OUTPUT WITHOUT WARRANT: SPECTRAL DIMENSION READOUTS CANNOT CERTIFY THEIR OWN HYPOTHESES

Anonymous authors
Paper under double-blind review

ABSTRACT

Local intrinsic dimension is routinely read off the local covariance spectra of neural representations. We distinguish a readout's output, the index it returns, from its warrant, the grounds for reading that index as a dimension. The output is cheap: for any data law, a spectral readout recovers its own population target at a sample size set by effective rank and spectral margin. The warrant is not. The hypotheses under which that target equals the dimension are lower bounds on reach, which no procedure can certify from a window of fewer than $(2^d - 1)\log(1/2\delta)$ samples under any hypothesis-independent observation channel. Moreover, the covariance does not identify dimension at all, so a stable readout can be wrong for every dimension but one. In planted windows matched to BERT and GPT-2 representations, the readout tracks where a decaying spectrum crosses the noise edge, moving by two orders of magnitude at fixed structure. We replace the dimension claim with a narrower one: a profile-depth statistic, calibrated against an anisotropy-matched null with measured power, that certifies detectable structure but not its dimension.

1 INTRODUCTION

The geometry of neural representations has become a standard object of study. Intrinsic dimension in particular is estimated layer by layer and read as evidence about abstraction, compression and generalisation (Ansuini et al., 2019; Valeriani et al., 2023; Razzhigaev et al., 2024; Cheng et al., 2025). A large share of these estimates come from the *local covariance spectrum*: collect the n_r points nearest an anchor, form their covariance, and read a dimension off its eigenvalues, as the index of the largest gap, the number of components needed to reach a fixed fraction of variance, or a derived spectral quantity. The practice rests on a picture in which a window of a d-dimensional manifold has d large tangent eigenvalues separated by a gap from a small normal block, so that the gap reveals d.

This paper asks when that picture licenses the readout. We separate two questions that the practice runs together. A readout has an *output*, the index it returns, and a *target*, the population quantity that output converges to. Reading the output as a dimension requires both that the output has reached its target and that the target is the dimension. We call the grounds for the second the readout's *warrant*. Our theory shows that the two requirements have very different sample costs, and our experiments show that on transformer representations it is the second that fails. Every measurement requires assumptions. Our claim is *not* that dimension estimation requires assumptions; it is that the *assumptions required to interpret a spectral readout as dimension cannot themselves be validated from the data consumed by the readout*.

Warrant cannot be certified from the window. Under a smooth manifold model with bounded third-order behaviour, the max-gap rule returns the local dimension once the window resolves its own spectrum (Theorem 3.1), while no estimator recovers principal curvature below $n_r = (2^d - 1)\log\frac{1}{2\delta}$ (Theorem 3.2). We show that this separation is an instance of a general dichotomy. A local functional pays the exponential cost exactly when it can be changed by modifying the window law on a set of mass 2^{-d} (Lemma Q.2). Curvature can; the covariance spectrum cannot (Proposition 4.1). The hypotheses under which Theorem 3.1 holds are lower bounds on reach, and reach

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.

bounds curvature pointwise, so certifying them from the window costs the same $(2^d - 1)\log\frac{1}{2\delta}$ samples (Theorem 4.2). A spectral readout thus sees, up to an exponentially small perturbation, the same spectrum on a warranted and an unwarranted window, and cannot validate itself (Corollary 4.3). This is the sense in which curvature is the right benchmark: a spectral readout never needs curvature to produce its output, but it needs a bound on curvature to be warranted.

Misspecification does not rescue the readout. Transformer representations violate the manifold model conspicuously: they are strongly anisotropic, heavy-tailed and cone-shaped (Ethayarajh, 2019; Sun et al., 2024). The two halves of the theory respond to this in opposite ways. The lower bounds survive any observation channel that does not depend on the hypothesis, with unchanged constants (Theorem 5.1). The dimension threshold does not survive, but it is replaced by a threshold defined for every law, at which a spectral readout recovers its own population target, whatever that target is (Theorem 5.3). Together these give a decomposition of readout error that needs no manifold assumption: a variance term that vanishes above the stability threshold, and a bias term, whether the target equals d, that no window can rule out by itself (Corollary 5.4). The bias term is not a technicality: for any covariance, every dimension up to its rank is consistent with an ideal latent manifold observed through a Gaussian channel, so a stable readout is wrong for all but one of them (Theorem 5.5).

Empirically, the target is not the structure. On a spectrum that decays through a noise edge with no gap, a split-point rule's target is the point where the spectrum crosses the edge, a function of signal-to-noise and decay exponent alone (Proposition 6.1). We confirm this directly: at fixed structure above the edge, the readout moves by nearly two orders of magnitude as the noise model changes (§N.2). Disjoint-half subsampling shows that the real windows do *not* clear the stability threshold (§N.7), so the variance explanation cannot be excluded; what the data do exclude is the warrant, since at nearly all the windows we test no candidate dimension meets the scope condition. The theory also says what an admissible alternative must supply: a reference external to the window. Our profile-depth gate is scored against an anisotropy-matched null with measured power (§7), and every absence claim it makes carries the disjunction its power licenses.

1.1 Contributions

We contribute, first, a formal separation of *output* from *warrant* (§3,4): a cap-modulus lemma distinguishing cheap local functionals from those paying an exponential sample cost, a proof that window spectra are cheap while the hypotheses of dimension identification are not, and a certification lower bound of $(2^d - 1)\log\frac{1}{2\delta}$ for those hypotheses, of which the curvature floor is an instance. Second, a theory that survives misspecification (§5): channel invariance of every lower bound, a stability threshold defined for any data law, the resulting bias–variance decomposition of readout error, and a proof that stability does not imply correctness, since the covariance does not determine the latent dimension even when the latent geometry is ideal. Third, a mechanism for the bias (§6): on a gapless spectrum the target of a split-point rule is $\lfloor(\Lambda/\lambda_+)^{1/a}\rfloor$, set by signal-to-noise and decay exponent rather than by any intrinsic dimension. Fourth, a calibrated gate and a pre-registered empirical programme on three transformers (§7, §8) that measures the two terms of the decomposition separately, with a density-ladder control calibrated at the geometry it ran on and an equivalence test carrying an explicit power bound. Fifth, a re-analysis of a published study (§9) whose two dimension estimates disagree by an order of magnitude, locating most of that disagreement in the spectral readout and the rest in two recoverable errors.

Appendices C and S table every claim against its status and against the measurement bearing on it.

2 Related work

Layer-wise intrinsic dimension is estimated for convolutional networks (Ansuini et al., 2019) and transformers (Valeriani et al., 2023; Razzhigaev et al., 2024; Cheng et al., 2025), usually with nearest-neighbour-ratio estimators (Facco et al., 2017; Denti et al., 2022; Levina & Bickel, 2004). Closest to our concerns, (Schulte & Rügamer, 2026) prove that pointwise and Hausdorff dimension cannot increase across a Lipschitz network, so the rise-then-fall profiles these estimators report cannot track dimension. Our work is complementary: it addresses the *spectral* family rather than the

nearest-neighbour family, gives a *finite-sample* identification theory rather than a population monotonicity argument, and supplies a calibrated gate rather than a necessary condition; the two agree where they overlap (§N.9). Minimax rates for manifold, tangent-space and second-fundamental-form estimation (Genovese et al., 2012; Kim & Zhou, 2015; Aamari & Levrard, 2019) give curvature the rate $n^{-1/d}$, whose inversion is exponential in d, and Theorem 3.2 is tight against it. The deformed spectral law we use as a null is due to (Silverstein, 1995), with the spatial-sign covariance of (Frahm & Jaekel, 2010) for elliptical data; transformer representations are a canonical instance of the anisotropy these address (Ethayarajh, 2019; Sun et al., 2024; Cai et al., 2021). Absence claims use two one-sided tests (Schuirmann, 1987; Lakens et al., 2018) with Holm correction (Holm, 1979). Appendix D gives the extended discussion.

3 Statistical identification theory

3.1 Sample-size regimes

Regime means a range of window populations, never a model family, an architecture or a depth. Writing $n_r^{\dim} := 64\, C_0^2\, \rho_T^2\, \kappa_T^2 \Big(\mathbf{r}(\Sigma) + \log \frac{1}{\delta}\Big)$, $\quad n_{curv} := (2^d - 1) \log \frac{1}{2\delta}$, for the thresholds of Theorems 3.1 and 3.2, a window population is *sub-dimensional* ($n_r < n_r^{\dim}$), in the *identification gap* ($n_r^{\dim} \le n_r < n_{curv}$), or *super-curvature* ($n_r \ge n_{curv}$). The gap width is $\rho_{\mathrm{gap}} := n_{curv}/n_r^{\dim}$. The third regime is the complement of a lower bound: it asserts only that the obstruction has been cleared, not that curvature is estimable there. Definition H.6 refines this taxonomy with one defined for every data law.

The observation model. We write $\mathfrak{M}(d, \tau, L_3)$ for the class of compact C^3 submanifolds of $\mathbb{R}^D$ of dimension d and reach τ, observed through i.i.d. windows $W_r(x) = \mathcal{M} \cap \bar{B}(x, r)$ with $r \le \tau/4$, sub-Gaussian window data, density bounded between $f_{\min}$ and $f_{\max}$ with density-conditioning number $\ell_f := f_{\max}/f_{\min}$, and a uniform third-order modulus L_3; the full statement is Definition E.1 in Appendix E. The third-order hypothesis is not cosmetic: without it the second fundamental form at x is not determined by the observable geometry at any finite sample size (Proposition F.4).

3.2 The identification hierarchy

Theorem 3.1 (Dimension identification, spectral form)**.** *Let Σ be the population window covariance, with eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots$, and fix an index $d \ge 1$ with $\lambda_d > 0$. Write*

$$\kappa_T := \frac{\lambda_1}{\lambda_d}, \qquad \rho_T := \max_{1 \le j < d} \frac{\lambda_j}{\lambda_{j+1}} \quad (\rho_T := 1 \text{ if } d = 1),$$

for the condition number and the largest adjacent ratio of the leading block. Suppose the spectral scope condition

$$\frac{\lambda_d}{\lambda_{d+1}} \;\ge\; 8\, \rho_T \tag{1}$$

holds, with $\lambda_d/0 := +\infty$. Let the window data satisfy the concentration condition (C) of Theorem 5.3 with sub-Gaussian constant $C_0 \ge 1$. If

$$n_r \;\ge\; 64\, C_0^2\, \rho_T^2\, \kappa_T^2 \Big(\mathbf{r}(\Sigma) + \log \frac{1}{\delta}\Big), \tag{2}$$

then the max-gap rule at window rank $q = d + 1$ returns d with probability at least $1 - \delta$.

Under Definition E.1, the returned index is the intrinsic dimension provided the leading d eigendirections are the tangent directions, that is, provided every tangent eigenvalue exceeds every normal one. Condition equation 3 below is a sufficient geometric condition for this and for equation 1. In that case $\mathbf{r}(\Sigma) \le d + \operatorname{tr} \Sigma_\perp/\lambda_1$, where $\Sigma_\perp$ is the normal block.

The scope condition is a condition on the spectrum, and it cannot be certified geometrically. Condition equation 1 is where the geometry enters: the gap across the tangent–normal boundary must exceed every gap inside the tangent block by a fixed factor. It is a condition on the data, not on the sample size, and no increase in n_r relieves it. Stated geometrically, without ambient noise,

the cross-block ratio is $\Theta(\tau^2/((d+2)r^2))$ and the within-block ratios are at most ℓ_f, so a sufficient condition is

$$\frac{r}{\tau} \;<\; \frac{c}{\sqrt{\ell_f\,(d+2)}} \tag{3}$$

for an absolute constant c (Appendix F.1). On our cohort, with median $\ell_f \approx 115$ and $d \approx 12$ estimated by the ball-mass exponent, equation 3 requires windows that are small relative to the reach, $r/\tau \lesssim 0.025c$, far below the model's admissible $r \le \tau/4$.

Three qualifications fix what this licenses. *First*, whether r/τ is that small is a lower bound on the reach at scale r, and Theorem 4.2 shows that no window-based procedure certifies such a bound below $(2^d-1)\log\frac{1}{2\delta}$ samples. The condition under which the max-gap readout is a dimension measurement is therefore itself uncertifiable from the window. *Second*, equation 3 is a worst-case sufficient condition. The within-block bound ℓ_f is loose by the same margin as the old constant, so failing equation 3 does not show that the readout fails. The evidence that it fails on this data is empirical (§N.2). *Third*, under ambient noise the cross-block ratio becomes $(\lambda_d^{\mathrm{sig}}+\sigma^2)/\sigma^2$ and no longer depends on τ alone. The spectral form equation 1 remains the operative condition in that case, and it is the one our planted experiments test. The estimate of ℓ_f is biased downward (Appendix M), so the true requirement on r/τ is, if anything, stricter than stated; the numerical value should nonetheless be read as indicative, since it also depends on the estimate of d.

Theorem 3.2 (Curvature lower bound)**.** *Let $\mathcal{M} \in \mathfrak{M}(d,\tau,L_3)$ with $d \ge 2$ and $r \le \tau/4$. For every $\varepsilon < \frac{1}{2}$ and $\delta < \frac{1}{2}$, every estimator of the principal-curvature vector at x that attains relative error ε with confidence $1-\delta$ uniformly over the class requires*

$$n_r \;\ge\; \big(2^d-1\big)\log\frac{1}{2\delta}. \tag{4}$$

The bound is independent of D, τ, r and ε. At target absolute accuracy $\varepsilon_{\mathrm{abs}}$ it sharpens to $n_r \ge \big((L_3 r/(3c_0\varepsilon_{\mathrm{abs}}))^d-1\big)\log\frac{1}{2\delta}$.

Corollary 3.3 (Existence and width of the gap)**.** *The identification gap is non-empty whenever $(2^d-1)\log\frac{1}{2\delta} > 64\,C_0^2\rho_T^2\kappa_T^2(\mathbf{r}(\Sigma)+\log\frac{1}{\delta})$. Its width*

$$\rho_{\mathrm{gap}} = \frac{(2^d-1)\log\frac{1}{2\delta}}{64\,C_0^2\rho_T^2\kappa_T^2\,(\mathbf{r}(\Sigma)+\log\frac{1}{\delta})}$$

is exponential in d over a factor polynomial in d at fixed density conditioning, because $\mathbf{r}(\Sigma) \le d + \operatorname{tr}\Sigma_\perp/\lambda_1$ and $\rho_T, \kappa_T \le \ell_f$ under Definition E.1. It has no explicit dependence on the ambient dimension: D enters only through the effective rank.

The exponential form is what the manifold-estimation literature predicts: the minimax rate $n^{-(k-2)/d}$ of (Aamari & Levrard, 2019) is $n^{-1/d}$ at $k=3$, and inverting it to a fixed accuracy is exponential in d.

4 Output and Warrant Theory

Theorem 3.2 concerns one functional, the principal-curvature vector. The mechanism behind it is more general. What makes curvature expensive is not that it is second-order, but that it can be changed by modifying the window law on a set of exponentially small mass. We make this precise, show that the population spectrum does *not* have this property, and then show that the hypotheses under which a spectral readout is a dimension measurement *do*. This is the sense in which curvature is the right benchmark: a spectral readout never needs curvature to produce its output, but it needs a bound on curvature to be warranted.

Call two laws on the window *p-cap-coupled* if they agree conditionally off a set of mass at most p, and write $\omega_F(p)$ for the largest change in a functional F across such a pair—its *cap modulus* (Definition Q.1). Le Cam's two-point method in conditional-coupling form then prices every functional by its cap modulus: any estimator attaining accuracy below $\omega_F(p)$ needs $n_r \ge (\frac{1}{p}-1)\log\frac{1}{2\delta}$, which at $p = 2^{-d}$ reads $n_r \ge (2^d-1)\log\frac{1}{2\delta}$ (Lemma Q.2, Appendix Q).

Lemma F.5 with $M=2$ is the special case of a two-valued F, and Theorem 3.2 is the instance in which F is a principal curvature at x and the coupled pair is the bump construction of Appendix F.2

with $\ell = r/2$, for which $\omega_F(2^{-d}) \geq c_0 L_3 r/2$. The technique is Le Cam's two-point method in conditional-coupling form. The content lies in which functionals have a non-vanishing cap modulus, to which we now turn.

Proposition 4.1 (Window spectra are cap-stable)**.** *Let P, Q be p-cap-coupled laws on $W_r(x)$ with covariances Σ_P, Σ_Q. Then*

$$\|\Sigma_P - \Sigma_Q\| \;\leq\; 12\, p\, r^2,$$

so every eigenvalue of Σ_Q lies within $12pr^2$ of the corresponding eigenvalue of Σ_P. If in addition the eigengap of Σ_P at index d satisfies $\delta_d \geq c_\delta\, r^2/(\ell_f(d+2))$, as in step (1) of the proof of Theorem 3.1, then the leading d-dimensional eigenspaces U_P, U_Q satisfy

$$\|\sin\Theta(U_P, U_Q)\|_{\mathrm{F}} \;\leq\; \frac{24\sqrt{d}\,\ell_f\,(d+2)}{c_\delta}\; p.$$

At $p = 2^{-d}$ the perturbation is exponentially small. Any functional of the population covariance that is Lipschitz in operator norm therefore has cap modulus $O(2^{-d}r^2)$, and Lemma Q.2 constrains it only at exponentially fine accuracy. Theorem 3.1 escapes the obstruction as the quantities a spectral readout consumes are window averages which cannot see a cap. Because reach bounds curvature pointwise, $\|\mathrm{II}_y\| \leq 1/\tau$ for every $y \in M$, a single cap of high curvature is enough to violate them.

Theorem 4.2 (Certifying the warrant of a spectral readout)**.** *Fix $d \geq 2$, $L_3 > 0$, $r > 0$, $\delta \in (0, \frac{1}{2})$ and a reach threshold $\tau^\star > 0$, and write $\mathfrak{M}(d, L_3) := \bigcup_{\tau>0} \mathfrak{M}(d, \tau, L_3)$ $W_r(x)$ defined as in Definition E.1 for every $r > 0$. Call a test $T \in \{0, 1\}$ computed from the n_r observations in $W_r(x)$ a* reach certificate at level $\tau^\star$ *if*

(S) (soundness) $P_M(T = 1) \leq \delta$ *for every $M \in \mathfrak{M}(d, L_3)$ with $\tau(M) < \tau^\star$, and*

(P) (power) $P_{M_0}(T = 1) \geq 1 - \delta$ *for at least one $M_0 \in \mathfrak{M}(d, 2\tau^\star, L_3/2)$ carrying the uniform law on its window.*

$$\textit{If} \quad c_0\, L_3\, r\, \tau^\star \;>\; 6, \;\textit{where } c_0 \textit{ is the constant of Appendix F.2,} \tag{5}$$

then every reach certificate at level $\tau^\star$ requires $n_r \;\geq\; (2^d - 1)\, \log \frac{1}{2\delta}$.

The instance that matters is $\tau^\star = 4r$, where the certified property is the window condition $r \leq \tau/4$ of Definition E.1(ii) and equation 5 reads $c_0 L_3 r^2 > 3/2$. The bound is attained at the most favourable density conditioning: M_0 carries the uniform law, $\ell_f = 1$, so poor conditioning cannot be blamed.

It is worth being precise about what this does *not* cover. The scope condition equation 1 is a functional of the population covariance, and Proposition 4.1 shows such functionals have cap modulus $O(2^{-d}r^2)$: they are certifiable at polynomial cost, and we certify this one directly from the spectrum in Section 8. Setting $\tau^\star = r\sqrt{\ell_f(d+2)}$ would certify a *geometric sufficient condition* for it, which is a reach bound and hence expensive; but the spectral condition itself is cheap, and nothing here says otherwise. The hypothesis that carries the exponential cost is the reach condition alone.

Corollary 4.3 (Spectral readouts are not self-validating)**.** *Let $\hat{k}$ be any split-point readout computed from the window covariance. Call a window* reach-warranted *if it satisfies the window condition $r \leq \tau/4$ of Definition E.1(ii), which together with equation 1 and equation 2 is what Theorem 3.1 requires to guarantee $\hat{k} = d$. Assume equation 5 holds at $\tau^\star = 4r$. Then:*
(i) Any procedure that certifies warrant from the window alone, with soundness and power as in Theorem 4.2, requires $n_r \geq (2^d - 1)\log\frac{1}{2\delta} = n_r^{\mathrm{curv}}$. Throughout regime (R2)*, therefore, the tangent block is identified (Theorem 3.1), but the warrant that makes the readout a dimension measurement cannot be certified from the data the readout is computed on.*
(ii) The failure of warrant is invisible to the readout. The warranted and unwarranted windows M_0, M_1 constructed in the proof of Theorem 4.2 have population covariances within $12 \cdot 2^{-d}r^2$ in operator norm (Proposition 4.1), so a spectral readout is presented, up to an exponentially small perturbation, with the same spectrum in both cases.

The theorem's hypotheses therefore divide, and the division is the useful content. The spectral hypotheses equation 1 and equation 2 are functionals of the population covariance, which by Proposition 4.1 are cap-stable and so estimable from the window at polynomial cost; we estimate them in

Section 8 and they are the part that can be falsified *by data. The reach hypothesis is not of that kind, and part (i) prices it at n_{curv}.*

Consequently the statement "$\hat{k}$ is the local dimension" is not produced by the window. One of its premises can be checked and the other cannot, so the claim as a whole needs a source outside the window: an assumption, a geometric prior, or calibration against a null of the kind developed in Section 7.

5 Stability, Bias and Variance Theory

This section addresses what survives when the data leave the model. The results so far are stated for Definition E.1, which excludes ambient noise, anisotropic distortion and heavy tails. Transformer representations exhibit all three. The two halves of the sample-size hierarchy respond to misspecification in opposite ways. The lower bounds survive any observation channel, unchanged (Theorem 5.1). The dimension threshold does not survive, but it can be replaced by a threshold defined for every law (Theorem 5.3). Together they give a decomposition of readout error that holds without the manifold model (Corollary 5.4).

5.1 Lower bounds are invariant under observation channels

Let $X_1, \dots, X_{n_r}$ be latent window samples, i.i.d. from a law on $W_r(x) \subset M$. An *observation channel* is a Markov kernel K_n from $(\mathbb{R}^D)^{n_r}$ to an observation space $\mathcal{Y}$, fixed in advance and not depending on the hypothesis. The analyst sees only $Y \sim K_n(X_{1:n_r}, \cdot)$. Additive noise of any law, heteroscedastic noise depending on X_i, heavy-tailed corruption, anisotropic or nonlinear distortion, quantisation and joint post-processing of the sample are all channels in this sense.

Theorem 5.1 (Channel invariance)**.** *Let P, Q be p-cap-coupled latent laws with $|F(P)-F(Q)| > \eta$, and let K_n be an observation channel. Every estimator $\hat{F}$ computed from Y whose error probability $\Pr(|\hat{F} - F| \geq \eta/2)$ is at most $\delta < \frac{1}{2}$ under both $K_n P^{\otimes n_r}$ and $K_n Q^{\otimes n_r}$ requires $(1-p)^{n_r} \leq 2\delta$, and hence $n_r \geq (1/p - 1)\log\frac{1}{2\delta}$. Consequently Lemma Q.2, Theorem 3.2, Theorem 4.2 and part (i) of Corollary 4.3 hold, with unchanged constants, for observations produced by any observation channel.*

5.2 A stability threshold defined for every law

The dimension threshold of Theorem 3.1 is defined through the manifold model and has no meaning when the model fails. What can be defined for any law is the sample size at which a spectral readout recovers its own *population target*, whatever that target is.

Lemma 5.2 (Exact tolerance of the max-gap rule)**.** *Let Σ be positive semidefinite with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_q > 0$ at window rank $q \geq 3$, and suppose the population target $k^\star := \arg\max_{1 \leq j < q} \lambda_j/\lambda_{j+1}$ is unique. There is an exact tolerance $\varepsilon^\star(\Sigma, q)$, defined in Appendix R by a pair of conditions monotone in ε, such that for every symmetric $\hat{\Sigma}$ with $\|\hat{\Sigma} - \Sigma\| < \varepsilon^\star(\Sigma, q)$, the max-gap index of $\hat{\Sigma}$ at rank q equals $k^\star$. Moreover, writing $\mu := \log\frac{\lambda_{k^\star}}{\lambda_{k^\star+1}} - \max_{j \neq k^\star} \log\frac{\lambda_j}{\lambda_{j+1}} > 0$ for the log-ratio margin, $\varepsilon^\star(\Sigma, q) \geq \tanh(\mu/4)\,\lambda_q$.*

Theorem 5.3 (Law-free stability threshold)**.** *Let $Y_1, \dots, Y_{n_r}$ be i.i.d. from* any *law on $\mathbb{R}^D$ with covariance Σ. No manifold structure is assumed. Let $\hat{\Sigma}$ be a covariance estimator satisfying the concentration condition*

$$\Pr\big(\|\hat{\Sigma} - \Sigma\| \geq \varepsilon_{n_r}(\delta)\big) \leq \delta, \tag{C}$$

and let $\hat{k}$ be the max-gap index of $\hat{\Sigma}$ at rank q. Define $n_{stab}(\Sigma, q, \delta) := \min\{\, n : \varepsilon_n(\delta) < \varepsilon^\star(\Sigma, q)\,\}$. Then $\Pr(\hat{k} \neq k^\star) \leq \delta$ for every $n_r \geq n_{stab}$. Two instantiations matter and are given in Appendix R: for sub-Gaussian data and the sample covariance $n_{stab} \leq C_0^2(\|\Sigma\|/\varepsilon^\star)^2(\mathbf{r}(\Sigma) + \log\frac{1}{\delta})$, and for heavy-tailed data robust estimators satisfy (C) under finite fourth moments with n_{stab} defined identically.

5.3 A bias–variance decomposition of readout error

Corollary 5.4 (Bias and variance of a spectral readout)**.** *Let $\hat{k}$ and $k^\star$ be as in Theorem 5.3.*

(i) Variance (any law). *If* $n_r \geq n_{stab}$*, then* $\Pr(\hat{k} \neq k^\star) \leq \delta$.

(ii) Bias (latent-manifold laws). *Suppose further that* $Y = K(X)$ *for an observation channel* K *and latent samples* X *from some* $M \in \mathfrak{M}(d, L_3)$*, and write* $b := \mathbf{1}\{k^\star \neq d\}$*. For* $n_r \geq n_{stab}$*,*

$$\Pr(\hat{k} \neq d) \leq \delta \text{ if } b = 0, \qquad \Pr(\hat{k} \neq d) \geq 1 - \delta \text{ if } b = 1.$$

Beyond n_{stab} *the readout's error is fixed by* b*, and further samples do not change it.*

(iii) Warrant (latent-manifold laws). *The hypotheses under which Theorem 3.1 gives* $b = 0$ *cannot be certified from the observations when* $n_r < (2^d - 1)\log\frac{1}{2\delta}$*, under the nondegeneracy condition equation 5, whatever the channel (Theorems 4.2 and 5.1).*

5.4 Stability does not imply correctness

For our final set of stability results, we present an abbreviated version of the main theorem here, and refer the reader to Appendix H for the full theorem statement, along with a detailed discussion of the theorem's implications.

Theorem 5.5 (Stability does not imply correctness: **Abbreviated**). *For any covariance* Σ *of rank* R *and every* $d \leq R$*, there is a latent* d*-manifold satisfying every hypothesis of Theorem 3.1, observed through a Gaussian channel, with observed covariance* Σ*. Consequently the stability threshold is identical across all* d*, and above it every covariance-based readout is wrong with probability at least* $1 - \delta$ *for all but one* d*.*

6 Why spectral readouts fail: threshold crossing

Theorem 3.1 presupposes a spectrum with a tangent block separated from a normal block. We now make precise what a spectral rule returns when that presupposition fails, and illustrate it in Figure 1.

The profile statistic. For a window $W \in \mathbb{R}^{n_r \times D}$ with centred spectrum $\lambda_1 \geq \cdots \geq \lambda_m$, $m = \min(n_r - 1, D)$, and a null with limiting spectral distribution F, define for each split point k

$$\mathrm{KS}(k) \;=\; \sup_t \left| \widehat{F}_{\log \lambda_{k+1:m}}(t) - F_{\log}(t) \right|, \qquad \hat{k} \;=\; \arg\min_k \mathrm{KS}(k), \tag{6}$$

with the null's scale refitted by moment matching at every k, so that no null can win by rescaling. Here $\hat{k}$ is the number of leading directions that must be removed before the remainder matches the null. Under Definition E.1 with small ambient noise it should equal d, up to the normal directions curvature makes visible. The max-gap rule of Theorem 3.1 and the variance-fraction rule used by (Cai et al., 2021) are both split-point selectors of this kind; equation 6 makes the null explicit.

Proposition 6.1. [1] *Let the signal spectrum be a continuum* $\lambda_j = \Lambda j^{-a}$*,* $\Lambda > 0$*,* $a > 0$*, with no gap, observed against a noise bulk of upper edge* λ_+*. Suppose the retained tail matches the null exactly when every retained signal eigenvalue lies at or below* λ_+*, and that retaining an eigenvalue above* λ_+ *strictly increases* KS*. Then* $\hat{k} = \lfloor (\Lambda/\lambda_+)^{1/a} \rfloor$*,which depends on the signal-to-noise ratio* Λ/λ_+ *and on the decay exponent* a*, and not on any intrinsic dimension.*

7 A calibrated gate: profile depth

Proposition 6.1 explains why $\hat{k}$ is not a dimension. It leaves open whether a window contains detectable structure at all, and the argmin of equation 6 cannot answer that: it returns a split point whether or not the profile has a genuine minimum. We therefore measure the depth of the profile rather than the location of its minimum.

Definition 7.1 (Profile depth)**.** For a window and a null, the *profile depth* is $G := \frac{\mathrm{KS}(0)}{\min_k \mathrm{KS}(k)} \geq 1$.

$G = 1$ means that removing no direction is already optimal: the window's spectrum is indistinguishable from the null. $G \gg 1$ means that leaving the leading directions in makes the fit markedly worse, so there is structure the null cannot absorb. Unlike $\hat{k}$, G is a statement about the *existence* of a boundary, not its location.

[1] We prove this in Appendix L.

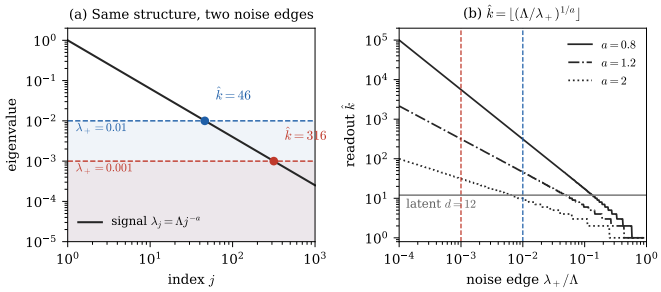


Figure 1: Threshold crossing (Proposition 6.1). **(a)** A gapless signal spectrum $\lambda_j = \Lambda j^{-a}$ ($\Lambda = 1$, $a = 1.2$) observed against two noise edges. The structure is identical in both cases; the split-point readout returns $\hat{k} = 46$ or $\hat{k} = 316$ depending only on where the edge sits. **(b)** The readout against the noise edge, for three decay exponents: it is set by signal-to-noise and decay alone, and by Corollary H.4 every d up to the rank of the spectrum is consistent with the same curve.

The null, and what the gate reads. An isotropic Marchenko–Pastur null rejects on anisotropic, heavy-tailed data whether or not a manifold is present. We score six nulls with the identical statistic equation 6, from that law to spatial-sign deformed laws whose population spectrum is estimated from held-out data; the last is our primary null, and Appendix J gives its construction, solver and calibration rows. On planted windows $G = 1.00$ on pure noise under every null and every noise family we test, and 1.99 to 116.97 on planted structure; under the cohort's own matched spectrum G is monotone in amplitude, so a failure to detect excludes a half-line rather than a band.

Remark 7.2 (What $G = 1$ licenses)**.** G is a detection statistic, and $G = 1$ is a failure to detect. Its meaning is fixed by its power, which we measure rather than assume: planted structure of dimension $d \in \{32, 64\}$ at amplitude $1.5\times$ the noise edge, under the cohort's own anisotropic spectrum, returns $G = 1.000$ (§N.4). The amplitude at which the gate begins to see is itself measured: under the matched spectrum $G = 1.000$ at $S = 1.5$ and $G = 1.663$ at $S = 1.75$, so the crossing of the equivalence margin is bracketed in $(1.5, 1.75)$ (§N.2). A window with G equivalent to 1 therefore licenses only the disjunction that local structure is *absent, or confined to dimension at least* 32 *within* $1.75\times$ *the noise edge*. We never report it as absence.

What the gate licenses, and what it does not. The profile depth G is computed from the window spectrum, so it is a covariance-based statistic, and Theorem 5.5 applies to it as much as to $\hat{k}$. G certifies that the window contains structure the null cannot absorb. It does not certify how many directions that structure has, and it does not repair the readout. Proposition K.1 in Appendix K states exactly what must be added for the profile readout to become a dimension measurement: isotropic Gaussian noise, fixed-rank latent structure, and idealized profile fit. Under these conditions $\hat{k} \to d$ if and only if every tangent direction clears the detection threshold and no curvature direction does. These conditions fail on our cohort, whose the noise is anisotropic, so (N) does not hold and the license does not transfer; extending it to a deformed null is open.

8 Empirical results

This section summarizes our empirical results; we present full details in Appendix N.

Cohort and protocol. We study BERT-base-uncased ($D = 768$, 12 layers, $N = 50{,}000$ tokens), GPT-2-large ($D = 1280$, 36 layers, $N = 46{,}875$) and GPT-2-XL ($D = 1600$, 48 layers, $N = 37{,}500$) on WikiText. The lookup layer is excluded throughout. Windows are k-nearest-neighbour patches at $n_r \in \{256, 1024\}$ about 64 anchors per layer. All thresholds, margins and grids are fixed before the runs and listed in Appendix T; deviations are reported as such.

Results. Six findings, ordered as Corollary 5.4 requires; Appendix N gives the designs and full tables.

The readout's target is not the structure (§N.2). With the content above the noise edge held fixed and the noise model varied, the readout spans $\hat{k} \in [3, 48]$ at one above-edge direction—two orders of magnitude at constant structure—and the spread contracts only as more directions clear the edge, which is what Proposition 6.1 predicts.

The decline with depth is not a bandwidth artefact (§N.3). A pre-registered thinning ladder of 400 rungs leaves the depth coefficient at $\beta = -1.406$ ($p = 0.001$), with sampling bandwidth accounting for 8.4% of the decline; the control is calibrated at the geometry it ran on, to a false-positive rate of 0.049 at power ≥ 0.998.

Absence is reported as equivalence, with measured power (§N.4). Three GPT-2-XL cells at 85–96% depth survive two one-sided tests at $\Delta = \log 1.35$ with Holm correction. Planted structure of dimension 32 and 64 returns $G = 1.000$ at $1.5\times$ the noise edge and 1.663 at $1.75\times$, so the claim is that structure is absent *or* confined to dimension at least 32 within $1.75\times$ the edge. The second disjunct is the content of the measurement, not a hedge.

The threshold of Theorem 3.1 is sound but loose (§N.5). Across 45 planted cells varying ρ_T, κ_T and $\mathbf{r}(\Sigma)$, the asserted bound holds in $45/45$ with a systematic $30\times$ margin; the measured scaling is $\rho_T^2(\mathbf{r}(\Sigma) + \log \frac{1}{\delta})$, so κ_T^2 is the conservative factor.

The hypotheses fail, and stability is not reached (§N.6–§N.7). On the cohort's own spectra the scope condition equation 1 fails at *every* candidate dimension in every window of $88/96$ cells at $n_r = 256$ and $45/96$ at $n_r = 1024$ (0.2% and 1.4% of windows). Disjoint-half subsampling, which isolates variance by construction, reaches the $(1 - \delta)^2$ stability level in $16/567$ cells and extrapolates to a window population three to seven times the one available, so Corollary 5.4does not apply; its premise $n_r \geq n_{stab}$ fails, the variance term cannot be treated as negligible, and the bias reading of part (ii) is unavailable.

No estimator is self-certifying (§N.9). Under a noise-floor criterion the two-nearest-neighbour estimator is inadmissible in all 96 cells. The ball-mass exponent survives both the planted worlds and the criterion, yet still reports a dimension near 15 at every cell where the gate is equivalent to noise, on both corpora; it must be read behind the gate, not in place of it.

9 Re-analysis of a published effective-dimension study

(Cai et al., 2021) report two dimension estimates for five architectures differing by more than an order of magnitude, and diagnose, for a different quantity, a per-cluster correction that would plausibly reconcile them. We pre-registered that prediction; it is disconfirmed in all five, and directionally, the correction moving the spectral estimate *away* from the density estimate. Two recoverable errors—clouds differing fortyfold in size, and local dimensions halved by an index returning squared distances—account for part of the published gap, but a residual of 9–$21\times$ survives both. Where the spectral estimate collapses to $d(\varepsilon) = 1$ the gate finds nothing either, while the density estimators report 12 to 25: what $d(\varepsilon)$ reads there is the global anisotropy, and nothing survives a matched null absorbing it. Appendix U gives the re-analysis in full.

10 Conclusion

A spectral dimension readout has an output and a warrant, and they cost different amounts. The output is stable once the window resolves the spectrum, at a sample size set by effective rank and margin, whatever the data law. The warrant rests on second-order hypotheses that no window certifies below $(2^d - 1) \log \frac{1}{2\delta}$ samples, under the manifold model or any hypothesis-independent corruption of it, because those hypotheses can be violated on a set of mass 2^{-d} that the spectrum cannot see. On transformer representations the readout's target tracks where a decaying spectrum crosses a noise edge, and moves by two orders of magnitude while the structure above it is held fixed. What survives calibration is narrower than a dimension: a gated, scale-resolved statement about whether structure is detectable at a stated bandwidth, against a stated null, with stated power.

Appendix V discusses where a local geometric certificate pays its $\exp(d)$, what should accompany a dimension reading, the relation to Lipschitz arguments, and the limitations.

REFERENCES

Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177–204, 2019.

Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E. Houle, Ken-ichi Kawarabayashi, and Michael Nett. Estimating local intrinsic dimensionality. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015.

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. Isotropy in the contextual embedding space: clusters and manifolds. In *International Conference on Learning Representations (ICLR)*, 2021.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Lei Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers. In *International Conference on Learning Representations (ICLR)*, 2025.

Francesco Denti, Diego Doimo, Alessandro Laio, and Antonietta Mira. The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12:20005, 2022.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, 2017.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3): 418–491, 1959.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Gabriel Frahm and Uwe Jaekel. A generalization of Tyler's M-estimators to the case of incomplete data. *Computational Statistics and Data Analysis*, 54(2):374–393, 2010.

Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax manifold estimation. *Journal of Machine Learning Research*, 13:1263–1291, 2012.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Arlene K. H. Kim and Harrison H. Zhou. Tight minimax rates for manifold estimation under Hausdorff loss. *The Annals of Statistics*, 43(6):2653–2675, 2015.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.

Daniël Lakens, Anne M. Scheel, and Peder M. Isager. Equivalence testing for psychological research: a tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269, 2018.

Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.

Stanislav Minsker. Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903, 2018.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry*, 39:419–441, 2008.

Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. The shape of learning: anisotropy and intrinsic dimensions in transformer-based models. In *Findings of the Association for Computational Linguistics: EACL*, 2024.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Rickmer Schulte and David Rügamer. Rethinking intrinsic dimension estimation in neural representations. *arXiv preprint arXiv:2604.20276*, 2026.

Jack W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339, 1995.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling (COLM)*, 2024.

William Timkey and Marten Van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, 2021.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

A Required Statements

A.1 AI Use Statement

In this article, we used AI to provide critiques of the paper and identify technical holes and literature gaps. Partial AI assistance was used for implementation and debugging.

A.2 Ethics Statement

We certify that we fully adhere to the ICLR Code of Ethics.

A.3 Reproducibility Statement

All efforts have been made to ensure reproducibility. Proofs for all theorems are presented in the Appendices. We provide complete experimental designs in the Appendix, together with all necessary parameters and hyperparameters required. Each hypothesis validated is stated clearly, and where appropriate we employ suitable statistical tests.

For novel models or algorithms, we attach anonymous downloadable source code as supplementary materials; for theoretical results, clear explanations of any assumptions and a complete proof of the claims has been included in the appendix; for any datasets used in the experiments, a complete description of the data processing steps is provided in the supplementary materials (as needed). For public data we provide the link/reference used to download the data.

All experiments were run on a single workstation with an NVIDIA GeForce RTX 4080 (16 GiB VRAM, driver 580.126.20), CUDA 12.0 and PyTorch with CUDA 12.x, under Linux. No multi-GPU or cluster resources were used. The heaviest single run is the density ladder of §N.3 at 400 rungs; the spectral audit of §N.6 and the stability measurements of §N.7 each complete in under an hour on this machine. Window extraction and the Gram reductions run on the GPU in `float32` with the spectra promoted to `float64`; the eigendecompositions are performed in `float64` throughout.

B Summary: What we claim, and what we do not

We claim that spectral dimension readouts cannot certify their own warrant from the window, under the manifold model and under any hypothesis-independent mis-specification of it; that on the transformer representations we study their target is set by the noise model rather than by the structure; and that a null-calibrated gate supplies the missing reference. We do not claim that these representations have, or lack, a latent manifold of any particular dimension: the gate detects structure at every shallow and intermediate layer we test, and where it finds none its power bounds what that licenses (§N.4). We do not claim that the bias of a readout is undetectable; only that the window cannot rule it out without an external reference. We do not claim that our lower bounds are achieved, and the concentration and minimax machinery is adapted from (Koltchinskii & Lounici, 2017; Aamari & Levrard, 2019). The cohort comprises three transformers with $D \leq 1600$.

C Summary of Coverage of Results

Table 1 lists each claim the paper makes or might be read as making, its status, what supports it, and what it does not license. Standing assumptions throughout are i.i.d. sampling within a window, or pooled i.i.d. sampling, and, for every latent statement, the existence of a latent manifold observed through a hypothesis-independent channel.

D Extended related work

Intrinsic dimension of representations. Layer-wise intrinsic dimension has been estimated for convolutional networks (Ansuini et al., 2019) and for transformers (Valeriani et al., 2023; Razzhigaev et al., 2024; Cheng et al., 2025), typically with nearest-neighbour-ratio estimators such

Claim	Status	Support	Does *not* license
Certifying the hypotheses of Thm 3.1 from a window needs $n_r \geq n_{curv}$, even under misspecification	Proved	Thms 4.2, 5.1; condition equation 5	That the readout is wrong; that b is undetectable
Spectral quantities are insensitive to cap-level changes	Proved	Prop. 4.1	That they equal any intrinsic quantity
Above n_{stab} the readout recovers its population target, for any law	Proved, given (C)	Thm 5.3	That the target is d; coverage of the profile rule
Readouts can be arbitrarily wrong at an identical stability certificate, with every latent hypothesis satisfied	Proved	Thm H.2, Rem. H.3	That real windows are such cases
The bias indicator b is not identified by any covariance-based statistic	Proved	Thm H.2(i)	That b is undetectable from the full law
The gate's matched null correctly models the channel	Assumption	Planted calibration (Sec. 7)	Validity where the held-out data carry structure
Real windows lie above n_{stab}	Empirical, measured: **no**	Disjoint-half agreement reaches $(1-\delta)^2$ in $16/567$ cells (§N.7); the plug-in is an upper bound and is not used for this row	A certified bound: plug-in values are estimates and are reported as diagnostics
Real windows lie below n_{curv}	Conditional	Holds if $d \geq 7$ ($n_r = 256$) or $d \geq 9$ ($n_r = 1024$)	Any claim without an external lower bound on d
In planted worlds, $k^\star$ tracks the noise model, not the structure ($b = 1$)	Empirical, ground truth known	Sec. 8	Transfer to real windows without a matched null
On real windows, the readout's target is not a function of the structure alone	Empirical, indirect	Dependence of $\hat{k}$ on the null and on n_r (Prop. 6.1)	A value for the true d, or $b = 1$ at a specific window
The failure on real windows is bias, not variance	Not claimed	The premise $n_r \geq n_{stab}$ is not met (§N.7), so Cor. 5.4(ii) does not apply	—
Real representations contain a latent manifold of dimension d	Not claimed	—	—

Table 1: Claims ledger. "Proved" means proved under the stated conditions. "Empirical" claims inherit the pre-registration and power statements of Section 8.

as TwoNN (Facco et al., 2017), its multiscale generalisation Gride (Denti et al., 2022), or the maximum-likelihood estimator of (Levina & Bickel, 2004). Closest to our concerns, (Schulte & Rügamer, 2026) prove that pointwise and Hausdorff dimension cannot increase across the layers of a Lipschitz network, so the rise-then-fall profiles these estimators report cannot track true dimension; they attribute the profiles to the growth of nearest-neighbour distances and relate them to spectral entropy. Our work is complementary in three respects: it addresses the *spectral* family of readouts rather than the nearest-neighbour family; it gives a *finite-sample* identification theory rather than a population monotonicity argument; and it supplies a calibrated gate rather than a necessary condition. The two lines of evidence agree where they overlap: we find nearest-neighbour-ratio estimators inadmissible on every cell we test, by an independent criterion (§N.9).

Manifold estimation. Minimax rates for manifold estimation (Genovese et al., 2012; Kim & Zhou, 2015) and for tangent spaces and second fundamental forms (Aamari & Levrard, 2019) establish that curvature estimation at order k has rate $n^{-(k-2)/d}$; for $k = 3$ this is $n^{-1/d}$, whose inversion to a target accuracy is exponential in d. Theorem 3.2 is consistent with, and tight against, that rate. Testing the manifold hypothesis itself is addressed by (Fefferman et al., 2016). The reach, which

governs our sampling precondition, is characterised by (Federer, 1959), and its relation to geodesic distortion by (Niyogi et al., 2008).

Spectra of anisotropic data. The Marchenko–Pastur law describes the spectrum of isotropic noise; for a general population spectrum the limiting law is given by the Silverstein equation (Silverstein, 1995). For elliptical and heavy-tailed data the sample covariance is not a suitable statistic, and spatial-sign covariance restores a tractable limit (Frahm & Jaekel, 2010). Transformer representations are a canonical instance: they occupy a narrow cone (Ethayarajh, 2019), carry a few massive activations (Sun et al., 2024), and form clusters and low-dimensional manifolds whose isotropy has been studied directly (Cai et al., 2021).

Equivalence testing. Our absence claims use two one-sided tests (Schuirmann, 1987; Lakens et al., 2018) with Holm correction (Holm, 1979), so that a failure to reject structure is never reported as its absence.

E THE OBSERVATION MODEL

Definition E.1 (Manifold model)**.** (i) $\mathcal{M} \subset \mathbb{R}^D$ is a compact C^3 submanifold of intrinsic dimension $d \geq 1$ with reach $\tau > 0$.

(ii) For $x \in \mathcal{M}$ and $r \leq \tau/4$, the *observation window* is the ambient ball section $W_r(x) := \mathcal{M} \cap \bar{B}(x, r)$, and n_r is the number of observations falling in it.

(iii) Observations are i.i.d. from a law μ on $W_r(x)$ with density f with respect to volume, $0 < f_{\min} \leq f \leq f_{\max}$; we write $\ell_f := f_{\max}/f_{\min}$ for the *density-conditioning number*. There is no additive ambient noise.

(iv) The centred window data are sub-Gaussian with constant C_0.

(v) *Uniform third-order modulus.* In any local graph parameterisation over $T_x\mathcal{M}$ the third derivatives of the defining map are bounded by L_3, uniformly over x and over the class.

We write $\mathfrak{M}(d, \tau, L_3)$ for the resulting class. Hypothesis (v) is not cosmetic. Without a uniform third-order bound the second fundamental form at x is not determined by the observable geometry at any finite sample size (Proposition F.4). Hypothesis (ii) uses the ambient window rather than the geodesic one because it is the object nearest-neighbour estimators form; the two agree to $O(r^3/\tau^2)$ for $r < \tau/4$.

F IDENTIFICATION BOUNDS

F.1 DIMENSION: THE UPPER THRESHOLD

Proof of Theorem 3.1. We show that $\varepsilon^\star(\Sigma, d{+}1) \geq \lambda_d/(8\rho_T)$ in Lemma 5.2; the claim then follows from Theorem 5.3(a), since $\lambda_1/\varepsilon^\star \leq 8\rho_T\kappa_T$. Put $\varepsilon = \eta\lambda_d$ with $\eta = 1/(8\rho_T) \leq 1/8$. For $j < d$ we have $\lambda_{j+1} \geq \lambda_d > \varepsilon$ and

$$\frac{\lambda_j + \varepsilon}{\lambda_{j+1} - \varepsilon} \leq \rho_T \frac{1+\eta}{1-\eta}, \qquad \frac{\lambda_d - \varepsilon}{\lambda_{d+1} + \varepsilon} \geq \frac{1-\eta}{\lambda_{d+1}/\lambda_d + \eta} \geq \frac{1-\eta}{2\eta},$$

using equation 1 in the last step. The tolerance condition therefore holds whenever $\rho_T(1+\eta)\cdot 2\eta < (1-\eta)^2$, that is $(1+\eta)/4 < (1-\eta)^2$, which is true for all $\eta \leq 1/8$ ($9/32 < 49/64$). By monotonicity in ε, every smaller ε is also admissible. The effective-rank bound uses $\operatorname{tr}\Sigma_T \leq d\lambda_1$. □

Remark F.1. The spectral form of the bounds is stated in the quantities that actually enter the argument, and $\kappa_T, \rho_T, \mathbf{r}(\Sigma)$ can be measured on planted windows. It makes a testable prediction: the measured threshold should scale with $\rho_T^2\kappa_T^2\,\mathbf{r}(\Sigma)$. The planted density ladder of §N.5 can test this by regressing the measured threshold on κ_T instead of ℓ_f. The bound remains a sufficient condition, and we expect its constant to be conservative.

The effective-rank concentration bound is free of the ambient dimension, whereas the earlier form carried a factor $\log(D/\delta)$, inherited from a matrix-Bernstein argument. In the spectral form, D enters only through $\mathbf{r}(\Sigma)$.

Remark F.2 (A scope condition the ratio rule needs)**.** The max-gap rule selects the largest *ratio*, so index d wins only if the cross-block ratio exceeds every within-block ratio. Within the tangent block adjacent eigenvalues differ by at most ℓ_f; across the blocks the ratio is $\lambda_d/\lambda_{d+1} = \Theta\big(\tau^2/((d+2)r^2)\big)$. The rule is therefore valid only when

$$\frac{\tau^2}{(d+2)\,r^2} \;>\; \ell_f, \qquad \text{i.e.} \qquad \frac{r}{\tau} \;<\; \frac{1}{\sqrt{\ell_f\,(d+2)}}, \tag{7}$$

which at the model's $r \le \tau/4$ requires $\ell_f(d+2) < 16$. This is a restriction on the data, not on the sample size, and no increase in n_r relieves it. On our cohort, with median $\ell_f \approx 115$ at $d \approx 12$ (§N.5), the left-hand side of equation 7 is violated by roughly two orders of magnitude. Since equation 7 is a worst-case sufficient condition, this does not by itself show that the readout fails; the evidence for that is empirical (§N.2).

Remark F.3 (The rank condition)**.** The argument perturbs from a population spectrum, which presupposes $n_r > q$ for the window rank q at which the spectrum is read. When $n_r \le q$ the window is rank-deficient and the hypothesis is absent: the rule still returns a number, but no result here covers it. In a noiseless window the normal block has rank at most $\binom{d+1}{2}$, so reading the ratio over the full spectrum locates the numerical-rank cliff rather than d; q must lie inside the populated spectrum. Cells failing either condition must be reported as such.

F.2 Curvature: the lower bound

Proposition F.4 (Without L_3, curvature is not estimable at any sample size)**.** *Drop hypothesis (v) of Definition E.1 and let $\mathfrak{M}_2$ be the resulting C^2 class. For every n_r, every $\eta > 0$ and every estimator $\hat\kappa$ there exist $\mathcal{M}_0, \mathcal{M}_1 \in \mathfrak{M}_2$ with identical reach and window laws up to total variation η, whose principal curvatures at x differ by an arbitrarily large factor.*

Proof. Fix $\mathcal{M}_0$ flat near x, let χ be a smooth cutoff supported in the ball of radius ℓ, and put $\mathcal{M}_1 = \operatorname{graph}\big(\tfrac{1}{2}\langle u, Su\rangle\chi(u/\ell)\big)$. Within the C^2 class $\|S\|$ may be taken arbitrarily large while ℓ shrinks, since the constraint is not uniform as $\ell \downarrow 0$. The two window laws differ only on the cylinder $\{|u| \le \ell\}$, of mass $(\ell/r)^d$; choosing ℓ with $n_r(\ell/r)^d \le \eta$ makes the experiments η-indistinguishable while $\|S\|$ is unbounded. Le Cam's two-point inequality gives the claim. □

The construction. For an admissible shape operator $S \in \operatorname{Sym}(d)$, a fixed smooth cutoff χ and a scale $\ell \in (0, r)$, define $\varphi_S(u) = \frac{1}{2}\langle u, Su\rangle\chi(u/\ell)$ and let Q_S be the law of a sample drawn uniformly from the ambient window of $\operatorname{graph}(\varphi_S)$. Admissibility in $\mathfrak{M}(d, \tau, L_3)$ requires $\|S\| \le c_0 L_3 \ell$. Two such hypotheses share support and have *identical conditional law off* the cylinder Cyl_ℓ, differing only inside it, where the mass is $p = (\ell/r)^d$.

Lemma F.5 (Conditional indistinguishability)**.** *Let $\mathbb{P}_j = Q_j^{\otimes n_r}$, $j = 1, \dots, M$, and suppose there is a common event E with $Q_j(E^c) \le p$ and $Q_j(\cdot \mid E) = \nu$ for every j. Then any selector $\hat J$ with $\max_j \mathbb{P}_j(\hat J \ne j) \le \delta$ satisfies $n_r \log \frac{1}{1-p} \ge \log \frac{1-1/M}{\delta}$.*

Proof. Let A be the event that all n_r samples fall in E; then $\mathbb{P}_j(A) \ge (1-p)^{n_r}$, and conditionally on A the sample has law $\nu^{\otimes n_r}$, independent of j. Hence $\beta := \mathbb{P}_j(\hat J = j \mid A)$ is the same for all j, and $\sum_j \mathbb{P}_j(\hat J = j \mid A) \le 1$ gives $\beta \le 1/M$. Therefore $\mathbb{P}_j(\hat J \ne j) \ge (1-p)^{n_r}(1 - 1/M)$; imposing $\le \delta$ and taking logarithms gives the claim. □

Two features deserve emphasis. The confidence δ enters *inside* the information budget rather than through an appended union bound; and the packing size M enters only through $\log(1-1/M)$, which saturates already at $M = 2$, so enlarging the hypothesis family buys nothing. The second point is why no repair of a Fano-style format is possible here: in the noiseless model a hypothesis is identified by a single sample landing in the bump, so $I(J; X^{1:n_r}) \le n_r p \log M$ and the $\log M$ cancels against the packing entropy. No factor proportional to the number of shape-operator components, in particular no $\Theta(d^2)$, can appear.

Proof of Theorem 3.2. Take $M = 2$ with $S' = 0$ and $\|S\| = c_0 L_3 \ell$, so any estimator with relative error $\varepsilon < \frac{1}{2}$ induces a selector with error probability at most δ. Applying Lemma F.5 with $E = \mathrm{Cyl}_\ell^c$ and $p = (\ell/r)^d$ gives $n_r \log \frac{1}{1-p} \geq \log \frac{1}{2\delta}$. Choose $\ell = r/2$, so $p = 2^{-d}$ and $\log \frac{1}{1-p} \leq \frac{p}{1-p} = \frac{1}{2^d-1}$; substituting yields $n_r \geq (2^d - 1) \log \frac{1}{2\delta}$. For the absolute statement, forcing a separation $3\varepsilon_{\mathrm{abs}}$ requires $c_0 L_3 \ell \geq 3\varepsilon_{\mathrm{abs}}$, giving $p^{-1} = (L_3 r/(3c_0\varepsilon_{\mathrm{abs}}))^d$. □

Corollary F.6 (Minimax risk for second-order functionals)**.** *Let $\phi : \mathrm{Sym}(d) \to \mathbb{R}$ be Lipschitz with $\|\nabla\phi\| \geq c_\phi > 0$ on the admissible range, and $f(\mathcal{M}) = \phi(\mathrm{I}\!\mathrm{I}_x)$. Then for every n_r and every estimator $\hat{f}$,*

$$\inf_{\hat{f}} \sup_{\mathcal{M} \in \mathfrak{M}(d,\tau,L_3)} \mathbb{P}_{\mathcal{M}}\Big(|\hat{f} - f(\mathcal{M})| \;\geq\; \tfrac{c_\phi}{2}\, c_0 L_3 \ell\Big) \;\geq\; \tfrac{1}{2}\,(1-p)^{n_r}, \qquad p = (\ell/r)^d,$$

so the risk is bounded below by a constant throughout the identification gap. The minimax risk, not a variance bound, is the meaningful object: a constant estimator has zero variance.

G Proofs for Section 4: Output and Warrant Theory

G.1 Proof of Lemma Q.2

Proof. Since $\eta < \omega_F(p)$ there are p-cap-coupled $P, Q \in \mathcal{P}$ with $|F(P) - F(Q)| > \eta$. Let E be the coupling set, ν the common conditional law, and A the event that all n_r observations fall in E. Then $P^{\otimes n_r}(A) \geq (1-p)^{n_r}$ and $Q^{\otimes n_r}(A) \geq (1-p)^{n_r}$, and conditionally on A the sample has law $\nu^{\otimes n_r}$ under both. The events $\{|\hat{F} - F(P)| < \eta/2\}$ and $\{|\hat{F} - F(Q)| < \eta/2\}$ are disjoint, so their conditional probabilities a, b given A, which are the same under P and Q, satisfy $a + b \leq 1$. Hence

$$\max\Big\{P^{\otimes n_r}\big(|\hat{F}-F(P)| \geq \tfrac{\eta}{2}\big),\; Q^{\otimes n_r}\big(|\hat{F}-F(Q)| \geq \tfrac{\eta}{2}\big)\Big\} \;\geq\; (1-p)^{n_r} \max\{1-a,\, 1-b\} \;\geq\; \tfrac{1}{2}(1-p)^{n_r}.$$

Imposing that the left-hand side is at most δ gives $(1-p)^{n_r} \leq 2\delta$, which is the first inequality. The second follows from $\log \frac{1}{1-p} \leq \frac{p}{1-p}$. □

Lemma F.5 with $M = 2$ is the special case of a two-valued F, and Theorem 3.2 is the instance in which F is a principal curvature at x and the coupled pair is the bump construction of Appendix F.2 with $\ell = r/2$, for which $\omega_F(2^{-d}) \geq c_0 L_3 r/2$. The technique is Le Cam's two-point method in conditional-coupling form. The content lies in which functionals have a non-vanishing cap modulus, to which we now turn.

G.2 Proof of Proposition 4.1

Proof. Write $P(B) - Q(B) = (P(E) - Q(E))\,\nu(B) + P(B \cap E^c) - Q(B \cap E^c)$. Each of the two differences is at most p in absolute value, so $d_{\mathrm{TV}}(P, Q) := \sup_B |P(B) - Q(B)| \leq 2p$, and $|\mathbb{E}_P g - \mathbb{E}_Q g| \leq 2 d_{\mathrm{TV}}(P, Q)\|g\|_\infty \leq 4p\,\|g\|_\infty$ for bounded g. For a unit vector v, applying this to $g(y) = \langle v, y - x\rangle^2 \leq r^2$ bounds the difference of second moments about x by $4pr^2$ in operator norm. Applying it to $g(y) = \langle v, y - x\rangle$, $|g| \leq r$, bounds the difference of the means m_P, m_Q (taken about x) by $4pr$. Since $\Sigma = \mathbb{E}[(y-x)(y-x)^\top] - mm^\top$ and $\|m_P m_P^\top - m_Q m_Q^\top\| \leq (\|m_P\| + \|m_Q\|)\|m_P - m_Q\| \leq 2r \cdot 4pr$, the total is at most $12pr^2$. The eigenvalue statement is Weyl's inequality. The subspace statement is the Davis–Kahan variant of Yu et al. (2015), $\|\sin\Theta\|_{\mathrm{F}} \leq 2\sqrt{d}\,\|\Sigma_P - \Sigma_Q\|/\delta_d$. □

At $p = 2^{-d}$ the perturbation is exponentially small. Any functional of the population covariance that is Lipschitz in operator norm therefore has cap modulus $O(2^{-d} r^2)$, and Lemma Q.2 constrains it only at exponentially fine accuracy. The tangent space, the tangent eigenvalues and the population location of a well-separated gap are all of this kind. This is why Theorem 3.1 escapes the obstruction: the quantities a spectral readout consumes are window averages, and window averages cannot see a cap.

The hypotheses of Theorem 3.1 are not window averages. The requirement $r \leq \tau/4$ of Definition E.1(ii), and the scope condition equation 1 on which the max-gap rule depends, are lower

bounds on the reach. Because reach bounds curvature pointwise, $\|\mathrm{II}_y\| \le 1/\tau$ for every $y \in M$, a single cap of high curvature is enough to violate them.

G.3 Proof of Theorem 4.2

Proof. Let M_0 be as in (P). In a graph parameterisation of M_0 over T_xM_0, add the bump φ_S of Appendix F.2 with $\ell = r/2$ in a fixed normal direction, taking $\|S\| = c_0(L_3/2)\ell = c_0L_3r/4$ so that the third-order modulus of the sum is at most L_3. Call the result $M_1 \in \mathfrak{M}(d, L_3)$. The gradient of the perturbation vanishes at x, so the second fundamental form of M_1 at x is $\mathrm{II}_x^{(0)} + S$, and

$$\|\mathrm{II}_x^{(1)}\| \;\ge\; \|S\| - \|\mathrm{II}_x^{(0)}\| \;\ge\; \frac{c_0L_3r}{4} - \frac{1}{2\tau^\star} \;>\; \frac{1}{\tau^\star},$$

the last step by equation 5. Hence $\tau(M_1) \le 1/\|\mathrm{II}_x^{(1)}\| < \tau^\star$, and (S) applies to M_1.

The two manifolds coincide outside the cylinder Cyl_ℓ. Their uniform window laws therefore have the same conditional law on $E = W_r(x) \setminus \mathrm{Cyl}_\ell$, and the cap mass is at most $p = (\ell/r)^d = 2^{-d}$ as in Appendix F.2, so the two laws are 2^{-d}-cap-coupled. Take $F = \mathbf{1}\{\tau(M) \ge \tau^\star\}$, which equals 1 at M_0 and 0 at M_1, and $\hat{F} = T$. Conditions (S) and (P) bound both error probabilities by δ, and Lemma Q.2, applied to the family $\{P_{M_0}, P_{M_1}\}$ with $\eta = \frac{1}{2}$, gives the claim. □

The instance that matters is $\tau^\star = 4r$, where the certified property is the window condition $r \le \tau/4$ of Definition E.1(ii) and equation 5 reads $c_0L_3r^2 > 3/2$. The bound is attained at the most favourable density conditioning: M_0 carries the uniform law, $\ell_f = 1$, so poor conditioning cannot be blamed.

It is worth being precise about what this does *not* cover. The scope condition equation 1 is a functional of the population covariance, and Proposition 4.1 shows such functionals have cap modulus $O(2^{-d}r^2)$: they are certifiable at polynomial cost, and we certify this one directly from the spectrum in Section 8. Setting $\tau^\star = r\sqrt{\ell_f(d+2)}$ would certify a *geometric sufficient condition* for it, which is a reach bound and hence expensive; but the spectral condition itself is cheap, and nothing here says otherwise. The hypothesis that carries the exponential cost is the reach condition alone.

Remark G.1 (Limitations to Theorem 4.2)**.** Part (ii) cuts both ways. Because the spectrum barely moves, the readout on M_1 may well still return d. Theorem 4.2 is a statement about *warrant*, not about *error*: it says the guarantee of Theorem 3.1 cannot be verified from the window, not that the readout is wrong whenever the guarantee fails. A readout that is robust to warrant violations on sets of small mass would need a separate theorem under hypotheses on window averages. Such hypotheses have vanishing cap modulus and are identifiable at polynomial cost, but they are not the hypotheses of Theorem 3.1. The failures we observe on transformer representations have a different source, Proposition 6.1. There the spectrum is gapless and the split tracks a noise edge. That departure is global, has non-vanishing effect on the population spectrum, and is detectable at polynomial cost by a matched null, which is why the gate of Section 7 can detect it. The two results are complementary: Proposition 6.1 explains why the readout fails on this data, and Theorem 4.2 explains why the readout could not have certified its own success even on data where it works.

Remark G.2 (The nondegeneracy condition)**.** Condition equation 5 says that the class is rich enough at window scale for curvature within a single cap to cross the threshold $1/\tau^\star$. When $c_0L_3r\tau^\star$ is small, the third-order bound already caps how far curvature can vary within the window, and the reach condition reduces to a statement about the global reach of the manifold. That statement is itself an external input, so the conclusion of Corollary 4.3 persists in a different form: warrant is then assumed rather than measured.

Remark G.3 (Relation to reach estimation and to Section N.8)**.** Minimax rates for estimating the reach are given by Aamari & Levrard (2019); Theorem 4.2 concerns the weaker one-sided problem of certifying a lower bound on reach from a single window, and its value is the constant $(2^d - 1)\log\frac{1}{2\delta}$, shared with Theorem 3.2. The detour certificate of Proposition I.1 is one such certificate, and the empirical finding of Section N.8, that no operating point both resolves geodesics and retains power at the cohort's dimension, is the behaviour Theorem 4.2 predicts for every certificate, not only geodesic ones. It therefore removes the caveat in Section 8 that a non-geodesic reach certificate is not excluded: any window-only certificate is excluded below n_r^{curv}.

H Statement and Proof of Theorem 5.5

The full statement of Theorem 5.5 is presented as Theorem H.2 below. We first present some introductory material before moving to the main Theorem.

Corollary 5.4 splits readout error into variance and bias, but does not show that the bias term can be large once variance is controlled. We now show that it can be as large as the rank of the spectrum allows. The latent geometry can satisfy every hypothesis of Theorem 3.1, and the stability certificate can be identical across every candidate dimension. The argument applies to any readout that sees the data only through their covariance, which includes every spectral rule considered in this paper.

Definition H.1 (Covariance-based readout)**.** A *covariance-based readout* is a map $\mathcal{R}$ from covariance matrices to indices, applied as $\hat{k} = \mathcal{R}(\hat{\Sigma})$. Its *population target* is $k^\star = \mathcal{R}(\Sigma)$. The max-gap rule at window rank q, the variance-fraction rule, and the population version of the profile rule equation 6 for a fixed null are all covariance-based.

Theorem H.2 (Stability does not imply correctness)**.** *Let $\Sigma \succeq 0$ on $\mathbb{R}^D$ have rank $R \geq 2$, eigenvalues $\lambda_1 \geq \cdots \geq \lambda_R > 0$ and eigenvectors $u_1, \ldots, u_R$, and fix $\theta \in (0,1)$. For each $d \in \{1, \ldots, R\}$, let $M_d \in \mathfrak{M}(d, \tau, L_3)$ be flat on the window $W_r(x)$, with $\tau \geq 4r$, and let X be uniform on $W_r(x)$. Then there is an observation channel $Y = A_d X + \xi_d$, with A_d linear and injective on $T_x M_d$ and $\xi_d \sim \mathcal{N}(0, \Sigma_{\xi,d})$ independent of X, such that:*

(a) $\mathrm{Cov}(Y) = \Sigma$ *for every d;*

(b) the latent tangent directions map onto $u_1, \ldots, u_d$, and in each of them the latent signal accounts for a fraction θ of the observed variance;

(c) the observed laws are sub-Gaussian with an absolute constant C_0, the same for every d.

Consequently, for every covariance-based readout $\mathcal{R}$:

(i) the population target $k^\star = \mathcal{R}(\Sigma)$ is the same for every $d \in \{1, \ldots, R\}$, so it equals d for at most one of them, and

$$\max_{1 \leq d \leq R} |k^\star - d| \;\geq\; \left\lceil \tfrac{R-1}{2} \right\rceil;$$

(ii) for the max-gap rule at window rank $q \leq R$, the stability threshold $n_r^{\mathrm{stab}}(\Sigma, q, \delta)$ of Theorem 5.3 is the same number for every d. Hence at every $n_r \geq n_r^{\mathrm{stab}}$ and every $d < q$ with $d \neq k^\star$, $\mathbb{P}_d(\hat{k} \neq d) \;\geq\; 1 - \delta$.

Proof. Place the flat window at the origin, with $T_x M_d = \mathrm{span}(e_1, \ldots, e_d)$. Then X has mean zero and covariance $c_d P_T$ with $c_d = r^2/(d+2)$. Set

$$A_d = \sum_{i \leq d} \sqrt{\theta \lambda_i / c_d}\, u_i e_i^\top, \qquad \Sigma_{\xi,d} = \sum_{i \leq d} (1-\theta)\lambda_i\, u_i u_i^\top + \sum_{i > d} \lambda_i\, u_i u_i^\top \;\succeq\; 0.$$

Then $\mathrm{Cov}(A_d X) = \theta \sum_{i \leq d} \lambda_i u_i u_i^\top$, and independence gives (a). Part (b) is immediate from the construction. For (c), the one-dimensional marginals of a uniform law on a ball and of a Gaussian are sub-Gaussian with absolute constants relative to their standard deviations, and the property is preserved for sums of independent terms. Part (i) holds because $\mathcal{R}(\Sigma)$ does not depend on d, and a single integer is at distance at least $\lceil (R-1)/2 \rceil$ from one of 1 or R. For (ii), n_r^{stab} depends only on Σ, q, δ and C_0, all independent of d. Theorem 5.3 then gives $\mathbb{P}_d(\hat{k} = k^\star) \geq 1 - \delta$, and $k^\star \neq d$. □

Remark H.3 (A well-conditioned instance)**.** Take $\Sigma = c\, I_k \oplus (c/8)\, I_{q-k}$ with $1 \leq k < q$. Then $k^\star = k$, the margin is $\log 8$, and $n_r^{\mathrm{stab}} \leq 281\, C_0^2 \left(q + \log \frac{1}{\delta} \right)$, linear in q and independent of d. The readout returns k for every $d \in \{1, \ldots, q-1\}$ at this sample size, wrong by up to $\max(k-1,\, q-1-k)$.

Corollary H.4 (Threshold crossing with a latent manifold)**.** *Let Σ have the gapless signal spectrum $\lambda_j = \Lambda j^{-a}$ of Proposition 6.1 on its leading R directions. For every $d \leq R$ there is a latent flat d-disk and a channel, as in Theorem H.2, with observed covariance Σ. Under the hypotheses of Proposition 6.1, the profile readout returns $\lfloor (\Lambda/\lambda_+)^{1/a} \rfloor$ for every such d.*

Proof. Apply Theorem H.2 to Σ and Proposition 6.1 to its spectrum, which does not depend on d. □

This corollary is a *population* statement. Stability of the profile rule is not covered by Theorem 5.3, whose target is a functional of Σ alone, because the profile rule's target moves with n_r through the noise edge. The proved "stable but wrong" statement is part (ii) of Theorem H.2, for the max-gap rule.

Remark H.5 (What Theorem H.2 does and does not show)**.** The construction is elementary; its value lies in what it rules out. *First*, every latent M_d is flat, uniformly sampled ($\ell_f = 1$) and satisfies $r \le \tau/4$. So even an oracle that certified the latent hypotheses of Theorem 3.1 would not rescue the readout: the bias can come entirely from the channel. The two obstructions are independent. Theorem 4.2 concerns the latent geometry and costs $\exp(d)$ samples. Theorem H.2 concerns the channel and is not resolved by any number of samples, for any covariance-based statistic. *Second*, the channels involved are not exotic. Where $k^\star > d$, the channel adds high-variance directions outside the tangent space, which is the "rogue dimension" and massive-activation phenomenon documented in transformers (Timkey & Van Schijndel, 2021; Sun et al., 2024). Where $k^\star < d$, it attenuates tangent directions into the floor. *Third*, the bias indicator $b = \mathbf{1}\{k^\star \neq d\}$ is not a function of Σ, so no covariance-based procedure can detect it. We do *not* claim that b is undetectable from the full observed law. The laws in the family differ in their higher moments. What the theorem shows is that ruling out bias requires information beyond the spectrum: a model of the channel. The matched null of Section 7 is exactly such a model, stated explicitly and calibrated. Its adequacy is an assumption, recorded as such in Table 1.

Definition H.6 (Law-free regimes)**.** For a window with covariance Σ and latent dimension d: (R1′) $n_r < n_{stab}$: the readout does not reliably recover its own target; (R2′) $n_{stab} \le n_r < n_{curv}$: the readout recovers its target, but two independent obstructions stand between that target and d—the warrant cannot be certified from the window (Theorem 4.2), and by Theorem H.2 certifying it would not suffice, since the target need not equal d even when every latent hypothesis holds; (R3′) $n_r \ge n_{curv}$: the obstruction of Theorem 4.2 is cleared.

As before, this does not assert that certification is possible. Under Definition E.1 with $q = d + 1$, (R1′) and (R2′) refine (R1) and (R2): the boundary n_{stab} is a functional of Σ rather than of (d, D), and the two agree in order of magnitude on the planted spectra of §N.5.

The two boundaries of the law-free taxonomy have different epistemic status. n_{stab} depends on Σ, which is observable, so a window's position relative to it can be estimated from data. n_{curv} depends on the latent d, which is not observed, so placing a window below it requires an external lower bound on d. At $\delta = 0.05$, $n_r < n_{curv}$ holds for $d \ge 7$ when $n_r = 256$ and for $d \ge 9$ when $n_r = 1024$.

I THE DETOUR CERTIFICATE

Proposition I.1 (Detour certificate)**.** *Let $\mathcal{M}$ have reach τ and let $x, y \in \mathcal{M}$ with chord $s = \|x - y\| \le \tau/2$ lie in a window of radius r. If*

$$\frac{d_{\mathcal{M}}(x, y)}{s} > \frac{2}{1 + \sqrt{1 - s/2r}}, \tag{8}$$

then $r > \tau/4$.

The proof inverts the geodesic-versus-chord bound of (Niyogi et al., 2008). The certificate is one-sided: firing proves the precondition is violated; not firing proves nothing.

Proof of Proposition I.1. By (Niyogi et al., 2008, Prop. 6.3), for x, y on a manifold of reach τ with chord $s = \|x - y\| \le \tau/2$,

$$d_{\mathcal{M}}(x, y) \le \tau\Big(1 - \sqrt{1 - 2s/\tau}\Big).$$

Suppose for contradiction that $r \le \tau/4$, i.e. $\tau \ge 4r$. The right-hand side is increasing in τ after division by s only through the combination s/τ, and is decreasing in τ; substituting the extremal

admissible value $\tau = 4r$ therefore gives the weakest bound consistent with the hypothesis,

$$\frac{d_{\mathcal{M}}(x,y)}{s} \;\le\; \frac{4r}{s}\Big(1-\sqrt{1-s/2r}\Big) \;=\; \frac{2}{1+\sqrt{1-s/2r}},$$

the last step by rationalising $1-\sqrt{1-u} = u/(1+\sqrt{1-u})$ with $u = s/2r$. This contradicts equation 8, so $r > \tau/4$. □

Implementability. The certificate is stated for the manifold geodesic. Replacing it by a k-NN graph geodesic d_G preserves soundness, since a path through samples shadowing the geodesic has chord-sum at most the geodesic length (each chord is at most its arc) and d_G is the minimum over paths, so $d_G \le d_{\mathcal{M}}$—*provided* the graph contains such a path. That proviso is not a technicality: it fails exactly when the sample spacing $r\,n_r^{-1/d}$ is comparable to r, which is the regime of interest. We therefore test it per band by measuring the detour of a curvature-free null at the same (n_r, d, k) and refusing any band where the null itself clears the threshold. Two degeneracies must also be excluded: a band whose null detour is identically 1 consists of single-hop pairs, for which $d_G = s$ by construction and the test has no power; and a band whose null detour exceeds the threshold indicates a graph too sparse to resolve geodesics. On our cohort no (n_r, d, k) avoids both (§N.8).

J THE INSTRUMENT AND ITS CALIBRATION

Every number in §8 is a comparison against a planted reference of known truth at matched (D, n_r). The statistic is equation 6, with the null's scale refitted by moment matching at every split point k, so that a null cannot win by rescaling.

The null hierarchy. Six nulls are scored with the identical statistic:

Null	Spectrum from	Law
`iso`	sample covariance	isotropic Marchenko–Pastur
`aniso-glob`	sample covariance	deformed, H from held-out global covariance
`aniso-loc`	sample covariance	deformed, H from pooled held-out local covariances
`aniso-nbr`	sample covariance	deformed, H from held-out points near the anchor
`sign-iso`	spatial-sign covariance	isotropic Marchenko–Pastur
`sign-aniso`	spatial-sign covariance	deformed

The deformed law is obtained from the Silverstein equation

$$\underline{m}(z) = \Big(\gamma \int \frac{t}{1+t\,\underline{m}(z)}\, dH(t) - z\Big)^{-1}, \qquad f_F(x) = \frac{1}{\pi\gamma}\,\Im\,\underline{m}(x+i0^+), \tag{9}$$

solved as a damped fixed point. Two implementation facts are load-bearing and were established by failure: a linear integration grid cannot resolve a density whose support spans decades, which is what a power-law population spectrum produces; and the imaginary offset must scale with the evaluation point rather than with the spectrum's maximum. With a fixed offset the solver mismatches a planted power-law spectrum at $\mathrm{KS} = 0.30$; with an offset proportional to the evaluation point it matches to 0.008, improving as D grows.

Tyler's M-estimator requires $n > D$ and its limiting law is proved only for aspect ratio below 1, which excludes most of our cells. The spatial-sign covariance is used at all aspect ratios and its behaviour above 1 is established here by planted calibration rather than by citation. Calibration rows are not decoration: a null that accepts planted signal has no power, and its acceptance of a real cell would mean nothing.

K A POSITIVE RESULT

Proposition K.1 (Conditional license)**.** *Let the window observations be $Y_i = S_i + \xi_i$, where:*

(N) Correct null. *The ξ_i are i.i.d. $\mathcal{N}(0, \sigma^2 I_D)$, independent of the S_i, and the null used in equation 6 is the Marchenko–Pastur law with the same σ^2 and aspect ratio.*

(L) Latent structure. *The S_i are i.i.d., bounded and centred, with covariance Σ_S of fixed rank m, where $d \le m \le d + \binom{d+1}{2}$. Its eigenvalues are $\theta_1 \ge \cdots \ge \theta_d$ on the tangent directions and $\vartheta_1 \ge \cdots \ge \vartheta_{m-d}$ on the curvature-induced normal directions.*

(P) Idealised profile fit. *The profile rule removes exactly the sample eigenvalues lying outside the null's bulk, as in the hypotheses of Proposition 6.1.*

Let $n_r, D \to \infty$ with $D/n_r \to \gamma \in (0, \infty)$, with m, the θ_i and the ϑ_i fixed, and none of them equal to the detection threshold $\sigma^2\sqrt{\gamma}$. Write $d_+ := \#\{i \le d : \theta_i > \sigma^2\sqrt{\gamma}\}$ and $m_+ := \#\{i : \vartheta_i > \sigma^2\sqrt{\gamma}\}$. Then, almost surely, $\hat{k} \to d_+ + m_+$, and $G > 1$ eventually if and only if $d_+ + m_+ \ge 1$. In particular $\hat{k} \to d$ if and only if: (1) every tangent direction is detectable: $d_+ = d$, and (2) no curvature direction is detectable: $m_+ = 0$.

L Why Spectral Readouts Fail

Proof of Proposition 6.1.

Proof. Under the hypotheses KS is minimised by removing exactly the eigenvalues exceeding λ_+, so $\hat{k} = |\{j : \Lambda j^{-a} > \lambda_+\}|$. The inequality $\Lambda j^{-a} > \lambda_+$ holds iff $j < (\Lambda/\lambda_+)^{1/a}$, and counting integers gives the claim. □

The hypotheses idealise the null fit, but the conclusion is robust to how they are relaxed: whatever a split-point rule does near the edge, it locates the edge, and on a gapless spectrum the edge's position is set by (Λ, λ_+, a). The readout should therefore move with n_r, exceed any density-based dimension, and move when the null changes while the structure above the edge does not; all three are confirmed in §8.

M Estimating the density-conditioning number

The constant C_1 depends on $\ell_f = f_{\max}/f_{\min}$ within a window. Local density obeys $f \sim k/(nV_d r_k^d)$, so at fixed k the dimension-dependent factors cancel in a ratio and $\ell_f = (r_k^{\max}/r_k^{\min})^d$, with r_k the distance from a point of the window to its k-th neighbour within the window. We use the 5th and 95th percentiles rather than the extremes, which makes the estimate a lower bound on the true ratio—the conservative direction for a placement that needs ℓ_f small.

One correction is essential. The k-th neighbour radius fluctuates as $1/\sqrt{k}$ from Poisson sampling alone, so at $k = 16$ a perfectly uniform window has a percentile ratio near 1.6; raising that to the power d yields $\ell_f \approx 15$ where the truth is 1. We therefore divide the observed ratio by the ratio a uniform sample of the same (n_r, d, k) produces before exponentiating. Validated against planted worlds, the corrected estimator returns $\ell_f = 1.00$ on a window uniform in a ball—at ambient noise floors 0, 0.54 and 1.0, and at $D = 1600$—and 2.38 on a window with a radial density gradient.

N Empirical results

N.1 Cohort and protocol

We study BERT-base-uncased ($D = 768$, 12 layers, $N = 50{,}000$ tokens), GPT-2-large ($D = 1280$, 36 layers, $N = 46{,}875$) and GPT-2-XL ($D = 1600$, 48 layers, $N = 37{,}500$) on WikiText. The lookup layer is excluded throughout. Windows are k-nearest-neighbour patches at $n_r \in \{256, 1024\}$ about 64 anchors per layer. All thresholds, margins and grids are fixed before the runs and listed in Appendix T; deviations are reported as such.

N.2 The max-gap readout is not identified

Proposition 6.1 predicts that the readout tracks the position of the noise edge. We test this by holding the structure above the edge fixed and varying the noise model. Planting a decaying tangent

spectrum at amplitude S times the edge, under four noise families (isotropic, globally anisotropic, elliptical, and the cohort's own matched spectrum), the number of planted directions clearing the edge is exactly 1 at $S = 1.5$ and exactly 2 at $S = 3$, invariant across planted $d \in \{4, 8, 16, 32, 64\}$ and across all four families. Over the same cells the readout spans

$$\hat{k} \in [0, 64] \text{ at } S = 1.5, \qquad \hat{k} \in [1, 87] \text{ at } S = 3.$$

At constant above-edge content the readout moves by nearly two orders of magnitude; it is reporting the noise model, not the structure. Profile depth on the same cells varies by a factor of 24–38 across families at identical planted structure, which is why the gate must be calibrated against a null matched to the data rather than against an isotropic reference.

Extending the amplitude grid to $S \in \{2, 2.5, 3, 4, 9, 20, 50, 100\}$ leaves this conclusion unchanged and sharpens it. The number of planted directions clearing the edge is $1, 2, 2, 3, 7, 15, 33, 33$ respectively, identical for planted $d = 32$ and $d = 64$; over the same cells the readout spans $\hat{k} \in [3, 48]$ at $S = 2$, a $16\times$ spread at single-direction above-edge content, contracting monotonically to $2.4\times$ at $S \geq 20$. The readout therefore tracks the position of the edge and not the planted dimension, and the spread it exhibits is largest exactly where the edge is least constrained.

The same grid measures the gate's sensitivity boundary, which Remark 7.2 requires. Under the cohort's matched spectrum the profile depth is

S	1.5	2	2.5	3	50	100
directions above edge	1	1	2	2	33	33
G, planted $d = 32$	1.00	2.00	2.49	2.98	36.90	49.56
G, planted $d = 64$	1.00	1.99	2.32	2.98	21.09	36.09

and $G = 1.000$ on the matched noise plant at every amplitude. The progression is monotone throughout, so a failure to detect excludes a half-line of amplitudes rather than a band.

Monotonicity is a property of the matched and isotropic families, not of the instrument in general. Under synthetic globally anisotropic planting G falls as amplitude rises—from 19.87 at $S = 4$ to 3.89 at $S = 20$ for $d = 32$, with non-overlapping 10–90 percentile intervals—because the minimum KS over split points rises $7.7\times$ across that range while $\mathrm{KS}(0)$ rises only $1.5\times$: planted energy leaks into the tail and no split point removes it. Power must therefore be calibrated against the null family in use, and this is a second reason, independent of Proposition 6.1, why an isotropic reference is inadequate.

The boundary is measured rather than interpolated. At $S = 1.75$, with one planted direction above the edge, the matched family returns $G = 1.663$ for planted $d = 32$ and $d = 64$ alike, with 10–90 percentile intervals $[1.514, 1.996]$ and $[1.499, 1.994]$; both lie wholly above the equivalence margin, while $G = 1.000$ at $S = 1.5$. The crossing is therefore bracketed in $(1.5, 1.75)$, and a window equivalent to noise excludes planted structure of dimension 32 or 64 at $1.75\times$ the edge and above. We quote the measured endpoint rather than the log-linear interpolant, which falls near $S = 1.65$.

N.3 Depth, and the sampling-density control

Profile depth declines with relative depth in GPT-2. Because n_r and the window radius jointly set the bandwidth at which the instrument resolves structure, a decline could be an artefact of sampling density rather than a property of the representation. We test this with a pre-registered ladder.

At each cell we form six rungs by thinning the window population by factors $c \in \{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$, giving 400 rungs across 40 layer clusters, and fit a collapse of $\log G$ onto a bandwidth coordinate θ. The unit of clustering is the physical layer, not the (layer, n_r) pair: rungs from the same layer at different n_r are not independent. Writing the collapse as $\log G = \alpha + \beta\theta + \text{depth}$, the depth coefficient survives the control on both collapse variables:

collapse variable	β	90% CI	p	heterogeneity
θ_{global}	-1.406	$[-1.931, -0.842]$	0.001	0.418
$\hat{\theta}$ (per window)	-3.993	$[-4.961, -3.011]$	0.001	0.422

The uncorrected depth slope is -1.534, so sampling bandwidth accounts for 8.4% of the decline and the remaining 91.6% has no density explanation.

The control is calibrated on the geometry it ran on rather than on a nominal model. Resampling the measured per-anchor dispersion (median 0.19 in $\log G$ over 398 cells) under the null of no depth effect gives a false-positive rate of 0.049 (95% CI $[0.040, 0.059]$, 2000 seeds) against a nominal 0.05, and power at least 0.998. The rate is flat across a $13\times$ range of dispersion (0.06–0.80) and across within-layer correlation from 0 to 0.9. An earlier version clustered on (layer, n_r) pairs and inflated to 0.125 at correlation 0.3; the cluster unit is the correction.

N.4 An equivalence claim, and its power

At the deepest layers of GPT-2-XL the gate finds no structure. Reporting this requires an equivalence test, not a failed detection: we use two one-sided tests at a pre-registered margin $\Delta = \log 1.35$ with Holm correction. Three cells at $n_r = 1024$ survive—layers 41, 43 and 46, at 85%, 90% and 96% relative depth, with $G = 1.000$, 1.000 and 1.080—against $G = 9.14$ at layer 4 of the same model.

What this licenses is fixed by Remark 7.2, and is measured rather than assumed. Planting structure of dimension 32 and 64 at amplitude $1.5\times$ the noise edge under the cohort's own anisotropic spectrum returns $G = 1.000$, numerically identical to pure noise, while the same structure at $1.75\times$ the edge returns $G = 1.663$; the amplitude grid of §N.2 therefore brackets the gate's sensitivity boundary in $(1.5, 1.75)$. The equivalence claim is: at these layers local structure is absent, or confined to dimension at least 32 within $1.75\times$ the noise edge. Stating the second disjunct is not a hedge; it is the content of the measurement.

N.5 Calibrating the spectral threshold

Theorem 3.1 states its threshold as a functional of the window spectrum, so the calibration varies the quantities it contains rather than d and D. We plant windows whose leading block is known by construction: $\lambda_j \propto j^{-a}$ for $j \le d$ above a floor, with the decay a setting ρ_T, the block size d setting κ_T, and the ratio λ_d/λ_{d+1} set directly. Across 45 cells ($d \in \{2, 4, 8, 12, 16\}$, $a \in \{0, \frac{1}{2}, 1\}$, gap $\in \{16, 64, 256\}$) we locate the smallest n_r at which the max-gap rule at rank $q = d + 1$ returns d with probability at least $1 - \delta$.

The theorem is sound and loose. The realised constant $n_r^\star/[\rho_T^2 \kappa_T^2(\mathbf{r}(\Sigma) + \log \frac{1}{\delta})]$ satisfies the asserted bound $64C_0^2$ in $45/45$ cells, with maximum 2.13: the guarantee holds with about $30\times$ to spare. But the looseness is systematic rather than uniform, and the prediction of a sharper scaling is not borne out:

predictor for $n_r^\star$	slope	R^2	spread of $n_r^\star/\text{pred}$
$\rho_T^2\kappa_T^2(\mathbf{r} + \log \frac{1}{\delta})$ (the theorem's)	0.290	0.376	249×
$\rho_T\kappa_T(\mathbf{r} + \log \frac{1}{\delta})$	0.831	0.735	8.0×
$\rho_T^2(\mathbf{r} + \log \frac{1}{\delta})$	1.043	0.677	9.3×

A tight bound would give slope 1 and a narrow band. The theorem's own predictor gives neither: the realised constant falls from 1.47 at $a = 0$ to 0.058 at $a = 1$, a $25\times$ drift driven by the squares. Empirically the threshold scales as $\rho_T^2(\mathbf{r}(\Sigma) + \log \frac{1}{\delta})$, with slope 1.04; κ_T^2 is the conservative factor. We report this because a sufficient condition loose by a systematic $30\times$ should not be read as the operative dependence. The earlier, geometric route through the density-conditioning number is looser still—grading the planted density over a $348\times$ range in ℓ_f moved the threshold by $4.0\times$ against the ℓ_f^3 prediction of 4.2×10^7—which is why the present form, stated on observable spectral quantities, replaces it. The ℓ_f estimator of Appendix M plays no further role.

N.6 Do real windows meet the theorem's hypotheses?

The calibration fixes what the theorem requires; the spectra of the cohort's own windows say whether it is met. For every window we compute $\mathbf{r}(\hat{\Sigma})$, and at every candidate $d \le 64$ the pair (ρ_T, κ_T), the scope condition equation 1 and the threshold equation 2. Nothing here needs the latent geometry: a window either exhibits a spectrum meeting the hypotheses at some d or it does not.

	$\mathbf{r}(\hat{\Sigma})$	$\hat{k}$	n_{stab}	$n_r \geq$ plug-in bound	any d meeting equation 1
$n_r = 256$	7.3	1	5.7×10^3	0.0%	0.2%
$n_r = 1024$	9.3	1	7.7×10^3	3.9%	1.4%

(medians over 96 cells per population; n_{stab} evaluated at the rank the readout is read at, $q = \hat{k} + 1$). Three readings. The effective rank is 7–9, two orders below the ambient 1600, so the threshold is not large for want of a small $\mathbf{r}(\Sigma)$. The plug-in bound on n_{stab} evaluates to 7 to 22 times n_r, rising from 2.0×10^4 to 3.9×10^5 across the rank grid $q \in \{8, 16, 32, 64\}$. Theorem 5.3(a) bounds n_{stab} from above, so we record this as a diagnostic and settle the question directly in §N.7. And the scope condition fails at *every* candidate dimension in every window of $88/96$ cells at $n_r = 256$ and $45/96$ at $n_r = 1024$, and for only 0.2% and 1.4% of windows overall, so for nearly all there is no d at which Theorem 3.1 could apply, whatever the sample size.

N.7 Is the readout stable at the populations we have?

Corollary 5.4 separates bias from variance only above n_{stab}, so whether the cohort clears it decides what may be asserted. Two routes are available and only one of them settles the question.

The plug-in route is inconclusive by construction. Condition (C) makes $\varepsilon_n(\delta)$ an *upper* bound on $\|\hat{\Sigma} - \Sigma\|$, so the expression $C_0^2(\|\Sigma\|/\varepsilon^\star)^2(\mathbf{r}(\Sigma) + \log\frac{1}{\delta})$ of Theorem 5.3(a) is an upper bound on n_{stab}, not an estimate of it. On the cohort it evaluates to 5.7×10^3 at $n_r = 256$ and 7.7×10^3 at $n_r = 1024$, and at ranks $q \in \{8, \ldots, 64\}$ from 2.0×10^4 to 3.9×10^5. Exceeding n_r is consistent with the window lying on either side of the true threshold, and we draw no conclusion from it.

The direct route needs a design that isolates variance. Comparing a subsample against the window that contains it will not do: the two share their points, and the agreement so obtained is inflated by the overlap. Two *disjoint* draws of size m from the same window are independent given the window, and share its population covariance by construction, so disagreement between them is sampling variance alone. If the readout is at its target with probability at least $1 - \delta$ at sample size m, two such draws agree with probability at least $(1 - \delta)^2 = 0.9025$.

	rank	m=32	64	128	256	384	512
$n_r = 256$	$q = 3$	64.8%	69.6%	74.7%	—	—	—
	$q = 8$	52.8%	57.5%	64.3%	—	—	—
$n_r = 1024$	$q = 3$	64.8%	68.8%	73.0%	78.7%	81.8%	83.8%
	$q = 8$	52.4%	57.6%	62.9%	69.1%	72.7%	74.8%

No configuration reaches the target. The median tops out at 83.8%, at the largest population, the most favourable rank and the largest m the design admits; across all cells and ranks only $16/567$ clear it at that point (2.8%). Agreement is close to log-linear in m, and extrapolating the two $n_r = 1024$ fits to 0.9025 gives $m \approx 1.4 \times 10^3$ at $q = 3$ and $m \approx 3.7 \times 10^3$ at $q = 8$. Since disjoint halves require $2m \leq n_r$, these correspond to window populations of roughly 3 to 7×10^3, against the 1024 we have: the readout would need between three and seven times the window population to reach the level Theorem 5.3 guarantees. The extrapolation is quoted as an order of magnitude, not a measurement; what is measured is that the target is not reached anywhere in the grid.

The cohort therefore lies below n_{stab}, so Corollary 5.4 does not apply: its premise $n_r \geq n_{stab}$ fails, the variance term cannot be treated as negligible, and the bias reading of part (ii) is unavailable. What the data do establish is the other half—that the scope condition equation 1 fails at every candidate dimension in every window of $88/96$ cells at $n_r = 256$—and that half needs no stability assumption at all.

N.8 The reach precondition cannot be certified

Every result above is conditional on the windows lying in $\mathfrak{M}(d, \tau, L_3)$ at $r \leq \tau/4$. That hypothesis is itself testable in principle, by a certificate needing neither tangent spaces nor curvature.

In practice $d_{\mathcal{M}}$ is unavailable and is replaced by a k-NN graph geodesic d_G. This is sound—$d_G \leq d_{\mathcal{M}}$, because a path shadowing the geodesic has chord-sum at most the geodesic length and

d_G is the minimum over paths—but only when the graph *contains* such a path. We test that condition directly, by measuring the detour a curvature-free window of the same (n_r, d, k) already produces. At the cohort's local dimension it exceeds the threshold of equation 8: at $d = 8$, $n_r = 1024$, $k = 20$ the median null detour is 1.41 against a threshold of 1.1716, and raising k drives the window into single-hop degeneracy, where the detour is 1 by construction and the certificate has no power. Across $(n_r, d) \in \{256, 1024\} \times \{4, 8, 12, 16, 21\}$ at $k = 20$ no operating point both resolves geodesics and retains power.

Sweeping the graph degree over $k \in \{6, 10, 16, 20, 24, 40, 64, 100\}$ at n_r up to 4097 shows that the obstruction has a sharp boundary in d rather than being uniform. Two bands out of 273 are usable, both at $d = 4$: at $(n_r, k) = (256, 64)$ the curvature-free null attains a maximum detour of 1.1476 and at $(1024, 100)$ of 1.1455, in each case below the threshold 1.1716 but by only two per cent. At $d \geq 8$ no k suffices. Writing each cell's best achievable ratio of null detour to threshold over all k,

d	4	8	12	16	21
$n_r = 256$	0.980	1.424	1.330	1.529	1.615
$n_r = 1024$	0.978	1.276	1.368	1.476	1.484
$n_r = 4097$	1.018	1.349	1.456	1.456	1.484

the step from $d = 4$ to $d = 8$ costs a factor of 1.3 to 1.5 at every window size and no choice of k recovers it. Two features of the table are worth separating. Enlarging the window does not help: at $d = 4$ the best ratio is worse at $n_r = 4097$ than at $n_r = 1024$, because more points at fixed k make the graph relatively sparser, so paths acquire more hops and more artefact; the certificate has an optimal k/n_r, not an optimal k. And the two usable margins rest on a single null draw each, so they should be read as locating the boundary between $d = 4$ and $d = 8$ rather than as establishing a usable operating point at $d = 4$.

This is a negative result with a structural cause. Resolving a geodesic needs sample spacing well below r, and in d dimensions that spacing is $r\, n_r^{-1/d}$—$0.42r$ at $n_r = 1025$, $d = 8$. The reach is uncertifiable by geodesics for the same exponential-in-d reason that curvature is unidentifiable in Theorem 3.2, even though Proposition I.1 is tangent-free and curvature-free by construction.

The cohort's local dimension, measured by the ball-mass exponent, is approximately 12. The reach precondition is therefore uncertifiable by this route on the data the paper studies, and by Theorem 4.2 it is uncertifiable by any window-only route below n_{curv}.

N.9 Which estimators are admissible

A density-based estimator is interpretable only if the ambient noise ball is small against the nearest-neighbour spacing it reads. We therefore report a noise-floor ratio r_{noise}/r_1 per cell, with $r_{\text{noise}} = \sigma_{\text{bulk}}\sqrt{D}$ estimated from the spectral bulk and r_1 the median nearest-neighbour distance. Across 96 layers the ratio has median 0.536 (range 0.46–1.36): the noise ball is about half the spacing. Under this criterion the two-nearest-neighbour estimator is admissible in $0/96$ cells, and the ratio of its estimate to a ball-mass exponent correlates with the noise floor at $r = +0.95$—the signature of an estimator reading the noise rather than the geometry. On planted worlds it returns 48.00 on a noisy two-dimensional Swiss roll. The ball-mass exponent survives both the planted worlds and the noise-floor criterion, and is the best behaved of the estimators we test.

It is nonetheless not self-certifying, and the gate shows why. Of the 80 undiluted cells in the cohort, five are equivalent to noise at the margin $\Delta = \log 1.35$—GPT-2-large layer 34 and GPT-2-XL layers 41, 43 and 46, which are the cells carrying the equivalence claim of §N.4—and at every one of them the ball-mass exponent reports between 13.2 and 15.1, with median 15.0, against a noise floor of 0.47 to 0.52. The same pattern appears on the second corpus: at the cells of §U where the gate is equivalent to noise, the ball-mass exponent reports 12.5 to 16.9. Two corpora, two samplings and the same number.

This is not a demonstration that the estimator is wrong at those cells. The gate's power bound (Remark 7.2) admits structure of dimension at least 32 above $1.75\times$ the noise edge, and the ball-mass exponent reports roughly 15, which is below that; the two statistics are not in contradiction but jointly uninformative. That is the substantive point. An estimator that returns a number at every

window, including those where a calibrated gate and an admissibility check both find nothing, cannot supply its own warrant. The recommendation is accordingly to report the ball-mass exponent behind the gate and the noise-floor check, not in place of them.

O RE-ANALYSIS PROTOCOL

Layer conventions. Two conventions had to be established empirically. For the four transformer models the released extractor indexes the hidden-state tuple after removing the embedding lookup, so a file's layer index equals the published one; this was confirmed by re-extracting from the models and matching stored clouds to candidate hidden states on two independent fingerprints, which agreed at margins of 5.9–$22\times$. ELMo is extracted through a separate tool that retains its character-convolution layer at index zero, so its published layers are offset by one; the offset is fixed by reproduction (-9.9% and -4.1% at offset one, against -38.9% and $+11.7\%$ at offset zero). The character-convolution layer is context-independent—every occurrence of a type receives the same vector—and is excluded throughout, as it is by the original authors. It is also the one cell in the study where the noise-floor criterion admits the two-nearest-neighbour estimator, and it does so spuriously: with duplicate vectors the median nearest-neighbour distance is zero and the ratio is undefined.

Reproduction. Under these conventions $d(\varepsilon)$ reproduces Table 1 to within 20% in 80%, 83%, 100%, 60% and 43% of cells. The shortfall for GPT-2 is uninformative rather than a failure: five of its twelve published values equal 1, where relative error carries no information. The shortfall for GPT includes one cell, layer 2, that we cannot reproduce (146 against a published 65, where adjacent layers read 141 and 76). DistilBERT's published Table 3 value is 31% above our token-level estimate and we do not account for it. ELMo contributes only two valid layers, too few to estimate a depth trend.

P THE GATE'S NULL HIERARCHY AND CALIBRATION

The null. Transformer representations are anisotropic and heavy-tailed: an isotropic null rejects on such data whether or not a manifold is present. We score six nulls with the identical statistic equation 6—an isotropic law, three deformed laws obtained from the Silverstein equation with population spectrum estimated from held-out data at global, pooled-local and per-anchor resolution, and spatial-sign variants (Silverstein, 1995; Frahm & Jaekel, 2010). The primary null is the spatial-sign deformed law; the others form a hierarchy of sensitivity checks. Appendix J gives the equation, the solver and the calibration rows.

Calibration. On planted windows at matched (D, n_r), $G = 1.00$ on pure noise under every null and every noise family we test, and G ranges from 1.99 to 116.97 on planted structure over the amplitude grid of §N.2. Under the cohort's own matched spectrum G is monotone in amplitude, rising from 2.00 at twice the noise edge to 49.56 at a hundred times it; the isotropic family behaves likewise. Under synthetic globally anisotropic planting it is not monotone, because planted energy leaks into the spectral tail and no split point removes it, so power must be calibrated per noise family rather than assumed to increase with signal. The gate is therefore sensitive to structure and insensitive to anisotropy, heavy tails and scale heterogeneity, which are the confounds that defeat an isotropic null.

Q THE CAP MODULUS

Definition Q.1 (Cap-coupled pair; cap modulus)**.** Let $\mathcal{P}$ be a family of laws on the window $W_r(x)$ and let $p \in (0, 1)$. Two laws $P, Q \in \mathcal{P}$ are *p-cap-coupled* if there is a measurable set E with $P(E^c) \le p$, $Q(E^c) \le p$ and $P(\cdot \mid E) = Q(\cdot \mid E)$. For a functional $F : \mathcal{P} \to \mathbb{R}$, the *cap modulus* of F is $\omega_F(p) := \sup\{\, |F(P) - F(Q)| \;:\; P, Q \in \mathcal{P} \text{ are } p\text{-cap-coupled} \,\}$.

Lemma Q.2 (Cap-modulus lower bound)**.** *Let $0 < \eta < \omega_F(p)$ and $0 < \delta < \frac{1}{2}$. Every estimator $\hat{F}$ computed from n_r i.i.d. observations that satisfies $\sup_{P \in \mathcal{P}} P^{\otimes n_r}\big(|\hat{F} - F(P)| \ge \eta/2\big) \le \delta$*

requires

$$n_r \;\geq\; \frac{\log \frac{1}{2\delta}}{\log \frac{1}{1-p}} \;\geq\; \Big(\frac{1}{p} - 1\Big) \log \frac{1}{2\delta}.$$

At $p = 2^{-d}$ *this reads* $n_r \geq (2^d - 1) \log \frac{1}{2\delta}$.

R Instantiations of the concentration condition

The exact tolerance of Lemma 5.2 is the supremum of $\varepsilon \geq 0$ such that, for every $j \in \{1, \ldots, q-1\} \setminus \{k^\star\}$,

$$\lambda_{j+1} > \varepsilon \qquad \text{and} \qquad \frac{\lambda_j + \varepsilon}{\lambda_{j+1} - \varepsilon} \;<\; \frac{\lambda_{k^\star} - \varepsilon}{\lambda_{k^\star+1} + \varepsilon}.$$

Both conditions are monotone decreasing in ε, so the admissible set is an interval and bisection locates $\varepsilon^\star$ to machine precision. Theorem 5.3 is then instantiated as follows.

(a) *Sub-Gaussian data, sample covariance.* With sub-Gaussian constant $C_0 \geq 1$ and effective rank $\mathbf{r}(\Sigma) = \operatorname{tr}\Sigma / \|\Sigma\|$, condition (C) holds with $\varepsilon_n(\delta) = C_0 \|\Sigma\| \max\{\sqrt{(\mathbf{r}(\Sigma) + \log \frac{1}{\delta})/n},\, (\mathbf{r}(\Sigma) + \log \frac{1}{\delta})/n\}$ (Koltchinskii & Lounici, 2017), so

$$n_{stab} \;\leq\; C_0^2 \Big(\frac{\|\Sigma\|}{\varepsilon^\star(\Sigma, q)}\Big)^2 \Big(\mathbf{r}(\Sigma) + \log \frac{1}{\delta}\Big).$$

(b) *Heavy-tailed data.* Robust covariance estimators satisfy (C) under finite fourth moments, with $\varepsilon_n(\delta)$ governed by a matrix-variance proxy in place of $\|\Sigma\|^2 \mathbf{r}(\Sigma)$; see (Minsker, 2018). The threshold n_{stab} is defined identically.

S Correspondence between claims and evidence

Table 2 pairs each formal statement with the measurement that bears on it, the outcome, and the section reporting it. It is the operational companion to the claims ledger of Appendix C: the ledger records what each claim does and does not license, this table records what was run and what came back. One row is marked *inconclusive by construction*: the available statistic bounds the quantity in the unhelpful direction.

T Pre-registered parameters

Quantity	Value
Window populations n_r	256, 1024
Anchors per cell	64
Window type	k-nearest neighbours, ambient metric
Confidence δ	0.05
Equivalence margin Δ	$\log 1.35$
Multiplicity correction	Holm, within collapse variable
Thinning ladder	$c \in \{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$
Ladder clustering unit	physical layer
Permutation replicates	1000 (collapse), 2000 (calibration)
Primary null	spatial-sign deformed Marchenko–Pastur
Graph degree for the detour certificate	$k = 20$
Density-conditioning k	16
Gap-closure decision rule	confirm $\geq 60\%$, disconfirm $\leq 20\%$
Excluded layers	lookup layer; ELMo character convolution

U Details of Re-analysis of a published effective-dimension study

(Cai et al., 2021) report, for five architectures on the Penn Treebank, two estimates of representation dimension that disagree by more than an order of magnitude: a spectral effective dimension $d(\varepsilon)$,

Formal statement	Measurement	Outcome	§
Thm 3.1, threshold equation 2	Planted spectra, 45 cells varying ρ_T, κ_T, $\mathbf{r}(\Sigma)$; smallest n_r recovering d at rank $q = d + 1$	Sound in $45/45$ (max constant 2.13 against $64C_0^2$); loose by a systematic $30\times$; measured scaling $\rho_T^2(\mathbf{r} + \log\frac{1}{\delta})$, slope 1.04	N.5
Thm 3.1, scope equation 1	Every candidate $d \leq 64$ on each real window's own spectrum	*Negative*: no d satisfies it in any window of $88/96$ cells at $n_r = 256$ and $45/96$ at $n_r = 1024$; 0.2% and 1.4% of windows admit any d	N.6
Thm 5.3, n_{stab} (plug-in)	$\hat{\varepsilon}^\star$ by bisection and $\mathbf{r}(\hat{\Sigma})$, per window, at ranks $q \in \{8, 16, 32, 64\}$ and $q = \hat{k} + 1$	*Inconclusive by construction*: (C) makes this an upper bound on n_{stab}, so exceeding n_r decides nothing; reported as a diagnostic only	N.7
Thm 5.3, n_{stab} (direct)	Disjoint-half agreement at 6–10 sample sizes, 3 ranks; halves share the window's covariance, so disagreement is variance alone	*Negative*: $(1-\delta)^2$ reached in $16/567$ cells; median tops out at 83.8%; extrapolates to 3–7× the available population	N.7
Cor. 5.4(i)–(ii)	The two rows above	Bias reading unavailable: the premise $n_r \geq n_{stab}$ is not met	N.7
Prop. 6.1	Planted spectra at eight amplitudes, four noise families, above-edge content held fixed	Confirmed: $\hat{k}$ spans $[3, 48]$ at one above-edge direction, contracting to $2.4\times$ once fifteen clear the edge	N.2
Def. 7.1, power of the gate	Planted structure $d \in \{32, 64\}$ under the matched spectrum, amplitudes 1.5–100	Boundary measured: $G = 1.000$ at $1.5\times$, 1.663 at $1.75\times$; monotone, so absence excludes a half-line	N.2
Prop. I.1	Curvature-free null at matched (n_r, d, k), $k \in \{6, \ldots, 100\}$, $n_r \leq 4097$	*Negative*: usable only at $d = 4$ and by 2%; no operating point at $d \geq 8$	N.8
Thm 4.2 (prediction for certificates)	The row above, read as an instance	Consistent: the geodesic certificate fails where the theorem says every window-only certificate must	N.8
Noise-floor admissibility	r_{noise}/r_1 per cell; gate cross-tabulated with the ball-mass exponent	Two-nearest-neighbour inadmissible in all 96 cells; ball-mass reports ≈ 15 where the gate is equivalent to noise, on both corpora	N.9

Table 2: Correspondence between the formal statements and the measurements that bear on them. Three of the ten outcomes are negative for the paper's hypotheses and one is inconclusive; all are reported as such.

the smallest k explaining a fraction $\varepsilon = 0.8$ of variance (their Table 1), and a density-based local intrinsic dimension (their Table 3). They diagnose, for a different quantity, the mechanism that would plausibly connect the two—global variance dominated by inter-cluster separation, which they remove by a per-cluster correction applied to cosine statistics and never to $d(\varepsilon)$. We pre-registered the prediction that applying that correction to $d(\varepsilon)$ would close the gap, with a fixed rule: at least 60% of the logarithmic gap closed to confirm, at most 20% to disconfirm.

The prediction is disconfirmed in all five architectures. Median closure is $+0.0\%$, $+0.0\%$, $+0.0\%$, -4.7% and -2.2% at the sampling that reproduces Table 1, and $+0.0\%$, $+14.9\%$, -0.3%, -4.6% and $+0.0\%$ at the sampling that reproduces Table 3. The failure is directional: where the clustering criterion selects more than one cluster, the correction moves $d(\varepsilon)$ *away* from the density estimate in 9 of 12 cells at the first sampling and 12 of 19 at the second. Partitioning removes the dominant inter-cluster eigenvalue and leaves flatter within-cluster spectra, so more components are needed to reach the same variance fraction; a correction diagnosed for cosine statistics does not transfer to a spectral one.

The two statistics were computed on different clouds. The released code exposes a per-type sampling parameter, and the two tables correspond to different settings of it. Table 1 is reproduced by a type-level cloud ($n \approx 2.9 \times 10^4$) to within 10% in all five models, and missed by the token-level cloud ($n \approx 1.2 \times 10^6$) by 28–$33\%$ in three of them; Table 3 is reproduced by the token-level cloud to within 6% in four models and missed by the type-level cloud by 52–87%. The two numbers whose disagreement is the paper's quantitative puzzle were evaluated on samples differing roughly fortyfold in size.

The published local dimensions are a factor of two low. The estimator of (Amsaleg et al., 2015) is defined on neighbour distances; the released code obtains them from an index returning *squared* Euclidean distances and passes them through unchanged. Since $\log(d^2/D^2) = 2\log(d/D)$, every estimate is halved. The as-coded variant matches Table 3 to within 6% in four models.

What remains is real. Recomputing both statistics on a common cloud with the estimator as defined reduces the published gaps—$48.9\times$, $30.2\times$, $45.2\times$, $31.4\times$ and $5.3\times$—by a factor of 2.1 to 2.8 in every model, leaving a residual of 9–$21\times$ (1.5–$1.9\times$ for GPT-2). Sampling mismatch and the distance error explain a consistent fraction; the part they do not explain is the divergence between spectral and density-based dimension that this paper concerns.

Which readouts track depth. Across the four architectures with enough layers to estimate a trend, the ball-mass exponent rises monotonically with depth at both samplings ($r = +0.71$ to $+0.99$), while $d(\varepsilon)$ falls in seven of the eight model–sampling combinations. The estimator of (Amsaleg et al., 2015) agrees with the ball-mass exponent at the token level ($r = +0.94$ to $+0.99$) but not at the type level, and the two-nearest-neighbour estimator is inadmissible in every valid cell— independently reproducing, on a second corpus and five further architectures, the admissibility finding of §N.9.

Layer conventions had to be established empirically and differ between the transformer extractor and the ELMo one; these, and the cells we cannot reproduce, are in Appendix O.

What the gate finds at the degenerate layers. Applying the gate of §7 to these windows settles what the published readout's collapse to $d(\varepsilon) = 1$ represents. At the four GPT-2 cells where the published value is 1 or 2, the profile depth is 1.00, 1.17, 1.15 and 1.00—below the equivalence margin in every case—while the Amsaleg estimator reports 15.6 to 25.0 and the ball-mass exponent 12.5 to 16.9 at the same cells. Nine of the twenty-five joined cells are equivalent to noise, so the finding is broader than the degenerate published values alone.

The agreement between $d(\varepsilon) = 1$ and $G = 1$ is not vacuous, because the two statistics measure different things. A value $d(\varepsilon) = 1$ records that one direction carries eighty per cent of the variance, which in these representations is the massive-activation direction; a value $G = 1$ records that once an anisotropy-matched null has absorbed that direction, nothing further remains. The gate therefore does not confirm that the local dimension is one. It confirms that what $d(\varepsilon)$ reads at these layers is global anisotropy, and that no additional local structure lies beneath it that the instrument can resolve at its bandwidth. The noise floor at these cells is 0.39 to 0.46 and the two-nearest-neighbour estimator is inadmissible in all 25, which is the same diagnosis reached independently in §N.9.

Profile depth does not trend consistently with depth across these architectures: $r = +0.94$ for BERT, $+0.79$ for GPT, $+0.43$ for GPT-2 and -0.44 for DistilBERT, against a monotone decline in the cohort of §N.3. The corpus and the sampling differ—Penn Treebank at type level against WikiText at token level—so the two are not in conflict, but the depth result of §N.3 should be read as a property of that cohort rather than of transformers generally.

V Discussion

One obstruction, in several places. The exponential dependence on intrinsic dimension that produces the curvature floor is not confined to curvature. It sets the gap width (Corollary 3.3); it makes the detour certificate unaffordable even though that certificate was constructed to avoid tangent spaces and curvature entirely (§N.8); it makes the density-conditioning number itself ill-conditioned to estimate, since ℓ_f is a ratio raised to the power d; and it is why a thinning ladder reaches only

a narrow range of bandwidth before the windows empty. A local geometric certificate that must resolve a d-dimensional neighbourhood pays $\exp(d)$ somewhere, and the useful design question is where, not whether.

What to report instead of a number. Our recommendation is concrete. A dimension readout should be accompanied by (i) a null matched to the data's anisotropy rather than an isotropic reference; (ii) a detection statistic such as G with a measured power bound, so that absence claims carry their disjunction; (iii) an admissibility check against the noise floor, which on our cohort disqualifies the most widely used estimator in every cell; and (iv) a scale profile rather than a single value, since Proposition 6.1 implies the value is a function of the scale at which the edge is crossed. Where a single number is wanted, the ball-mass exponent is the one our planted worlds and noise-floor criterion both support.

Relation to the Lipschitz argument. (Schulte & Rügamer, 2026) reach a compatible conclusion by a different route, showing that layer-wise dimension cannot increase under Lipschitz maps and therefore that rising nearest-neighbour estimates cannot track dimension. Their argument is at the population level and applies to the nearest-neighbour family; ours is finite-sample and applies to the spectral family. Read together they leave little room for either family to be interpreted as a dimension measurement on transformer representations without calibration of the kind we describe.

Limitations. Our cohort is three transformers with $D \leq 1600$; the scaling of these effects to frontier models is untested, though Corollary 3.3 predicts the gap narrows only logarithmically in D. The calibration of the spectral threshold varies ρ_T, κ_T and $\mathbf{r}(\Sigma)$, which are the quantities the threshold contains, but at a single ambient dimension $D = 1600$; since D does not appear in equation 2 this is not a gap in the test of that bound, though it does mean we report no evidence on ambient scaling. The cohort's windows do not clear the stability threshold, so the bias reading of Corollary 5.4(ii) is unavailable and we do not claim it. And the detour result concerns one family of implementations: a reach certificate not based on geodesics—via the bottleneck term of (Federer, 1959), say—is not excluded by it.