ACTIVATION DENSITY IN TRANSFORMER FEED-FORWARD NETWORKS IS SET BY THE NONLINEARITY, NOT THE WIDTH

Anonymous authors
Paper under double-blind review

ABSTRACT

Two thirds of a Transformer's parameters sit in its feed-forward blocks, whose width follows from an expansion ratio r fixed at 4 since 2017. Work on activation sparsity has tied that width to the fraction of the block a model uses. We find the two nearly independent. We measure the fraction of FFN units carrying the output, the activation density, with a participation ratio, which needs no threshold and so works for gated activations. Across nine Transformers trained to differ only in r, density moves 7% over a fourfold range, reconciling an earlier report of sparser wide MLPs that confounded width with model scale. Density is set by the nonlinearity. Two moments of the pre-activation, with the activation function, predict it to 0.49% across non-gated nonlinearities whose densities differ by a factor of 3.6; gated activations carry a residual we quantify and leave open. The prediction's closed forms, such as $\alpha = 1/\pi$ for ReLU, hold only at zero pre-activation mean, which training destroys. Constraining the first-layer rows to average to zero restores that mean: ReLU density returns to $1/\pi$ within 0.65%, and the width slope vanishes, as pre-registered, for both activations whose gate is positively homogeneous. Applied to a covariance spectrum, the same participation ratio understates effective rank by a median factor of four, and by up to twenty-six, because trained networks learn a handful of massive activations. Utilization stays put while width changes, so it cannot be what sets the width; any case for a ratio must be made on capacity.

1 INTRODUCTION

The position-wise feed-forward network holds roughly two thirds of a Transformer's parameters and a comparable share of its compute. Its width is set by the expansion ratio $r = d_{\text{ff}}/d_{\text{model}}$, which the original Transformer fixed at 4 (Vaswani et al., 2017). Since 2017 d_{model} has grown from 512 to past $12{,}000$ and nearly every other architectural choice has been revisited at least once, yet the ratio has scarcely moved.

Much of the work on the FFN assumes that its width is tied to how much of it a model uses. The activation-sparsity literature reports that few units respond to any given token (Li et al., 2022; Szatkowski et al., 2025), and efficiency methods trade on that fraction. If utilization and width were coupled, measuring what the FFN uses would tell us how wide it should be.

The two are decoupled. The fraction of FFN units that carry the output depends on the *activation function* and is almost blind to the expansion ratio. The insensitivity has a structural cause. Activation density is a symmetric function of the activation magnitudes, so it sees only *how many units take each value*; which unit takes which value, and how the units covary, never enter. Width can reach density only by moving that distribution, so no argument from utilization can settle how wide the FFN should be.

We measure utilization by the activation density α. Writing the FFN output as a superposition of per-unit contributions,

$$\text{FFN}(x) \;=\; \sum_{i=1}^{d_{\text{ff}}} h_i \, w_2^{(i)}, \qquad h_i = \phi\Big(w_1^{(i)\top} x\Big), \tag{1}$$

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.

α is the participation ratio of $|h|$ divided by d_{ff}, the effective fraction of the d_{ff} terms that carry the sum. No threshold enters, so α stays meaningful for GELU and for gated activations, which produce no exact zeros.

The evidence. Our central experiment fixes data, depth, width, seed, optimizer and schedule, and varies only $r \in \{2, 4, 8\}$, with three seeds each. Over that fourfold change α moves 7% (§5.1). The earlier report that wider MLPs are sparser (Li et al., 2022) compared models that differed at once in size, data and recipe; with those held fixed, most of the dependence disappears. Density instead tracks the nonlinearity. If the pre-activations a token induces are roughly Gaussian and roughly independent, the law of large numbers gives $\alpha = (\mathbb{E}|h|)^2/\mathbb{E}[h^2]$ for $h = \phi(z)$, so two moments of z fix a nonlinear functional of $\phi(z)$. The Gaussian model holds to 0.49% across ReLU and GELU, whose densities differ by $3.6\times$. For gated activations it breaks, and an exact decomposition traces the gap to coupling between the gate and up streams and to their non-Gaussian shapes (§5.6).

Effective counts are also harder to measure than they look. On the covariance spectrum of a trained Transformer the participation ratio understates effective rank by a median factor of 4, and by up to $26\times$, because a few units carry activations orders of magnitude above the rest (Sun et al., 2024). §3.1 settles the instrument before anything depends on it.

Why the ratio is worth explaining. The value 4 was never a measured optimum. Vaswani et al. (2017) did vary it, and the largest ratio they tried was best on both of their metrics but was not the one they shipped (Appendix N). Our runs, an order of magnitude smaller, show the same direction. Extra width keeps helping, and utilization cannot be the reason. For efficiency work, activation sparsity cannot be bought by changing FFN width. It is a property of the activation function, and its size follows from two measured moments of the pre-activation.

Contributions.

- Activation density barely responds to the expansion ratio, moving 7% over a $4\times$ range across nine models and three seeds under full control. This reconciles a prior report confounded by model scale (§5.1). We explain the invariance and measure the one condition it needs, which holds with room to spare in eleven pretrained models and strengthens with width (§5.2).
- Density follows from the activation function and two measured moments of the pre-activation, to 0.49% for non-gated activations whose densities differ by $3.6\times$ (§5.3). Gated activations leave a residual, which we decompose exactly into gate shape, up-stream shape and gate–up magnitude coupling (§5.6).
- We test this account causally. Constraining the first-layer rows to average to the zero vector holds the pre-activation mean at zero for every token. Across twenty-seven models, the width slope vanishes, as pre-registered, for the two activations whose gate is positively homogeneous, and stays positive for SwiGLU, whose SiLU gate leaves the scale channel open. ReLU density returns to $1/\pi$ within 0.65%, and a ReLU-gated arm reaches $2/\pi^2$ within 5.1% (§5.5).
- The participation ratio, the standard effective-count estimator, is unsafe on trained Transformers because they learn massive activations. We quantify the error and supply a working alternative (§3.1).

Together these results close off utilization-based accounts of the expansion ratio (§6.1).

We mark throughout which analyses were specified in advance. Three predictions were registered before the runs that tested them. One succeeded and one failed. The third, written down after the failure was diagnosed and before the intervention it suggested, held for the two homogeneous arms and held only qualitatively for SwiGLU.

2 Related Work

Activation sparsity. Li et al. (2022) showed that trained Transformers fire few MLP units per input, and reported the dependence on width that we revisit in §5.1. Szatkowski et al. (2025) measure sparsity across contemporary gated models with a top-p mask on the ℓ_1 norm and conclude that

model size alone does not determine it. Neither study isolates the expansion ratio as a controlled variable, and neither gives a closed form for the fraction it measures.

Gaussian pre-activations. The analytical tools we use were developed for networks at initialization. Price et al. (2024) derive sparsity in closed form for sparsity-inducing activations in a Gaussian-process limit at random initialization, and Wolinski & Arbel (2025) map out when the Gaussian assumption holds, finding it fragile at finite width. We apply the same tools to trained networks. The single term separating the two regimes is the pre-activation mean, which symmetry sets to zero at initialization and training does not preserve (§5.4).

What the FFN represents. Geva et al. (2021) read FFN layers as key-value memories. Elhage et al. (2022) show that networks pack more features than they have dimensions by placing them in superposition, at a cost in interference; Appendix L specializes that argument to the FFN. These works ask which features get represented. We ask what fraction of units carry them, and whether that fraction depends on width.

3 METHODOLOGY

The paper rests on two quantities, the activation density of a single token across FFN units and the effective rank of the FFN output across tokens. The second is harder to estimate than it appears, so we settle how to measure both before drawing conclusions from either, and then show why width has so little purchase on the first.

3.1 THE STANDARD ESTIMATOR FAILS ON TRAINED TRANSFORMERS

Equation 1 writes the FFN output as a sum of d_{ff} contributions, each a fixed vector $w_2^{(i)}$ scaled by a data-dependent activation h_i. This raises two questions that must be kept apart. The first is how many *units* are appreciably active for a given token. We measure it by the participation ratio of $|h|$, normalized by width,

$$\alpha \;=\; \frac{1}{d_{\text{ff}}} \cdot \frac{\left(\sum_i |h_i|\right)^2}{\sum_i h_i^2}, \tag{2}$$

which equals $1/d_{\text{ff}}$ when one unit carries everything and 1 when all units contribute equally. This is the activation density, the paper's central object. Needing no threshold, it remains meaningful where counting exact zeros does not, and as an effective fraction it should not be read as a thresholded count. For ReLU at the measured $\mu/\sigma = -0.555$, 29% of units are strictly positive while $\alpha = 0.178$, since the participation ratio weights each unit by its magnitude.

The second question is how many *directions* the outputs span across tokens. The answer lies in the spectrum of the covariance $C = \mathbb{E}[hh^\top] - \mathbb{E}[h]\,\mathbb{E}[h]^\top$, and here the natural estimator breaks down. We treat it now because this spectral quantity feeds the analyses of Appendices H and L.

The participation ratio is unsafe on spectra. Applied to eigenvalues, the participation ratio $(\sum_k \lambda_k)^2 / \sum_k \lambda_k^2$ weights directions by variance, and the squared denominator lets a single large eigenvalue dominate. Trained Transformers contain exactly such directions. Table 1 in Appendix A compares it with the entropy-based effective rank $\mathrm{erank} = \exp(-\sum_k p_k \log p_k)$, $p_k = \lambda_k / \sum \lambda$ (Roy & Vetterli, 2007), on the same centred covariance for Pythia-70m (Biderman et al., 2023), and the two disagree by a factor of 2.8 to 12.6 across its six layers. Over all 228 layers of the eleven models the median factor is 4.1 (5.2 for Pythia, 3.6 for the gated models) and the largest 26. At layer 2, whose largest activation is 138.5 against 4.8–10.2 in neighbouring layers, the participation ratio reads 6.3 and erank 79.3. Figure 3 isolates the mechanism on a synthetic spectrum with a known true count.

In pre-norm architectures these outliers are learned. At random initialization the ten pre-norm models we measure peak at 2.6–17.6, and after training at 62–5133, a 9- to 512-fold increase (Table 3). OLMo-2 is the exception: it normalizes each block's output instead of its input, so its FFN sees the raw residual stream and already peaks at 470 at initialization. This is the massive-activation phenomenon of Sun et al. (2024), and it disqualifies the participation ratio as an effective-rank

estimator, so we report erank for every spectral quantity. Density is computed per token over units, never over a spectrum, and is unaffected.

Controls. A null control tests whether erank responds to training. Erank per d_{model} separates trained from randomly initialized models of the same architecture by 13.0%, 29.6% and 48.8% at Pythia-70m, 160m and 410m, a gap that widens with model size (Table 2). The token budget caps what can be measured, since a covariance built from N tokens has rank at most N. We use $16{,}384$ tokens, and the code rejects any layer whose estimate approaches that bound. Appendix A lists the analytic sanity checks each estimator passes and the independent reimplementation against which they were verified.

3.2 Why width is invisible to activation density

The near-invariance reported in §5.1 follows from the algebraic form of α.

Proposition 1 (Density is a functional of the empirical measure)**.** *Fix a token and let $a_1, \ldots, a_D$ be the pre-activations it induces, $h_i = \phi(a_i)$. Let $\nu_D = \frac{1}{D}\sum_i \delta_{a_i}$ be their empirical measure. Then, exactly and for every D,*

$$\alpha_D \;=\; \Psi[\nu_D], \qquad \Psi[\rho] \;\equiv\; \frac{\left(\int |\phi|\,\mathrm{d}\rho\right)^2}{\int \phi^2\,\mathrm{d}\rho}. \tag{3}$$

Proof sketch. α_D is a ratio of two sums over units, each of which is D times an average against ν_D; the factors of D cancel. Full proof in Appendix M.

Activation density is therefore permutation-symmetric. It depends on the distribution of magnitudes across units, whichever units carry them. Since D enters Eq. 3 only through ν_D, the expansion ratio can move α only by reshaping that distribution.

Corollary 1 (Width-invariance)**.** *Let ϕ be continuous. If $\nu_D \to \nu$ weakly for some D-independent limit ν with $\int \phi^2\,\mathrm{d}\nu > 0$, and ϕ^2 is uniformly integrable under $\{\nu_D\}$ (for instance, $\sup_D \int |\phi|^{2+\epsilon}\,\mathrm{d}\nu_D < \infty$ for some $\epsilon > 0$), then $\alpha_D \to \Psi[\nu]$, independent of D and hence of r at fixed d_{model}.*

The corollary does not assume independent pre-activations. This matters because once $D > d_{\text{model}}$ the rows $w_1^{(i)}$ cannot be linearly independent, and neither can the a_i. Linear dependence constrains the row *set*, whereas α sees only the row *measure*. The corollary does need ν_D to *concentrate*, which strong common-mode dependence prevents. In simulation at equicorrelation $\rho = 0.2$ with the marginal held fixed, α moves from 0.175 to 0.188 (Appendix M).

The measured invariance is therefore informative. It places trained FFNs in the weak-dependence regime where ν_D concentrates despite the linear dependence that $D > d_{\text{model}}$ imposes, and it predicts that a network driven into strong common-mode correlation would lose the invariance. Because concentration is an empirical property of trained networks, we measure it directly in §5.2.

4 Experiments

Every causal claim in the paper rests on controlled sweeps over models we train ourselves. Forward passes over public checkpoints extend the measurements to larger models, at the cost of confounds.

Controlled sweeps. Public models that differ in expansion ratio also differ in laboratory, data, tokenizer and training budget, so the published zoo cannot show whether width moves activation density. We therefore train our own. All sweeps share one recipe: TinyStories (Eldan & Li, 2023), $d_{\text{model}} = 256$, six layers, eight heads, 2500 steps on a fixed schedule, and 5–12M parameters per model. Each sweep varies exactly one factor. The seed governs weight initialization and batch sampling together, so a given seed sees byte-identical data at every setting. Cells have three seeds unless noted. Parameter count varies with r by design, and we record validation loss for context.

The four sweeps vary the expansion ratio under SwiGLU (Table 4), the activation function at $r = 4$ (Table 5), the expansion ratio under a non-gated ReLU FFN (Table 10), and the zero-mean-row

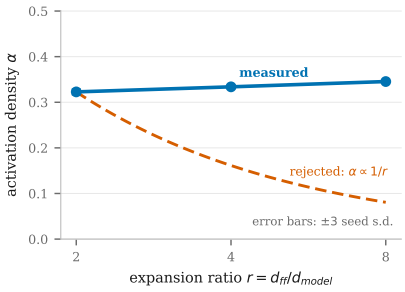

Figure 1: Activation density does not track the expansion ratio. Nine models identical in data, depth, width, seed and schedule, varying only r. The dashed line is $\alpha \propto 1/r$, the expectation if the network conserved a fixed number of active units per output dimension. Error bars (± 3 seed standard deviations) are smaller than the markers.

constraint across three activation functions and three ratios, twenty-seven models in all (Table 8). Appendix B gives the optimizer, schedule and rationale for each.

Pretrained checkpoints. Forward passes over eleven public checkpoints measure activation density, effective rank and the common-mode component. The set comprises the Pythia ladder (Biderman et al., 2023) from 70m to 1.4b and six gated models from the SmolLM2 (Ben Allal et al., 2025), TinyLlama (Zhang et al., 2024), OLMo-2 (OLMo Team, 2025) and Qwen2.5 (Qwen Team, 2024) families, spanning expansion ratios from 2.667 to 5.833. Each model is evaluated on 16,384 tokens of WikiText, a budget that keeps the rank cap of §3.1 well clear of the estimated quantities. Randomly initialized models of matching architecture serve as the null control.

Pre-registration. Three predictions were written down before the runs that tested them. The depth-versus-width test of Appendix H succeeded. The prediction $\alpha = 1/\pi$ for trained ReLU networks (§5.4) failed. The three constrained slopes of §5.5, registered once that failure had been diagnosed, were met exactly by the two homogeneous arms; SwiGLU's stayed positive as predicted, though above the forecast value.

5 Results

5.1 Activation density does not respond to the expansion ratio

The nine-model sweep of §4 holds data, depth, width, seed, optimizer and schedule fixed, and varies only r.

Figure 1 and Table 4 in Appendix C give the result. Across a fourfold change in r, activation density moves by 7%, from 0.3227 to 0.3338 to 0.3454 at $r = 2, 4, 8$, with seed standard deviations of 0.0006 to 0.0020. The fitted slope is $\mathrm{d}\log\alpha/\mathrm{d}\log r = +0.049$.

For scale, the strongest coupling one could posit is a fixed number of active units per output dimension. It implies $\alpha \propto 1/r$, a slope of -1 and a 400% change over this range. Instead, αr varies by $4.28\times$ while α varies by $1.07\times$. The data rule out a fixed budget of activity spread over the extra width.

Reconciling prior work. Li et al. (2022) report that Transformers with wider MLP hidden dimensions are sparser, measuring sparsity as the proportion of exact zeros after a ReLU. Their comparison changes MLP width together with model size, data and recipe. Ours changes width alone, and the dependence largely disappears. We read their observation as correct for their setting and confounded by scale. Their measure also requires exact zeros, which the gated activations in current models never produce.

A residual slope, and where it comes from. The invariance is approximate. The $+0.049$ slope is small but real, lifting α by 0.011 per doubling of r, or five to ten seed standard deviations. Section 3.2

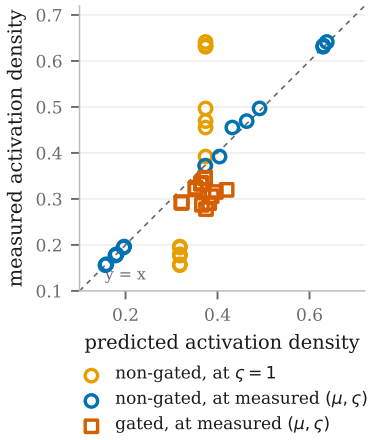

Figure 2: Predicted against measured activation density for eleven pretrained models and our controlled runs. Eq. 4 evaluated at the measured within-token moments (blue) puts non-gated activations on or near the identity line, the largest miss being 5% for Pythia-1b; evaluated at $\mu = 0$, $\sigma = 1$ (orange) it does not. Gated activations (red squares) retain a residual, which Table 12 decomposes.

explains why width can barely reach α, and Proposition 1 implies that whatever remains must come from drift in the pre-activation distribution, since nothing else can reach α. Evaluating the two-moment decomposition at each model's own measured drifts accounts for the slope to within 12% with nothing fitted. It also correctly predicts that ReLU's slope has the *opposite sign* (-0.158 measured), because $\partial \log \alpha / \partial k$, with $k = \mu/\sigma$, changes sign between the two activations. A finite-size explanation fails on separate grounds, since the $O(1/D)$ correction has the wrong sign and is several times too small. Appendices I and J give both arguments in full.

5.2 The concentration precondition, measured

Corollary 1 holds only where ν_D concentrates, and §3.2 assumes this of trained networks without establishing it. The strength of the shared component can be recovered from statistics the instrument already accumulates. For equicorrelated units, $R = \mathrm{Var}_t(\mu_t)/\mathbb{E}_t[\sigma_t^2]$ gives $\rho = R/(1+R)$, where μ_t and σ_t are the within-token mean and standard deviation that the density model of §5.3 already averages. The measurement therefore costs only forward passes. Table 7 and Figure 4 report it for eleven pretrained models, and Appendix D gives the full measurement and its two caveats.

The shared component is small in every model, from 0.008 to 0.079 with a median of 0.035. That shifts α by under 0.1% at the median and 0.4% at worst, compared with the 7.7% shift at $\rho = 0.2$, so the precondition holds with a wide margin. The shared component also *falls* as the layer widens, with $\mathrm{d}\log\rho/\mathrm{d}\log d_{\mathrm{ff}} = -0.64$ across the eleven models. The concentration that Corollary 1 needs therefore strengthens in the direction in which the corollary takes its limit.

5.3 Activation density follows from the nonlinearity

Suppose the pre-activations a token induces across the d_{ff} units are approximately independent draws from a common distribution. The law of large numbers applied to Eq. 2 then gives

$$\alpha \;=\; \frac{(\mathbb{E}|h|)^2}{\mathbb{E}[h^2]}, \qquad h = \phi(z), \quad z \sim \mathcal{N}(\mu, \sigma^2). \tag{4}$$

By itself Eq. 4 is the limiting form of the participation ratio. Its content lies in the modelling assumption that z is Gaussian with approximately independent coordinates. The claim under test is that this assumption is *adequate*, meaning that two moments of the pre-activation, together with ϕ, determine a nonlinear functional of $\phi(z)$. The claim is falsifiable, and we report below where it fails.

Table 5 in Appendix C tests the claim on models that differ only in the activation function, with three seeds per arm. Measured densities range from 0.178 for ReLU to 0.635 for GELU, a factor of 3.6.

Evaluated at the measured within-token mean and standard deviation, Eq. 4 predicts ReLU to 0.55% and GELU to 0.43%, or 0.49% on average. Figure 2 extends the comparison to the eleven pretrained models, whose expansion ratios run from 2.667 to 5.833.

5.4 A PREDICTION WE REGISTERED IN ADVANCE, AND WHY IT FAILED

We expected more. Equation 4 reduces to $1/\pi$ for ReLU and $2/\pi$ for the identity, apparently with no free parameter, because positive homogeneity makes α independent of σ. We therefore predicted that a trained ReLU network would measure $1/\pi \approx 0.3183$ regardless of its pre-activation statistics, and trained one to check. It measured 0.1781, a 79% error.

The error was in our analysis. Homogeneity gives invariance to the pre-activation *scale* and says nothing about its *mean*. Trained networks sit at $\mu/\sigma \approx -0.55$, which pushes most units below threshold and pulls density well below $1/\pi$. The parameter-free constants describe the zero-mean case, which holds at initialization and which training removes. This marks where existing initialization analyses (Price et al., 2024; Wolinski & Arbel, 2025) stop applying to trained models. The Gaussian assumption is not at fault. The failure lies in the vanishing mean that symmetry supplies at initialization.

5.5 REMOVING THE CAUSE MAKES THE PREDICTION TRUE

This diagnosis attributes the failure to one quantity, the pre-activation mean. To test it causally we removed that quantity, against predictions registered before the runs.

The within-token mean is a fixed linear functional of the first-layer weights. Writing $\bar{w} = \frac{1}{d_{\text{ff}}} \sum_i w_1^{(i)}$ for the mean row and $\bar{b}$ for the mean bias,

$$\mu_z \;=\; \frac{1}{d_{\text{ff}}} \sum_i \Big[\langle w_1^{(i)}, \mathrm{LN}(x) \rangle + b_i \Big] \;=\; \langle \bar{w}, \mathrm{LN}(x) \rangle + \bar{b}, \tag{5}$$

so constraining $W_1^\top \mathbf{1} = 0$, which forces the rows to average to the zero vector, and disabling the MLP bias gives $\mu_z = 0$ *exactly, for every token and every width*. The average runs over *units*, so the constraint differs from Weight Standardization, which averages over input dimensions. It costs d_{model} of $d_{\text{ff}} \times d_{\text{model}}$ parameters, about 0.1%. Appendix E gives the enforcement details.

By construction, the constraint closes the mean channel of the two-moment slope decomposition (Eq. 6, Appendix I). The scale channel is separate, and it is what lets the experiment discriminate. Positive homogeneity closes the scale channel identically. For a homogeneous nonlinearity both terms then vanish and the slope should reach zero; otherwise the scale term survives and the slope should fall only partway. We ran three arms with three different predictions. ReLU and ReGLU have both channels closed, the mean channel by the constraint and the scale channel by homogeneity, and should give slope 0 at $\alpha = 1/\pi$ and $2/\pi^2$ respectively. SwiGLU keeps its scale channel open, so its slope should fall only to $+0.027$.

SwiGLU answers the obvious objection that the constraint might flatten every slope it touches. ReGLU serves a different purpose. It is the only gated activation whose gate is positively homogeneous, so with the mean pinned and the scale irrelevant the independent-stream factorization gives $\alpha = \alpha_\phi \alpha_u = (1/\pi)(2/\pi) = 2/\pi^2$ with no free parameter, and any departure from it isolates the gate–up dependence. Table 8 and Figures 5 and 6 report all three arms, twenty-seven models built as in Tables 4 and 10 except for the constraint. Figure 5B confirms that the constraint holds in each layer individually.

Outcome. The two homogeneous arms landed at their predicted zero, and SwiGLU's slope stayed positive, as predicted, though above the forecast.

arm	slope before	slope predicted	slope measured	resolution
ReLU	-0.1583	0	$+0.0022$	± 0.0017
ReGLU	—	0	-0.0012	± 0.0008
SwiGLU	$+0.0371$	$+0.027$	$+0.0341$	± 0.0005

For ReLU over the fourfold width range, these slopes correspond to a -19.7% change in density before the constraint and $+0.30\%$ after. The two homogeneous-gate arms are indistinguishable from zero, at 1.3σ and 1.5σ. SwiGLU is not. Its slope stays positive, because its scale channel remains open and σ still drifts from 0.489 to 0.405, and the constraint lowers it only slightly, from $+0.0371$ in unconstrained models trained on the same hardware (Appendix K). A constraint that flattened every slope indiscriminately would have sent SwiGLU to zero as well.

On levels, ReLU returns to 0.3162 against $1/\pi = 0.3183$, an error of 0.65%, where the unconstrained network had missed by 79%. ReGLU gives 0.1922 against $2/\pi^2 = 0.2026$, an error of 5.1%. This is the first parameter-free prediction we have for a gated activation, and its shortfall can be measured directly: 56% of it is magnitude coupling between the gate and up streams and 38% the gate's own shape (Appendix F).

The other pre-registered checks also pass. Excess kurtosis stays within 0.04 of zero throughout, so once denied the mean the network does not recover a threshold through the shape of ν. The common-mode component collapses, with ρ_{shared} falling to 10^{-16} against 0.0199 in the unconstrained control. This also supplies the causal test of Corollary 1 that §3.2 could only state.

Scope of the result. The intervention shows that the mechanism is understood well enough to switch off on demand, completely where homogeneity closes the scale channel and only slightly where it does not. It is not a proposed repair. The constrained models are a different architecture, and the invariance claim of §5.1 stands as measured. Validation loss barely moves at any ratio, so at this scale the mean row that training builds appears close to free to remove. That leaves open, and sharpens, the question of why training builds it. Both homogeneous-gate arms leave a small unfitted residual, reported with the loss comparison in Appendix F, where the ReGLU residual is decomposed.

5.6 Where the model fails: gated activations

Gated activations carry a systematic residual of 10.2% for GEGLU in our controlled runs and 26–35% for SwiGLU in pretrained models, against roughly 0.5% for ReLU and GELU. For $h = \phi(g) \cdot u$ the prediction makes three assumptions: the gate is Gaussian at its measured moments, the up stream is Gaussian, and the two magnitudes are independent, so that density factorizes as $\alpha_\phi \cdot 2/\pi$. The exact per-token identity $\alpha = \alpha_\phi \, \alpha_u \, C_1^2/C_2$ of Appendix F tests each assumption separately. Its three factors account for the residual to within 2% of density in every model but Qwen2.5-1.5B, where they covary across tokens and leave 6% (Table 12). Under control, the GEGLU residual is almost entirely magnitude coupling between the streams ($C_1^2/C_2 = 0.936$, 88–99% of the gap), which the four tests of Appendix K, signed correlation among them, could not detect. In the six pretrained SwiGLU models the gate's and up stream's departures from Gaussian shape carry 55–91% of it, and coupling most of the rest. This locates the residual; why training couples the streams and distorts their shapes remains open.

6 Discussion

6.1 Utilization cannot set the expansion ratio

Sections 5.1 and 5.3 together show that how much of the FFN a model uses is a property of the activation function and is insensitive to width. An account of the expansion ratio that starts from utilization, holding that a ratio is right when it matches how much of the FFN gets used, therefore appeals to a quantity that does not move when the ratio does. Whatever fixes r has to come from the capacity side, from what additional width buys. In our runs that side is real and far from exhausted. Validation loss falls monotonically from 3.572 to 3.422 as r goes from 2 to 8, in the direction the original ablation found nine years ago (Appendix N).

One capacity-side argument works through interference between the active units and finds the cost of widening to be scale-free in width. Appendix L records it only as an argument, since we never measure interference directly.

For efficiency work. Methods that exploit activation sparsity inherit a constraint. Choosing a different FFN width cannot increase the sparsity available to them. That sparsity belongs to the activation function, and two measured moments of the pre-activation predict its magnitude.

6.2 Limitations

Our controlled experiments use models of 5–12M parameters trained on a synthetic children's-story corpus, and every causal claim in this paper rests on them. The pretrained measurements reach further but are confounded, so we cannot rule out that the decoupling behaves differently at frontier scale. The intervention of §5.5 shares this limit. Its twenty-seven models are no larger than the sweeps, so it demonstrates the mechanism at small scale and says nothing about frontier width. Equation 4 is an adequacy claim about a Gaussian model, not a derivation, and it fails for exactly the gated activations that current large models use. The invariance result of §5.1 does not depend on it, but the predictive claim of §5.3 is established for non-gated activations only. The interference argument of Appendix L is a sketch that relies on a strong readout assumption and has no direct measurement behind it. Activation density is also one notion of utilization among several, and a different measure might couple to width where this one does not. Density also shifts by 1.5–3% with the training hardware at matched validation loss, more than seed noise, and because the shift depends on width it moves the fitted slope as well (Appendix K); the comparison of §5.5 is therefore made on matching hardware.

The decomposition of ReGLU's 5.1% shortfall is measured on retrained replicas of the published models, which reproduce their densities to within 1.1%, and it locates the shortfall without explaining why training couples the two streams. The zero-mean-row constraint alters the architecture, so the constrained arms establish that the mean channel is causal and license no further claim about unconstrained networks. Finally, the common-mode measurement assumes an equicorrelated shared component and regresses across six laboratories with no controlled sweep. It bears on the precondition of Corollary 1, and not on the invariance itself.

7 Conclusion

Activation density in a Transformer FFN is nearly independent of its width. In controlled runs that vary only the expansion ratio, a fourfold change moves density by 7%. The cause is structural. Density is a functional of the empirical measure of the pre-activations, width can act on it only by moving that measure, and in trained networks the measure concentrates. The activation function, given two measured pre-activation moments, predicts density to 0.49% across non-gated nonlinearities whose densities differ by a factor of 3.6. Gated activations leave a residual, which we can locate in the coupling and shape of their two streams but not yet explain.

The small width dependence that remains is drift in those two moments. Constraining the first-layer rows to average to the zero vector removes the drift in the mean. ReLU density then returns to $1/\pi$ within 0.65%, and the width slope vanishes for both activations whose gate is positively homogeneous, as predicted in advance.

The implication for the expansion ratio is negative. A quantity that barely moves with width cannot select among widths. The case for the value of 4, in use for nine years, must rest on what added width buys in capacity, and in our runs validation loss still falls as the ratio grows.

Reproducibility statement

All results use public checkpoints and public datasets. Appendix A gives the analytic checks each estimator passes and the independent reimplementation it was verified against. Appendices B and C specify every trained model; the controlled sweeps fix the seed for both initialization and batch order, so each configuration sees byte-identical data, and seed standard deviations are reported throughout. Appendix M gives full proofs. Every quantitative claim traces to a data file produced by our code, and every figure regenerates from those files with a single script; the code will be released with the paper.

REFERENCES

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, et al. SmolLM2: When smol goes big — data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023. PMLR v202; arXiv:2304.01373.

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

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Zonglin Li, Chong You, Srinadh Bhojanapalli, Daliang Li, Ankit Singh Rawat, Sashank J. Reddi, Ke Ye, Felix Chern, Felix Yu, Ruiqi Guo, and Sanjiv Kumar. The lazy neuron phenomenon: On emergence of activation sparsity in transformers. *arXiv preprint arXiv:2210.06313*, 2022. A short version appeared at ICLR 2023.

OLMo Team. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025. COLM 2025.

Ilan Price, Nicholas Daultry Ball, Samuel C.H. Lam, Adam C. Jones, and Jared Tanner. Deep neural network initialization with sparsity inducing activations. In *International Conference on Learning Representations*, 2024. arXiv:2402.16184.

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

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, 2007.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, 2017.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Filip Szatkowski, Patryk Będkowski, Alessio Devoto, Jan Dubiński, Pasquale Minervini, Mikołaj Piórczyński, Simone Scardapane, and Bartosz Wójcik. Universal properties of activation sparsity in modern large language models. *arXiv preprint arXiv:2509.00454*, 2025. ICLR 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Pierre Wolinski and Julyan Arbel. Gaussian pre-activations in neural networks: Myth or reality? *Transactions on Machine Learning Research*, 2025. arXiv:2205.12379.

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

A INSTRUMENT VALIDATION

Each estimator is checked against cases with a known analytic answer before being applied to any model. The participation ratio of a one-hot vector is 1; of a uniform vector over n entries is n; of k equal non-zero entries is k; and it is invariant to rescaling. The spectral estimators return r for a rank-r identity embedded in d dimensions, 1 for a rank-one matrix, and approach d for i.i.d. Gaussian data when the sample count greatly exceeds d. All pass. A further check confirms the rank cap: with $N < d$ samples the estimate cannot exceed N, which is why the token budget must be kept well above the quantity being measured.

We implemented the spectral estimators twice and independently, once inside the main measurement script and once as a standalone diagnostic. On Pythia-70m the two agree to four decimal places, at $\mathrm{erank} = 251.6$.

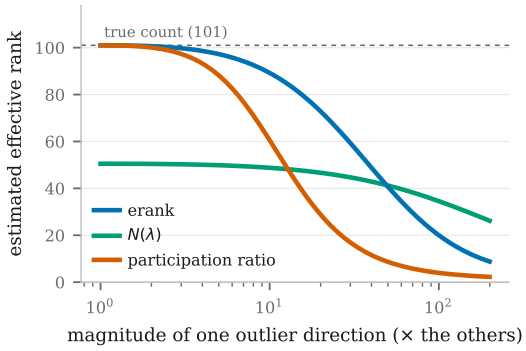

Figure 3: Why the participation ratio is unsafe as an effective-rank estimator. A synthetic spectrum of 100 equal directions plus a single larger one; the true count is 101. The participation ratio is variance-weighted and collapses once the outlier exceeds roughly $10\times$, while the entropy-based effective rank degrades far more slowly. $N(\lambda)$ (Rudi & Rosasco, 2017) is shown for reference and reads about half the true count for a flat spectrum by construction.

Table 1: Three effective-rank estimators on the same centred covariance spectrum, Pythia-70m, 16,384 tokens of WikiText. Layer 2 carries a massive activation and the participation ratio collapses accordingly. $N(\lambda) = \sum_k \lambda_k/(\lambda_k + \lambda)$ evaluated at the mean eigenvalue is shown for reference; it reads roughly half the true count for a flat spectrum by construction and so is comparable across models but is not a literal count.

layer	PR (centred)	erank	$N(\lambda)$	top-1% var	max $\lvert h \rvert$
0	60.6	255.7	417.0	10.2%	4.8
1	170.9	558.8	615.4	3.7%	10.2
2	**6.3**	**79.3**	428.1	**39.4%**	**138.5**
3	112.9	320.9	457.6	4.4%	7.2
4	35.9	161.9	404.6	11.2%	21.5
5	31.8	133.1	371.4	10.3%	31.6

Table 2: Null control. Effective rank per d_{model}, trained versus identical-architecture untrained models. The separation widens with scale and the two series move in opposite directions.

model	trained	random init	rel. diff
Pythia-70m	0.491	0.427	13.0%
Pythia-160m	0.498	0.351	29.6%
Pythia-410m	0.573	0.294	48.8%

Table 3: Largest FFN activation $\max|h|$ over all layers, at the library's default random initialization and after training, 16,384 tokens of WikiText. Every pre-norm architecture grows its peak 9- to 512-fold in training. OLMo-2 normalizes each block's output instead of its input, so its FFN sees the un-normalized residual stream and already peaks above its trained value at initialization. Random-initialization values are one draw each.

model	random init	trained	growth
Pythia-70m	2.61	138.5	53×
Pythia-160m	3.17	104.8	33×
Pythia-410m	3.73	108.4	29×
Pythia-1b	5.12	61.9	12×
Pythia-1.4b	5.53	98.2	18×
SmolLM2-360M	6.02	3084.4	512×
SmolLM2-1.7B	15.35	2984.5	194×
TinyLlama-1.1B	17.63	161.0	9×
Qwen2.5-0.5B	6.36	1933.6	304×
Qwen2.5-1.5B	12.45	5133.4	412×
OLMo-2-1B	470.36	171.5	0.36×

B THE CONTROLLED SWEEPS

Four sweeps are reported. The *expansion-ratio* sweep (Table 4) is the paper's central control: nine SwiGLU models varying only $r \in \{2, 4, 8\}$. The *nonlinearity* sweep (Table 5) holds $r = 4$ and varies the activation function over ReLU, GELU, GEGLU and SwiGLU. The *ReLU ratio* sweep (Table 10) repeats the first with a non-gated ReLU FFN, whose positive homogeneity makes one term of the slope decomposition vanish identically; its seeds are $2/3/2$ rather than $3/3/3$ because the sweep was interrupted. The *zero-mean-row* sweep (Table 8) adds twenty-seven models, three activation functions (ReLU, ReGLU, SwiGLU) by three ratios by three seeds, built exactly as the others except for the constraint of §5.5.

Training details. All models use the Hugging Face Llama implementation for gated FFNs and GPT-NeoX for non-gated ones, with $d_{\text{model}} = 256$, six layers and eight heads, and a byte-pair tokenizer with a vocabulary of 4096 trained on TinyStories. We train with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$) and weight decay 0.1 on weight matrices, with biases and norm parameters exempt. The learning rate warms up linearly to 3×10^{-4} over 150 steps and then follows a cosine decay to 10% of its peak. Gradients are clipped at norm 1.0, and each of the 2500 steps uses a batch of 32 sequences of 256 tokens. Validation loss is averaged over 20 fixed held-out batches, identical across runs.

C MEASUREMENTS

Table 4: Controlled expansion-ratio sweep. Nine models on TinyStories, $d_{\text{model}} = 256$, six layers, identical data order per seed; only r varies. Three seeds per ratio.

r	n	α	seed s.d.	αr	val. loss
2	3	0.3227	0.0020	0.645	3.5723
4	3	0.3338	0.0013	1.335	3.4955
8	3	0.3454	0.0006	2.763	3.4215

D THE COMMON-MODE COMPONENT

The measurement of §5.2 in full. ρ is the strength of the component shared across units within a token, recovered from the within-token moments the instrument already accumulates; $\mathrm{sd}_t(k_t)$ is its token-to-token fluctuation.

For scale, a slope of 0 would mean concentration never improves with width, and -1 is the rate a rank-one weight component of fixed magnitude would give.

Table 5: Controlled nonlinearity sweep. Models identical in data, depth, width, seed and schedule at $r = 4$, differing only in the activation function; three seeds per arm. SwiGLU is measured but its moment-conditioned prediction is not reported, as those runs predate the pre-activation instrumentation.

activation	n	α	seed s.d.	err @ μ=0, σ=1	err @ meas. (μ, σ)	μ/σ
ReLU (non-gated)	3	0.1781	0.0014	78.6%	**0.55%**	-0.555
GELU (non-gated)	3	0.6352	0.0055	41.1%	**0.43%**	-0.586
GEGLU (gated)	3	0.2921	0.0007	18.1%	10.19%	-0.065
SwiGLU (gated)	3	0.3338	0.0013	14.3%	—	—
mean, non-gated arms				59.8%	**0.49%**	

Table 6: Activation density across eleven pretrained models, 16,384 tokens of WikiText.

model	act.	d_{model}	r	depth	α
Pythia-70m	GELU	512	4.00	6	0.3704
Pythia-160m	GELU	768	4.00	12	0.3913
Pythia-1b	GELU	2048	4.00	16	0.4553
Pythia-1.4b	GELU	2048	4.00	24	0.4731
Pythia-410m	GELU	1024	4.00	24	0.4977
SmolLM2-360M	SwiGLU	960	2.667	32	0.3061
TinyLlama-1.1B	SwiGLU	2048	2.750	22	0.3197
OLMo-2-1B	SwiGLU	2048	4.000	16	0.2852
SmolLM2-1.7B	SwiGLU	2048	4.000	24	0.2908
Qwen2.5-0.5B	SwiGLU	896	5.429	24	0.3155
Qwen2.5-1.5B	SwiGLU	1536	5.833	28	0.2779

Two caveats attach to it. The estimator assumes the shared component is equicorrelated across units, so it quantifies dependence under a model of that dependence rather than characterizing its form; a component concentrated on a few units would register differently. And the eleven models differ in laboratory, data and recipe, so the width trend is a cross-family regression rather than a controlled sweep. Reassuringly, this pass reproduces the densities of Table 6 to within 2%, so the extended instrument agrees with the one used elsewhere in the paper.

Table 7: The common-mode component in trained networks, 16,384 tokens of WikiText. ρ is the strength of the component shared across units within a token and $\mathrm{sd}_t(k_t)$ its token-to-token fluctuation. For contrast, the constrained arms of §5.5 give $\rho \sim 10^{-16}$ by construction.

model	act.	d_{ff}	ρ	$\mathrm{sd}_t(k_t)$	α
Pythia-70m	GELU	2048	0.0794	0.146	0.3704
Pythia-160m	GELU	3072	0.0572	0.143	0.3913
Pythia-410m	GELU	4096	0.0345	0.135	0.4977
Pythia-1b	GELU	8192	0.0547	0.157	0.4553
Pythia-1.4b	GELU	8192	0.0414	0.135	0.4731
SmolLM2-360M	SwiGLU	2560	0.0414	0.167	0.3061
Qwen2.5-0.5B	SwiGLU	4864	0.0184	0.098	0.3155
TinyLlama-1.1B	SwiGLU	5632	0.0163	0.124	0.3197
OLMo-2-1B	SwiGLU	8192	0.0345	0.096	0.2852
SmolLM2-1.7B	SwiGLU	8192	0.0083	0.105	0.2908
Qwen2.5-1.5B	SwiGLU	8960	0.0215	0.114	0.2779

E THE ZERO-MEAN-ROW SWEEP

Per-arm measurements for the intervention of §5.5, and the check that the constraint holds at every layer rather than on average.

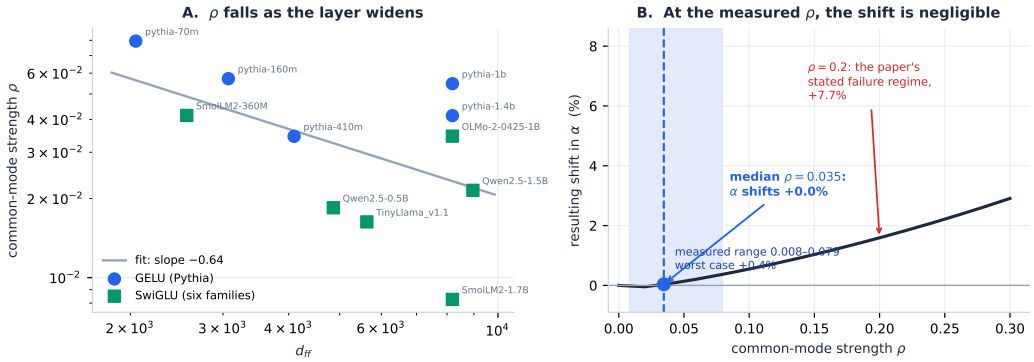

Figure 4: The precondition of Corollary 1, measured. **A.** The common-mode strength across eleven pretrained models, spanning d_{ff} from 2048 to 8960 and two activation families. It falls with width, slope -0.64 in logs. **B.** What that costs: reproducing the equicorrelated simulation of Appendix M, the shift in α at the measured median is below 0.1%, and 0.4% at the largest ρ observed, against $+7.7\%$ at the $\rho = 0.2$ where the corollary is stated to fail.

The constraint is enforced by projection at initialization, on the gradient, and after each optimizer step. The layer norm's β needs no treatment, since it reaches μ_z only through $\langle \bar{w}, \beta \rangle$, which vanishes with $\bar{w}$.

Table 8: The zero-mean-row constraint across three activation functions, three seeds per cell. k confirms the constraint held at every layer rather than on average. Skewness is reported because, with the mean pinned, it is the only remaining route to a threshold, and it is the moment the existing instrumentation did not record. σ still drifts in every arm while α moves only for SwiGLU, which exhibits homogeneity rather than assuming it.

arm	r	α	seed s.d.	k	σ	skew	exc. kurt
ReLU	2	0.3164	0.0004	-5.5×10^{-10}	0.383	$+0.031$	-0.009
(non-gated)	4	0.3162	0.0009	$+4.0 \times 10^{-11}$	0.365	$+0.023$	$+0.010$
	8	0.3174	0.0003	-1.5×10^{-10}	0.352	$+0.007$	$+0.008$
ReGLU	2	0.1929	0.0013	-1.1×10^{-10}	0.466	$+0.088$	-0.040
(ReLU gate)	4	0.1922	0.0003	-1.9×10^{-10}	0.431	$+0.087$	-0.031
	8	0.1926	0.0004	-6.0×10^{-10}	0.402	$+0.079$	-0.011
SwiGLU	2	0.3286	0.0009	$+4.5 \times 10^{-10}$	0.489	$+0.023$	$+0.038$
(SiLU gate)	4	0.3383	0.0003	$+2.3 \times 10^{-9}$	0.441	$+0.042$	$+0.029$
	8	0.3445	0.0019	-1.4×10^{-10}	0.405	$+0.035$	$+0.037$

F THE RESIDUALS OF THE CONSTRAINED ARMS

Validation loss is essentially unaffected by the constraint: constrained SwiGLU gives 3.5747, 3.4970, 3.4227 against 3.5723, 3.4955, 3.4215 unconstrained, within seed noise at every ratio. ReGLU costs about 0.015 in loss relative to SwiGLU and is used here as an instrument, not proposed as an architecture. The common-mode collapse is to $\rho_{\text{shared}} \sim 10^{-16}$ and $\mathrm{sd}_t(k_t) \sim 10^{-8}$ in every arm, against 0.0199 and 0.143 unconstrained.

Two residuals are reported as measured rather than fitted. ReLU sits 0.65% below $1/\pi$, larger than its 0.28% seed noise and consistent in sign with the measured positive skew, which concentrates magnitude into fewer units; the finite-width correction of Proposition 2 has the opposite sign and is an order of magnitude smaller. ReGLU's 5.1% shortfall is the more interesting of the two, and it can be measured rather than inferred. Within a token, write $\alpha_\phi = (\mathbb{E}|\phi|)^2/\mathbb{E}[\phi^2]$ and $\alpha_u = (\mathbb{E}|u|)^2/\mathbb{E}[u^2]$ for the gate and up streams, and $C_1 = \mathbb{E}[|\phi||u|]/(\mathbb{E}|\phi|\,\mathbb{E}|u|)$, $C_2 = \mathbb{E}[\phi^2 u^2]/(\mathbb{E}[\phi^2]\,\mathbb{E}[u^2])$, both unity when the two magnitudes are independent. Then $\alpha = \alpha_\phi \alpha_u C_1^2/C_2$ holds exactly for every

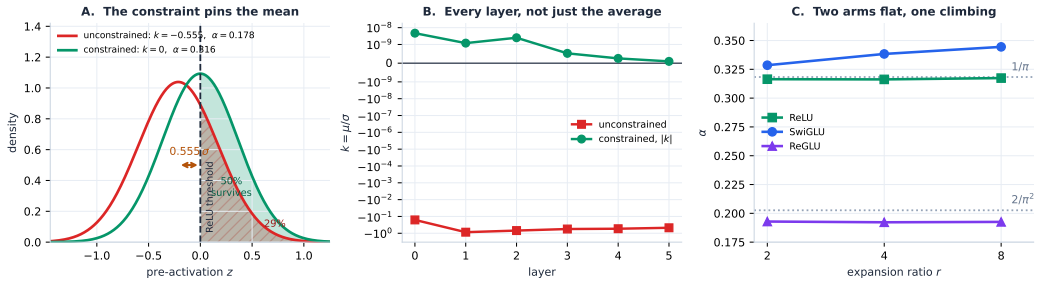

Figure 5: What the constraint does. **A.** Training places the pre-activation mean 0.555σ below the ReLU threshold, so 29% of units survive and $\alpha = 0.178$; the constraint returns the mean to zero, half the units survive, and $\alpha = 0.316$. Curves are the Gaussian model at the measured moments. **B.** The constraint holds *per layer*, not on average: $|k|$ falls from 0.16–0.86 to below 5×10^{-9} at every layer. Note the symmetric-log axis. **C.** The three constrained arms against their parameter-free closed forms.

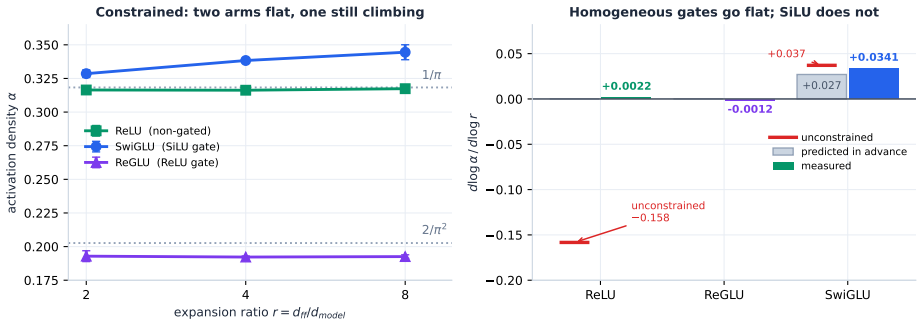

Figure 6: The result, and why three arms rather than one. **Left:** the two arms whose gate is positively homogeneous are flat in width and sit on their closed forms; SwiGLU, whose scale channel stays open, still climbs. Error bars are ± 3 seed standard deviations. **Right:** the slopes. The two homogeneous arms land on their pre-registered zero; SwiGLU stays positive, as predicted, above the forecast $+0.027$. Unconstrained slopes are -0.158 for ReLU and $+0.037$ for SwiGLU trained on the same hardware.

token, and the independent-Gaussian prediction is $\alpha_\phi = 1/\pi$, $\alpha_u = 2/\pi$ and $C_1 = C_2 = 1$. The trained weights of the published models were not retained, so Table 9 measures each factor on replicas retrained with the same recipe and seeds; they reproduce the published densities to within 1.1% per model and 0.2% on average. In log terms, magnitude coupling accounts for 56% of the shortfall ($C_1^2/C_2 = 0.970$), the gate's departure from a Gaussian shape for 38% (α_ϕ at 0.980 of $1/\pi$, in the direction its skewness of $+0.08$ predicts), and the up stream for the remaining 6%. The shares barely move with width, although both couplings weaken as r grows. The signed stream correlation, 0.002 to 0.014 in this arm, does not constrain C_1 or C_2, which are magnitude couplings, and this is why the elimination of §5.6 could not detect them.

G ReLU EXPANSION-RATIO SWEEP

A second controlled sweep, identical in construction to Table 4 but with a non-gated ReLU FFN. Seeds are $2/3/2$ rather than $3/3/3$: the sweep was interrupted, leaving one run without a saved configuration and one unstarted. The effect being measured is roughly $100\times$ the seed standard deviation.

For reference, the SwiGLU sweep's own moments, measured the same way: $k = +0.027, -0.023, -0.059$ and $\sigma = 0.493, 0.443, 0.407$ at $r = 2, 4, 8$, giving $\mathrm{d}k/\mathrm{d}\log r = -0.0615$ and $\mathrm{d}\log\sigma/\mathrm{d}\log r = -0.1371$.

Table 9: Exact per-token decomposition $\alpha = \alpha_\phi\, \alpha_u\, C_1^2/C_2$ for the constrained ReGLU arm, each factor relative to its independent-Gaussian value, averaged over three seeds of retrained replicas. The last column is $\alpha/(2/\pi^2)$; the product of the three ratios before it matches it to within 0.2%, the gap being the token-to-token covariance of the factors.

	α	$\pi\alpha_\phi$	$\frac{\pi}{2}\alpha_u$	C_1	C_2	C_1^2/C_2	$\frac{\pi^2}{2}\alpha$
$r = 2$	0.1914	0.977	0.997	1.054	1.183	0.968	0.944
$r = 4$	0.1922	0.980	0.997	1.035	1.128	0.971	0.949
$r = 8$	0.1929	0.981	0.998	1.023	1.087	0.972	0.952
all	0.1922	0.980	0.997	1.037	1.133	0.970	0.948

Table 10: ReLU controlled expansion-ratio sweep, with the pre-activation moments that Eq. 6 requires. $k = \mu/\sigma$.

r	n	α	seed s.d.	k	σ
2	2	0.1954	0.0015	-0.476	0.407
4	3	0.1781	0.0014	-0.555	0.384
8	2	0.1569	0.0014	-0.669	0.369

H Effective rank, depth and width

Effective rank per d_{model} across the Pythia ladder. Sorted by depth the sequence is monotonic; sorted by width it is not. Table 11 gives the numbers. They support the width-invariance direction of the interference sketch in Appendix L, and separately indicate a depth dependence for which we have no account.

Pythia-1.4b served as a pre-registered test: it shares its width with Pythia-1b and its depth with Pythia-410m, so the two hypotheses give different predictions. We recorded 0.573 for depth and 0.538 for width before measuring, and observed 0.582.

Table 11: Effective rank per d_{model} across the Pythia ladder.

model	d_{model}	depth	erank/d_{model}
Pythia-70m	512	6	0.491
Pythia-160m	768	12	0.498
Pythia-1b	2048	16	0.538
Pythia-410m	1024	24	0.573
Pythia-1.4b	2048	24	0.582

I The residual slope is drift in the pre-activation moments

The residual slope is not a finite-size effect. A natural objection is that the $+0.049$ slope is simply the $O(1/D)$ error in Corollary 1. It is not, on two independent grounds.

Proposition 2 in Appendix J gives that correction to $O(1/D)$. Evaluated at our measured moments it has the wrong sign. For ReLU, $c = +2.56$, so Eq. 7 predicts α *decreasing* with D where we measure an increase, and at $D \approx 1000$ its magnitude is several times too small in any case. Independently, our per-doubling log-slopes are 0.0488 and 0.0493, constant, i.e. a power law, whereas a $1/D$ term must halve with each doubling. The residual is drift in the pre-activation moments, not a finite-size artefact.

Proposition 1 says it must come from drift in the pre-activation distribution, since nothing else can reach α, and Proposition 2 rules out a finite-size artefact. That leaves the two moments, and both are measurable:

$$\frac{\mathrm{d}\log\alpha}{\mathrm{d}\log r} = \frac{\partial\log\alpha}{\partial k}\frac{\mathrm{d}k}{\mathrm{d}\log r} + \frac{\partial\log\alpha}{\partial\log\sigma}\frac{\mathrm{d}\log\sigma}{\mathrm{d}\log r}, \qquad k \equiv \mu/\sigma. \tag{6}$$

We test this with a second controlled sweep using ReLU, chosen because it is positively homogeneous, so $\partial \log \alpha / \partial \log \sigma \equiv 0$ exactly and Eq. 6 collapses to its first term. Its predictions are sharply different from SwiGLU's: $\partial \log \alpha / \partial k$ is $+1.27$ for ReLU against -0.27 for SwiGLU, so the two slopes should have *opposite signs*. They do. Table 10 gives the sweep: ReLU's slope is -0.158, and its pre-activation mean drifts monotonically negative with width, $k = -0.476, -0.555, -0.669$ at $r = 2, 4, 8$.

Evaluated at each model's own measured drifts and sensitivities, Eq. 6 predicts -0.1766 for ReLU against -0.1583 measured (mean term -0.1766, scale term identically zero), and $+0.0435$ for SwiGLU against $+0.0490$ measured (mean term $+0.0168$, scale term $+0.0267$). Both are accounted for to within 12% with no fitted quantity, and the sign difference follows entirely from $\partial \log \alpha / \partial k$ changing sign between the two activations. Note the decomposition is not the same in both cases: for ReLU the scale term vanishes identically and the slope is pure mean-drift, whereas for SwiGLU the scale term is the larger of the two. What remains unexplained sits one level down. Why training drifts the pre-activation moments with width at all is a question we do not address.

J THE FINITE-SIZE CORRECTION

Proposition 2 (Finite-size correction)**.** *For i.i.d. pre-activations with $m_1 = \mathbb{E}|h|$, $m_2 = \mathbb{E}[h^2]$,*

$$\mathbb{E}[\alpha_D] \;=\; \alpha_\infty \left(1 + \frac{c}{D} + O(D^{-2})\right), \quad c = \frac{\mathrm{Var}|h|}{m_1^2} - \frac{2\,\mathrm{Cov}(|h|, h^2)}{m_1 m_2} + \frac{\mathrm{Var}(h^2)}{m_2^2}. \tag{7}$$

Proof sketch. Second-order delta method on $f(S_1, S_2) = S_1^2/S_2$ about (m_1, m_2), with $\mathrm{Var}(S_k) = O(1/D)$. Full proof in Appendix M.

Evaluated at our measured moments, $c = +2.56$ for ReLU. That is positive, so Eq. 7 predicts α decreasing toward α_∞ as D grows, while our slope is positive; the correction has the wrong sign, and at $D \approx 1000$ its magnitude is several times too small in any case. A second argument is independent of the first. Our measured per-doubling log-slopes are 0.0488 and 0.0493, constant, which is a power law, whereas a $1/D$ term must halve with every doubling of D. The residual is moment drift and not a finite-size artefact.

K THE GATED-ACTIVATION RESIDUAL

Gated activations carry a systematic residual: 10.2% for GEGLU in our controlled runs and up to 26–35% for SwiGLU in pretrained models, against $\approx 0.5\%$ for ReLU and GELU. For $h = \phi(g) \cdot u$ with g and u independent, density should factorize as $\alpha_\phi \cdot 2/\pi$. We tested four ways that independence could fail and eliminated all four: a mean offset in the up-projection stream (measured -0.002 to 0.000); signed correlation between the streams (-0.002 to $+0.001$); non-Gaussianity (GEGLU's excess kurtosis is 0.10, essentially Gaussian, and it carries the residual regardless); and per-unit scale heterogeneity from correlated projection row norms, which is present in the right direction ($\mathrm{corr}(\|w_{\text{gate},i}\|, \|w_{\text{up},i}\|) = +0.09$ to $+0.16$) but moves the prediction away from the measurement when simulated with the empirical norm pairs.

None of these tests can see a dependence between the *magnitudes* of the two streams, and none examined the up stream's shape. The per-token identity $\alpha = \alpha_\phi\, \alpha_u\, C_1^2/C_2$ of Appendix F measures both. Relative to the prediction, the gate factor is α_ϕ against its Gaussian value at the measured moments, the up factor is α_u against $2/\pi$, and the coupling factor is C_1^2/C_2 against 1. Table 12 reports them for the controlled GEGLU arm and all six pretrained SwiGLU models. The three factors reproduce the residual to within 2% of density except in Qwen2.5-1.5B, whose peak activation is the largest we measure (Table 3); there the factors covary across tokens and leave 6% unaccounted, so its shares are indicative only.

The GEGLU arm is measured on replicas retrained with the same recipe and seeds, and their densities sit 1.3–2.7% above the published ones. This offset reflects the training hardware. Retraining the SwiGLU sweep of §5.1 with the same recipe and seeds on the GPU hardware used for the constrained arms, where our replicas reproduce the constrained ReGLU densities to 0.2%, reproduces its validation losses exactly (3.5723, 3.4955 and 3.4215 at $r = 2, 4, 8$) but raises its density by 3.2%,

2.5% and 1.5%. The shift depends on width, so it also changes the fitted slope, to $+0.0371$ (seed s.d. 0.0008) from $+0.049$. Density still moves only 5% over the fourfold range, so the width-invariance of §5.1 holds on both, and the comparison of §5.5 uses $+0.0371$, the value on matching hardware. The GEGLU replicas' residual is accordingly 7.4% where the published models give 10.2%; the share carried by coupling is the same in all three seeds, and it is that share, not the size, that we report.

Table 12: Decomposition of the gated residual. $\hat{\alpha}$ is the two-moment prediction; the gate, up and coupling columns are the measured factors relative to their values under that prediction, whose product is $\alpha/\hat{\alpha}$ up to token-to-token covariance. The last column gives each factor's share of $\log(\alpha/\hat{\alpha})$ as gate/up/coupling, in percent. †Mean over three retrained replicas. ‡The factors leave 6% of density unaccounted, so these shares do not sum to 100.

model	α	$\hat{\alpha}$	residual	gate	up	coupling	shares
GEGLU, controlled†	0.2984	0.3204	7.4%	0.996	1.000	0.936	6/0/93
SmolLM2-360M	0.3061	0.3874	26.6%	0.868	0.930	0.981	60/31/8
OLMo-2-1B	0.2852	0.3661	28.4%	0.931	0.910	0.935	29/38/27
Qwen2.5-0.5B	0.3155	0.3964	25.7%	0.883	0.923	0.996	54/35/2
SmolLM2-1.7B	0.2908	0.3819	31.3%	0.899	0.935	0.906	39/25/36
TinyLlama-1.1B	0.3197	0.4199	31.3%	0.912	0.945	0.883	34/21/46
Qwen2.5-1.5B‡	0.2779	0.3750	34.9%	0.860	0.906	1.012	50/33/−4

L AN INTERFERENCE SKETCH

One tempting capacity-side argument for the expansion ratio runs through interference. We record it here as a sketch, not as a result. Projecting the FFN output onto column j recovers h_j plus $\sum_{i\neq j} h_i \langle w_2^{(i)}, w_2^{(j)}\rangle$; assuming unit-norm generic columns with $\mathbb{E}[\langle w_2^{(i)}, w_2^{(j)}\rangle^2] = 1/d_{\text{model}}$ and $k = \alpha d_{\text{ff}}$ active units of comparable magnitude and independent sign, the signal-to-interference ratio is $\sqrt{\alpha r}$, in which d_{model} cancels. With §5.1 this would make interference scale-free in width. Consistent with that, our effective-rank measure moves 1.6% when width doubles at fixed depth against 8.2% when depth changes at fixed width (Appendix H). The argument is indirect on every count. We never measure interference, the readout assumption is strong, and what is priced here is only the cost of widening; the benefit sits outside it (§6.1). Figure 7 gives the geometry.

M PROOFS

PROPOSITION 1 (DENSITY IS A FUNCTIONAL OF THE EMPIRICAL MEASURE)

Fix a token and let $a_1, \ldots, a_D$ be the pre-activations it induces across the $d_{\text{ff}} = D$ units, with $h_i = \phi(a_i)$. Let $\nu_D = \frac{1}{D}\sum_{i=1}^{D} \delta_{a_i}$.

By definition of ν_D, for any measurable g, $\int g \,\mathrm{d}\nu_D = \frac{1}{D}\sum_i g(a_i)$. Applying this to $g = |\phi|$ and $g = \phi^2$,

$$\sum_i |h_i| \;=\; D\int |\phi| \,\mathrm{d}\nu_D, \qquad \sum_i h_i^2 \;=\; D\int \phi^2 \,\mathrm{d}\nu_D.$$

Substituting into Eq. 2,

$$\alpha_D = \frac{\left(\sum_i |h_i|\right)^2}{D\sum_i h_i^2} = \frac{D^2\left(\int |\phi| \,\mathrm{d}\nu_D\right)^2}{D \cdot D\int \phi^2 \,\mathrm{d}\nu_D} = \frac{\left(\int |\phi| \,\mathrm{d}\nu_D\right)^2}{\int \phi^2 \,\mathrm{d}\nu_D} = \Psi[\nu_D].$$

The three factors of D cancel exactly. No distributional assumption is used, and the identity holds for every D and every realisation.

Numerically, α_D and $\Psi[\nu_D]$ agree to machine precision ($< 10^{-12}$) at $D = 256, 1024, 4096$.

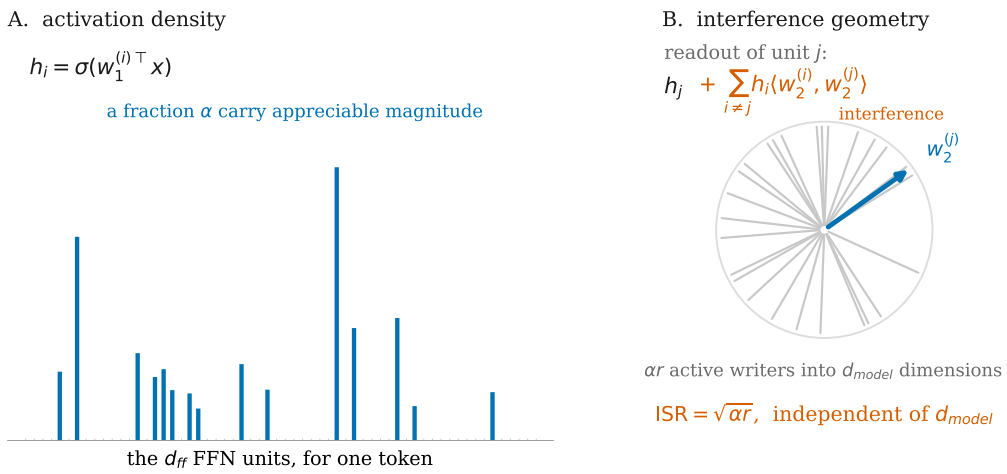

Figure 7: The interference sketch, recorded as an argument and not a result. **A.** For a given token a fraction α of the d_{ff} units carries appreciable magnitude. **B.** Those $\alpha\, d_{\text{ff}}$ active units each write a vector into a d_{model}-dimensional space; because $\alpha r > 1$ there are more active writers than dimensions, so reading a unit back recovers its own activation plus interference from the others. Under the assumptions stated in the text that interference scales as $\sqrt{\alpha r}$, in which d_{model} cancels. We do not measure interference directly.

COROLLARY 1 (WIDTH-INVARIANCE)

Weak convergence moves integrals of bounded continuous functions only, and $|\phi|$ and ϕ^2 are unbounded for ReLU, GELU and SiLU; the integrability condition closes that gap. Since ϕ is continuous and ϕ^2 is uniformly integrable under $\{\nu_D\}$, weak convergence gives $\int \phi^2 \,\mathrm{d}\nu_D \to \int \phi^2 \,\mathrm{d}\nu$. Because $|\phi| \le (1+\phi^2)/2$, $|\phi|$ is uniformly integrable too, so $\int |\phi| \,\mathrm{d}\nu_D \to \int |\phi| \,\mathrm{d}\nu$. The map $(m_1, m_2) \mapsto m_1^2/m_2$ is continuous wherever $m_2 > 0$, and $\int \phi^2 \,\mathrm{d}\nu > 0$, so $\Psi[\nu_D] \to \Psi[\nu]$; by Proposition 1, $\alpha_D = \Psi[\nu_D]$. The limit does not involve D.

The integrability condition has a concrete meaning here. It fails when a vanishing fraction of units carries a non-vanishing share of $\sum_i h_i^2$, since then $\frac{1}{D}\sum_{i:\, h_i^2 > K} h_i^2$ does not vanish as K grows. Massive activations are the natural candidate, so what the corollary needs of them is that their share of $\sum_i h_i^2$ shrinks as the layer widens.

On Corollary 1 and dependence. Because ν_D is an empirical measure, α_D is blind to the *order* of the pre-activations and to any dependence structure that leaves ν_D unchanged. It is not, however, blind to dependence that prevents ν_D from concentrating. Take equicorrelated pre-activations $a_i = \mu + \sigma(\sqrt{\rho}\,\xi + \sqrt{1-\rho}\,\epsilon_i)$ with ξ, ϵ_i standard normal: the marginal law of each a_i is $\mathcal{N}(\mu, \sigma^2)$ for every ρ, but the shared component ξ shifts the whole empirical measure together, so ν_D does not converge to the marginal. Simulation at $D = 4096$, $\phi = \mathrm{ReLU}$, $\mu/\sigma = -0.555$ gives $\alpha = 0.1745$ at $\rho = 0$ and 0.1879 at $\rho = 0.2$, a 7.7% shift; at $\rho = 0.8$ the quantity becomes ill-defined on a non-negligible fraction of draws, since a large negative common shift can zero every unit.

This bears on how §3.2 should be read. Corollary 1 supplies a mechanism for width-invariance, but the invariance itself is an empirical fact about trained networks, namely that they sit in the weak-dependence regime, and not an identity holding regardless. A network with strong common-mode pre-activation correlation would violate it.

PROPOSITION 2 (FINITE-SIZE CORRECTION)

Let $S_1 = \frac{1}{D}\sum_i |h_i|$ and $S_2 = \frac{1}{D}\sum_i h_i^2$, so that $\alpha_D = S_1^2/S_2 = f(S_1, S_2)$ with $f(u, v) = u^2/v$. Write $m_1 = \mathbb{E}|h|$, $m_2 = \mathbb{E}[h^2]$, $\delta_k = S_k - m_k$.

The partial derivatives of f at (m_1, m_2) are

$$f_1 = \frac{2m_1}{m_2}, \quad f_2 = -\frac{m_1^2}{m_2^2}, \quad f_{11} = \frac{2}{m_2}, \quad f_{12} = -\frac{2m_1}{m_2^2}, \quad f_{22} = \frac{2m_1^2}{m_2^3}.$$

A second-order Taylor expansion gives

$$\mathbb{E}[f] = \frac{m_1^2}{m_2} + \tfrac{1}{2} f_{11}\mathrm{Var}(S_1) + f_{12}\mathrm{Cov}(S_1, S_2) + \tfrac{1}{2} f_{22}\mathrm{Var}(S_2) + O(D^{-2}),$$

since $\mathbb{E}[\delta_k] = 0$. Dividing through by $\alpha_\infty = m_1^2/m_2$,

$$\frac{\mathbb{E}[\alpha_D]}{\alpha_\infty} = 1 + \frac{\mathrm{Var}(S_1)}{m_1^2} - \frac{2\,\mathrm{Cov}(S_1, S_2)}{m_1 m_2} + \frac{\mathrm{Var}(S_2)}{m_2^2} + O(D^{-2}).$$

For i.i.d. units, $\mathrm{Var}(S_1) = \mathrm{Var}|h|/D$, $\mathrm{Cov}(S_1, S_2) = \mathrm{Cov}(|h|, h^2)/D$ and $\mathrm{Var}(S_2) = \mathrm{Var}(h^2)/D$, which yields Eq. 7.

Numerical check and the sign. At $\phi = \mathrm{ReLU}$, $\mu/\sigma = -0.555$: the analytic values are $\alpha_\infty = 0.17411$ and $c = +2.5552$. Monte Carlo over 4000 draws gives $\alpha_D = 0.17737$, 0.17490, 0.17439 at $D = 128, 512, 2048$, against Eq. 7 predictions of 0.17759, 0.17498, 0.17433. Every value lies *above* α_∞ and decreases toward it, confirming $c > 0$: the finite-size correction makes α a *decreasing* function of D, opposite in sign to the slope measured in §5.1.

N WHERE THE EXPANSION RATIO CAME FROM

The usual telling has it that the ratio was never examined, which is not quite right. Vaswani et al. (2017) give no justification in the text, writing only that "the inner-layer has dimensionality $d_{ff} = 2048$". Their Table 3 does vary it, however. At $d_{\text{model}} = 512$ they report $d_{\text{ff}} \in \{1024, 2048, 4096\}$, that is $r \in \{2, 4, 8\}$, with development perplexities of 5.12, 4.92 and 4.75, BLEU of 25.4, 25.8 and 26.2, and parameter counts of 53M, 65M and 90M.

The largest ratio was best on both metrics and was not adopted. The base model kept $r = 4$ at 65M parameters. The constant is therefore a parameter-economy choice made at one model size on one translation task, not a measured optimum, and as far as we can determine it has not been revisited.

The gated variants' $r = 8/3$ is more mechanical still: Shazeer (2020) reduces d_{ff} by exactly two thirds when a third projection is introduced, so as to hold the parameter count fixed against the baseline.