0.005]$	0.76
TinyLlama-1.1B	sequential	-0.102	$[-0.192, -0.017]$	0.80

checkpoints do not provide enough sizes to satisfy C1, so such a comparison would not identify an exponent.

G AI USE STATEMENT

This section is required by the ICLR 2027 AI policy and does not count toward the page limit.

A large language model assistant was used throughout this project, across all versions of the work from its first draft to the present manuscript. Its role was substantial enough that we describe it precisely rather than summarily.

Where it was used. Code development and review for the extraction probe and the analysis scripts; literature search and bibliographic retrieval; drafting and editing of the text, including compressing an earlier long-form manuscript into this one; and construction of the quality-assurance scripts that recompute reported quantities from the archived data.

Where it was not. No experimental design decision, hypothesis, pre-registered decision rule, or interpretation of a result originated from the assistant. The pre-registered protocol and its GREEN/YELLOW/RED thresholds were fixed by the author before the confirmatory runs. No reported number was produced or estimated by a language model: every quantity comes from the released code applied to the archived trajectories. A one-command checker automatically re-derives five headline values from the shipped fit-grid trajectories and, for the two extrapolation residuals among them, the shipped 6.9B/12B held-out trajectories. It verifies that they match the manuscript source and that superseded values do not reappear; further analyses recompute through their own commands, and three larger analyses are provided as archived reports, as the Reproducibility statement sets out.

Review and responsibility. All AI-assisted work was reviewed. Code was verified by re-deriving reported quantities from the archived data with checker and analysis scripts; retrieved references were checked against the primary sources, which is how an incorrect publication year and four unsupported claims from an earlier version were caught and withdrawn. No language model is an author. The author takes full responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI, and for the corrections and withdrawals it reports.