CERTIFIED OUT-OF-DISTRIBUTION SAFETY FOR GENERATIVE MAPS FROM IN-SUPPORT GEOMETRY ALONE

Anonymous authors
Paper under double-blind review

ABSTRACT

Every generative model is trained on a finite set of examples, and every generative model is eventually asked for something a little outside that set. What happens next is usually found out the hard way: you sample, you look, and you decide whether the output is still trustworthy. The awkward part is that checking whether a model behaves outside its training support seems to demand data from outside that support, which is exactly the data nobody has. We show the check can be done entirely from the inside. Near the edge of a model's support we fit a small atlas of local linear approximations to the map and read off three quantities that never require leaving the support: how well the atlas already fits, how rough the map looks at the finest scale the atlas can see, and how sharply it bends. Those three numbers name a radius: a distance beyond the edge within which the map's extrapolation error is guaranteed to stay under a tolerance the user picks. We test the prediction against ground truth on synthetic warped manifolds, on maps fitted to real tabular data, and on flow-matching networks trained on three image datasets. On the synthetic families the predicted radius sits inside the true one on every cell we measured. The certificate is conservative rather than optimistic, and tight enough to be useful rather than vacuous. On trained networks it does too, but only after one constant is recalibrated for the model class, and we report exactly what goes wrong when it is not. The algebra underneath is machine-checked in a proof assistant, so the parts that are theorems are theorems. Two findings surprised us. The bending term must be measured as the curvature of the map itself and not the curvature of the surface it draws. The second, more natural choice quietly smuggles in a dependence on how much the map contracts, and breaks the guarantee. And the constant tying the two roughness measurements together, genuinely constant on synthetic maps, becomes a per-model quantity with a heavy tail on trained networks, so it must be calibrated per model class and at a high quantile rather than at a typical value. We also report where the method stops: it certifies smooth deterministic samplers, and a diffusion-transformer sampler we tried falls outside that class for reasons we measure rather than guess.

1 INTRODUCTION

A trained generative map $F : \mathcal{Z} \to \mathcal{X}$ is only ever constrained where it saw data. Query it a little outside the training support and its behaviour is, in the strict sense, unspecified: nothing in the training objective says anything about that region. Practitioners still query there constantly: latent interpolation and extrapolation, guidance that pushes samples off the data manifold, conditioning on rare attributes, and simple sampling noise all reach past the support boundary.

The natural way to find out how far one can go is to look: sample outside, measure the damage, stop when it becomes unacceptable. This works, and it is also circular, because it needs exactly the out-of-distribution (OOD) data whose absence created the problem, and for a deployed model that is usually the data nobody has.

This paper asks whether the question can be answered without leaving the support, and answers yes for a well-delimited class of maps. The construction is geometric and small (Figure 1). Near the

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.

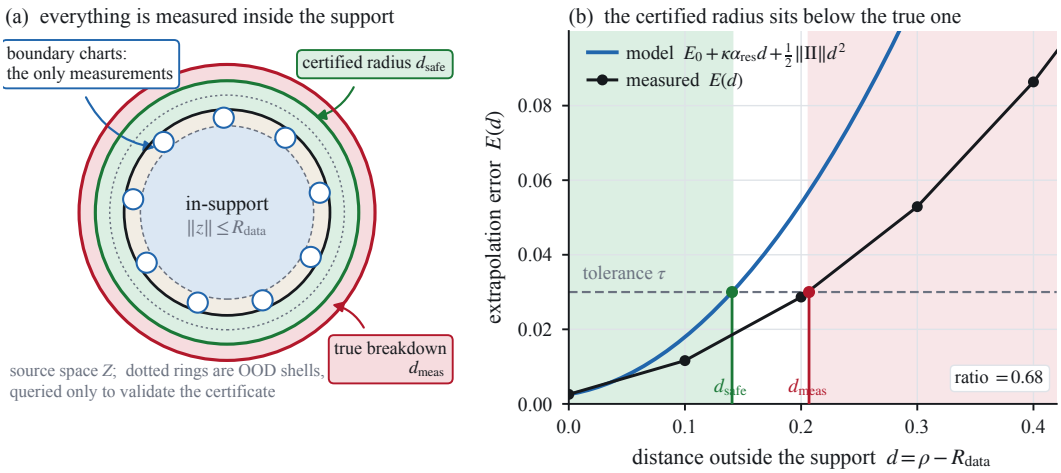


Figure 1: The certificate at a glance. (a) All measurements happen inside the support, on charts in a thin boundary band. The certificate returns a radius d_{safe}; the true breakdown radius d_{meas} is further out. The dotted OOD shells are queried only to *validate* the prediction, never to build it. **(b)** For one synthetic cell: the measured extrapolation error $E(d)$, the linear-floor model built from three in-support measurables, and the two radii. The ratio $d_{\text{safe}}/d_{\text{meas}} = 0.68 < 1$ is what safety means here.

boundary of the source support we cover a thin band with local charts and fit each chart with an affine map. Three numbers come out, all measured strictly inside the support: the residual error E_0 of the resulting atlas at the boundary; a scale α_{res} describing how much the map deviates from affine within a chart; and an aggregate $\|\text{II}\|$ of the fitted second derivatives over the boundary charts. From these we predict the radius at which the atlas' extrapolation error crosses a user-chosen tolerance τ:

$$d_{\text{safe}} \;=\; \text{the positive root of } \; E_0 + \kappa\,\alpha_{\text{res}}\,d + \tfrac{1}{2}\|\text{II}\|\,d^2 \;=\; \tau, \tag{1}$$

with one calibration constant κ. We then check the prediction against the truth (for validation only, never for fitting) by sampling shells outside the support and locating the real crossing d_{meas}. The certificate is *safe* on a cell when $d_{\text{safe}} < d_{\text{meas}}$, and its *tightness* is the ratio $d_{\text{safe}}/d_{\text{meas}}$.

Contributions.

1. **A fully in-support certificate** for the OOD safety of a generative map, with the closed form equation 1 and three in-support inputs (§3). Its soundness, uniqueness, monotonicity and scaling law are machine-checked in Lean 4/Mathlib (§3.4, Appendix A).
2. **The linear-floor model.** Modelling the error as a floor plus a linear term plus curvature, rather than curvature alone, gives a strictly tighter certificate at equal safety: median ratio 0.675 versus 0.53 with every input measured in-support, zero violations on 72 pre-registered fresh cells (§4.1).
3. **The raw-Hessian convention.** $\|\text{II}\|$ must be the norm of the map's fitted Hessian, not the normal-projected second fundamental form of the image. The projection reintroduces a dependence on the map's contraction that the certificate forbids; it inflates the median contraction spread from 0.043 to 0.234 and produces the only overclaiming variant in our discovery grid (§4.2).
4. **An exact in-support proxy for the linear term.** On the synthetic family, α_{res} and its OOD counterpart are both exactly proportional to the warp amplitude and exactly invariant to contraction and tolerance, so their ratio κ is a function of the chart radius alone (10.30 at r=0.12, 9.67 at r=0.08; §4.3).
5. **Transfer, and the κ heavy tail.** The certificate transfers in the median to LOT+TPS maps on six real tabular datasets and to flow-matching networks on MNIST, Fashion-MNIST and EMNIST, with two corrections: a maximum aggregate when the band is small, and κ set per map class at a high quantile, because its per-cell distribution is heavy-tailed. This restores safety on MNIST but not yet on every class (§4.5, §4.6).
6. **A measured scope boundary.** We report where the certificate does not apply, namely a converged diffusion-transformer sampler, and why, in measurements rather than speculation (§5).

2 BACKGROUND AND RELATED WORK

Detecting OOD versus bounding OOD. A large literature predicts *whether* an input is out of distribution: softmax baselines (Hendrycks & Gimpel, 2017), temperature and input perturbation (Liang et al., 2018), Mahalanobis scores (Lee et al., 2018); it also shows that generative likelihoods can be actively misleading (Nalisnick et al., 2019). Those methods classify points. We ask instead *how far* outside the support a given map stays accurate, and answer with a radius computed without OOD samples.

Guarantees of other kinds. Conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2023) and calibration or ensembling (Guo et al., 2017; Lakshminarayanan et al., 2017) give distributional guarantees whose exchangeability assumption is exactly what the shift we study breaks. Lipschitz certificates (Szegedy et al., 2014; Cissé et al., 2017; Fazlyab et al., 2019; Weng et al., 2018) and randomized smoothing (Cohen et al., 2019) bound worst-case behaviour globally and are typically vacuous at the radii of interest here; ours is local, quantile-based, and measured where data exist, closer in spirit to neural certificates in control (Chang et al., 2019).

Charts, and the maps we certify. Covering a manifold with local linear pieces is classical (Tenenbaum et al., 2000; Roweis & Saul, 2000; Belkin & Niyogi, 2003; Brand, 2002); we use an atlas not to embed data but as a measuring instrument for a map's second-order geometry (do Carmo, 1992), which is the right reference object given that networks extrapolate along essentially linear directions (Xu et al., 2021). Our real targets are flow-matching and rectified-flow maps (Lipman et al., 2023; Liu et al., 2023b; Albergo & Vanden-Eijnden, 2023) and optimal-transport maps (Villani, 2009; Peyré & Cuturi, 2019) with thin-plate-spline interpolation (Bookstein, 1989); diffusion samplers (Ho et al., 2020; Song et al., 2021; Peebles & Xie, 2023; Liu et al., 2023a) are stochastic, and §5 reports what happens when we try to certify one. Appendix J gives the extended discussion.

3 THE CERTIFICATE

3.1 SETUP

Let $F : \mathcal{Z} \to \mathcal{X} \subseteq \mathbb{R}^D$ be a deterministic, queryable map from a d_0-dimensional source with support radius R_{data}: a sphere cap for the synthetic family, an empirical latent cloud for trained models. Write $S_\rho = \{z : \|z\| = \rho\}$ for the shell at radius ρ and $d = \rho - R_{\text{data}} \geq 0$ for the distance outside the support.

Boundary atlas. Sample chart centres c by farthest-point sampling at spacing $0.7r$ and keep those in the boundary band $|\,\|c\| - R_{\text{data}}\,| \leq r$ (on the synthetic family the band half-width is 0.12 at both radii). On each chart we fit, in local d_0-dimensional tangent coordinates t, both an affine model $F \approx b + Mt$ and a quadratic model $F \approx b + Mt + \frac{1}{2}t^\top Ht$ by least squares. The *atlas* F_{atlas} is the distance-blended combination of the affine charts; it is the object whose extrapolation error we certify.

The extrapolation error. For $\rho \geq R_{\text{data}}$ define $E(d) = Q_{0.99}\,\|F_{\text{atlas}}(z) - F(z)\|$ over $z \in S_{R_{\text{data}}+d}$, the 99th percentile discrepancy between the atlas and the true map on the shell. E is increasing in practice and $E(0) = E_0$ is the in-support floor.

3.2 THREE IN-SUPPORT MEASURABLES

Figure 2 shows what one chart supplies.

- $E_0 = Q_{0.99}\|F_{\text{atlas}} - F\|$ evaluated *at* R_{data}: the error the atlas already makes where data exist.
- $\alpha_{\text{res}} = Q_{0.90}$ over boundary charts of $\text{std}(\text{affine residual})/r$: the scale of the map's departure from affine, per unit chart radius.
- $\|\mathrm{I\!I}\|_{q90} = Q_{0.90}$ over boundary charts of $\|H\|$, the norm of the *fitted map-Hessian* in tangent coordinates. Not the normal-projected second fundamental form of the image; see §4.2. On real and trained maps the aggregate switches to $\max$ when the band holds fewer than 100 charts (§4.5); the synthetic bands (42-62 charts) keep $Q_{0.90}$, validated in §4.4.

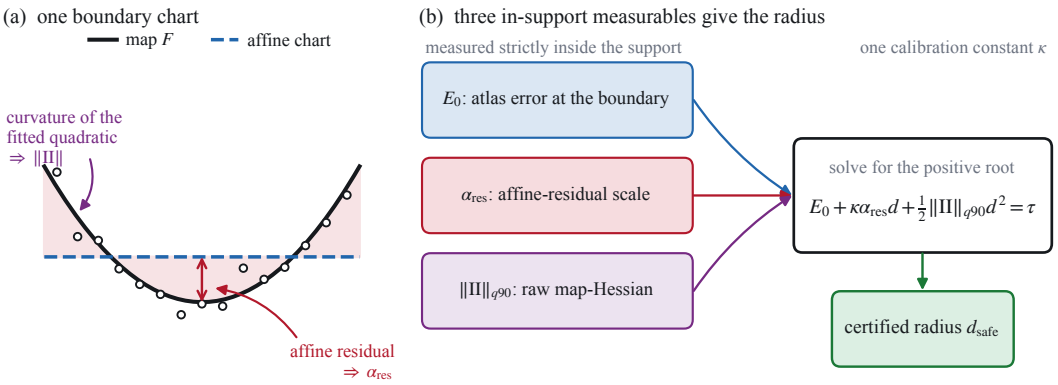


Figure 2: The measurement pipeline. (a) One boundary chart supplies two of the three inputs: the spread of the affine residual gives α_{res}, and the curvature of the fitted quadratic gives $\|\mathrm{II}\|$. **(b)** The three in-support measurables enter one quadratic whose positive root is the certified radius. Only κ is calibrated, and it is fixed before any validation run.

All three are computed from samples with $\|z\| \le R_{\mathrm{data}}$ only.

3.3 The linear-floor model and the certified radius

We model the extrapolation error as a floor, a linear term and a curvature term,

$$\widehat{E}(d) \;=\; E_0 \;+\; \kappa\,\alpha_{\mathrm{res}}\,d \;+\; \tfrac{1}{2}\|\mathrm{II}\|_{q90}\,d^2, \tag{2}$$

and define the certified radius as the positive root of $\widehat{E}(d) = \tau$,

$$d_{\mathrm{safe}} \;=\; \frac{-\kappa\alpha_{\mathrm{res}} + \sqrt{(\kappa\alpha_{\mathrm{res}})^2 + 2\|\mathrm{II}\|_{q90}\,(\tau - E_0)}}{\|\mathrm{II}\|_{q90}}. \tag{3}$$

The root exists and is unique whenever $\tau > E_0$ (Prop. 1). On the synthetic family τ is a pre-registered grid variable (Appendix B); on real and trained maps, which have no natural error scale, $\tau = \max(0.02,\, 2E_0)$, so the certificate asks how far the error may double. The quadratic coefficient is the Taylor constant $\frac{1}{2}$ for a map with second derivative bounded by $\|\mathrm{II}\|$; the linear term models the atlas' own fit noise, which a curvature-only bound cannot see and which dominates at small d (Figure 3a).

Why one constant. α_{res} is measured *inside* a chart of radius r while the linear term of $E(d)$ grows *outside* the support, so the two are related by a factor κ of the chart construction. Fixing κ before validation is what keeps the certificate predictive rather than fitted. §4.3 shows κ is an exact function of r on the synthetic family; §4.6 shows what happens on trained networks.

The pure-curvature form. With $E_0 = \kappa\alpha_{\mathrm{res}} = 0$, equation 3 reduces to $\sqrt{2\tau/\|\mathrm{II}\|}$ (Prop. 4). Our early rounds used it with a fixed, never-refit offset, $d_{\surd} = \sqrt{2\tau/\|\mathrm{II}\|} - 0.6r$; the Law-A, aggregator and source-dimension tests, which concern $\|\mathrm{II}\|$ alone, use $d_{\surd}$, and Table 2 marks the form behind every synthetic number.

3.4 What is proved, and what is measured

We separate the two cleanly. The algebraic and analytic backbone of equation 2-equation 3 is formalised in Lean 4 with Mathlib (de Moura & Ullrich, 2021; The mathlib Community, 2020); all twenty-four theorems compile against the kernel and depend only on the three standard Mathlib axioms, with no `sorry`. The empirical claims (the measured constants, the measured ratios) are experiments, and are reported as such. Appendix A gives the statements and the audit trail; the illustration is Figure 9.

Proposition 1 (Existence and uniqueness, `model_dsafe`, `dsafe_unique_pos`)**.** *For $\|\mathrm{II}\| > 0$, $\kappa\alpha_{\mathrm{res}} \ge 0$ and $\tau > E_0$, the radius equation 3 is strictly positive, satisfies $\widehat{E}(d_{\mathrm{safe}}) = \tau$, and is the unique nonnegative solution.*

Proposition 2 (Soundness, `certificate_sound`, `dsafe_le_dmeas`)**.** *If $E(d) \le \widehat{E}(d)$ for all $d \in [0, d_{\text{safe}})$, then $E(d) < \tau$ for every such d; consequently $d_{\text{safe}} \le d_{\text{meas}}$, i.e. the measured safety ratio is at most* 1.

Proposition 3 (Monotonicity and safety inheritance, `dsafe_le_of_inputs_le`)**.** *d_{safe} is antitone in each of E_0, $\kappa\alpha_{\text{res}}$ and $\|\mathrm{II}\|$, and monotone in τ. Replacing any input by an upper bound can therefore only shrink the certified radius.*

Proposition 3 is why the aggregator question has a clean answer. Since $\text{median} \le Q_{0.90} \le \max$ pointwise on any finite sample, a certificate that is sound at the median is automatically sound at $Q_{0.90}$ and at the max, and never the reverse (`dsafe_quantile_anti`). §4.4 shows the median is in fact *not* sound, which is what forces the choice.

Proposition 4 (The $+1/2$ law, `dsafe_half_power`)**.** *With $E_0 = \alpha_{\text{res}} = 0$, $d_{\text{safe}} = \sqrt{2\tau/\|\mathrm{II}\|}$, so scaling τ by c^2 scales d_{safe} by c: on a log-log plot the exponent is exactly* $1/2$.

Proposition 5 (Law A, `secondDeriv_shrink_invariant`)**.** *For any continuous linear L, vector v and twice-differentiable q, the second derivative of $z \mapsto Lz + v + q(z)$ equals that of q. On a flat source the raw map-Hessian, and hence its norm and d_{safe}, cannot depend on the map's affine part; a curved source adds a chart-coordinate term that we measure in §4.2. The normal projection does not have this property (`projected_not_invariant`).*

4 Experiments

Protocol and discipline. Every protocol change was written into the run script's docstring *before* the run; discovery and fresh-seed cells are always reported separately; and κ and τ are fixed before any validation. Safety means $d_{\text{safe}} < d_{\text{meas}}$ on every cell (the measurable form of Prop. 2), tightness a median ratio of at least 0.5. Full grids are in Appendix B, every round's numbers in Appendix C (Table 2).

Map families. (i) *Synthetic*: $F(u) = u\,D_{\text{lin}}^{\top} + a_w \sin(u\cdot w/\ell_w)\,v$ on a d_0-sphere cap embedded in $\mathbb{R}^6$, where $D_{\text{lin}} = I - s\,u_2 u_2^{\top}$ contracts along one direction by s ("shrink") and a_w sets the warp amplitude; $d_0 \in \{2, 3, 4\}$. (ii) *Real 2D*: linear-optimal-transport plus thin-plate-spline maps fitted to six tabular datasets. (iii) *Trained networks*: conditional flow-matching maps $\mathbb{R}^2 \to \mathbb{R}^{784}$ on MNIST, Fashion-MNIST and EMNIST letters. (iv) *Diffusion transformer* (§5).

4.1 The linear floor buys tightness at no cost in safety

Fitting $E(d) = E_0 + \alpha d + \beta d^2$ on the train grid gives $\beta/(\frac{1}{2}\|\mathrm{II}\|_{q90}) = 0.497$: the certificate's curvature coefficient is about twice the measured one, the conservatism a bound should have. But a curvature-only model has $\widehat{E}(0) = 0$ and under-predicts the error at small d, where the atlas' own fit noise dominates (Figure 3a,b). Adding the floor and the linear term, all inputs in-support and $\kappa = 9.99$ fixed on train seeds, raises the median ratio from 0.53 ($d_{\surd}$) to 0.675 on two pre-registered fresh seed sets (36 cells each, max 0.845, zero violations; Figure 3c, Table 2). The certificate becomes tighter while remaining conservative, which is the direction that matters.

4.2 Law A, and why the projected second fundamental form is wrong

The safe radius must not depend on how much the map contracts: contraction changes the affine part of F, and an affine map has no second derivative (Proposition 5). This is a falsifiable prediction, and it discriminates sharply between two candidate definitions of $\|\mathrm{II}\|$.

The ground truth obeys it exactly: with the sample fixed, the atlas error $E(d)$ is identical at every contraction level (Appendix D), and no fresh cell's certified ratio moves by more than 0.025 across contraction (Figure 4a). The estimators differ. Varying *only* the contraction, the $Q_{0.90}$ aggregate of the raw map-Hessian moves by at most 0.41% at the standard chart radius and the normal-projected form by up to 3.16%, a 6-12$\times$ separation. Downstream, the median ratio spread on the discovery grid rises from 0.043 to 0.234 (Figure 4c), and the projected variant is the only one there that overclaims (max 1.325 at r=0.12, 1.349 at r=0.08; Appendix E).

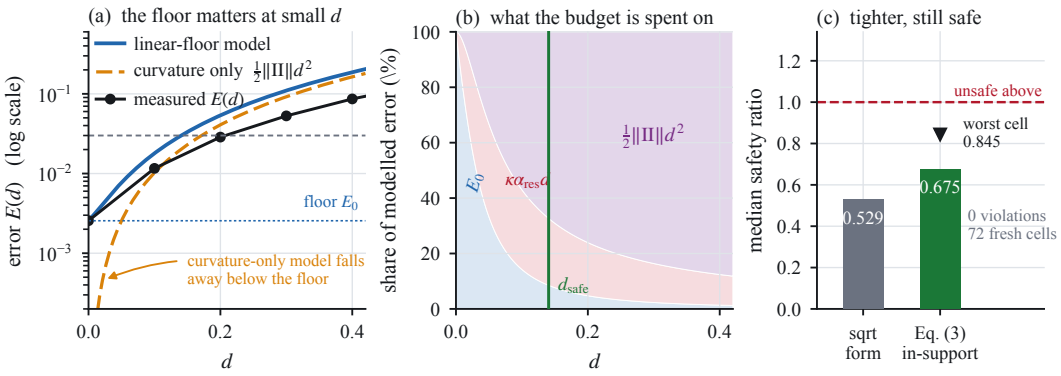


Figure 3: The linear-floor model. (a) Log-scale view of one cell: the curvature-only model falls away below the in-support floor E_0 at small d, while the linear-floor model tracks the measured curve. **(b)** Share of the modelled error contributed by each term; at the certified radius the floor and linear terms still carry a substantial fraction. **(c)** On two pre-registered fresh seed sets (36 cells each) the fully in-support root, equation 3, is tighter than the square-root form $d_{\surd}$ at equal (zero-violation) safety.

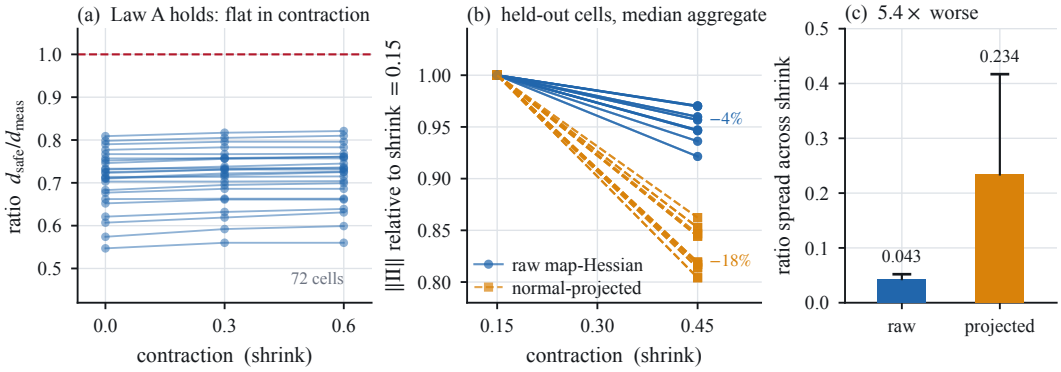


Figure 4: Law A and the raw-Hessian correction ($d_{\surd}$). **(a)** The certified ratio is flat across contraction on all 72 fresh cells. **(b)** Held-out cells, median aggregate: the raw map-Hessian moves by $\approx 4\%$ (source curvature), the projected form by $\approx 18\%$. **(c)** Median ratio spread on the discovery grid.

Honest caveat. Proposition 5 is exact on a flat source. On the sphere cap an ambient-linear map is not affine in a chart's tangent coordinates, so single raw Hessians carry a contraction-dependent source-curvature term (norm $\approx \sqrt{2}$ at $a_w = 0$, where the atlas error is exactly zero; up to 35% change per chart). The $Q_{0.90}$ aggregate screens it out, the median does not (Figure 4b), and at $r{=}0.08$ even the $Q_{0.90}$ spread reaches 4.4% (Appendix D).

4.3 The linear term has an exact in-support proxy

The linear coefficient α_{curve} of the true error curve is an OOD quantity; α_{res} is its in-support proxy. Over the 36-cell grid the two correlate at $+0.998$, but the structure is sharper than a correlation (Figure 5). Because the synthetic map is an affine part plus a warp of amplitude a_w, both α_{res} and α_{curve} are *exactly* proportional to a_w and *exactly* invariant to contraction and to the tolerance, to six decimal places. Their ratio therefore depends on the chart radius alone:

$$\kappa(r{=}0.12) = 10.2995, \qquad \kappa(r{=}0.08) = 9.6729.$$

The value $\kappa = 9.99$ used throughout is the median over a grid mixing the two radii, and its "spread" $[9.67, 10.30]$ is this r-dependence, not seed noise; we keep the single pre-fixed value because real and trained maps offer no r-resolved calibration. A re-run of the estimator reproduces the logged correlation and both κ values exactly.

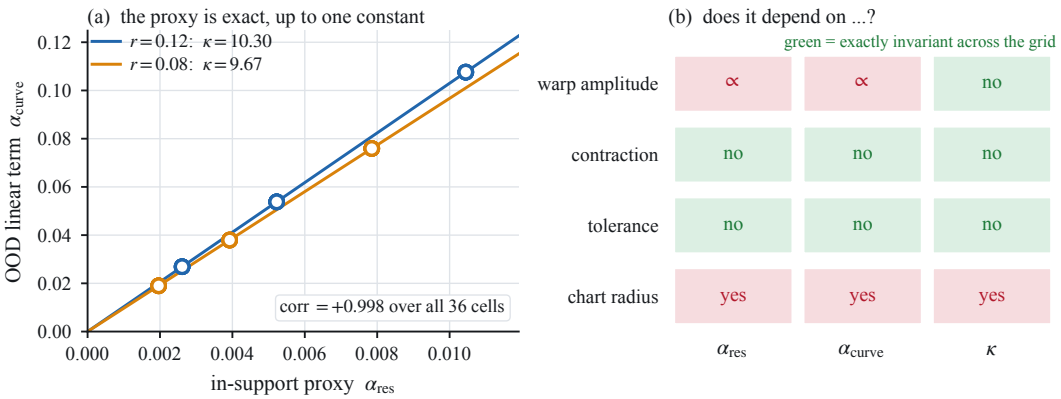


Figure 5: The non-circular linear-term estimator. **(a)** Each point is a grid cell; the 36 cells collapse onto six distinct measurements lying exactly on two rays, one per chart radius. **(b)** What the two α's and their ratio depend on. Green cells are exact invariances, not approximate ones.

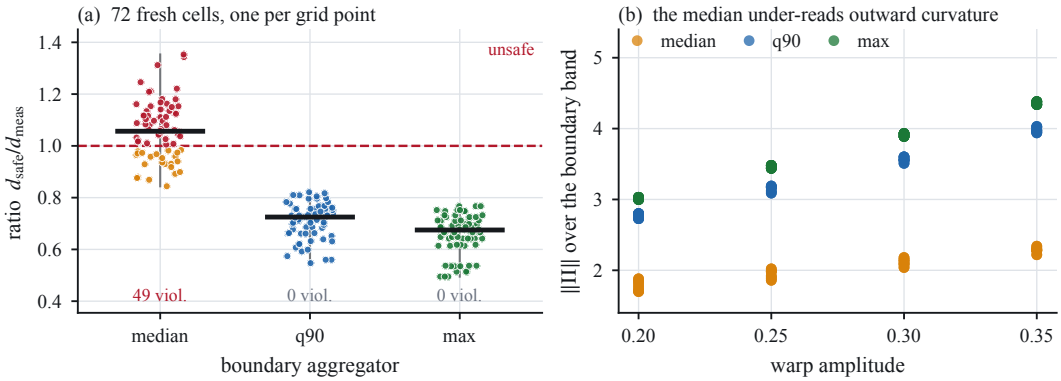


Figure 6: Choosing the boundary aggregate. **(a)** All 72 fresh cells under the three aggregators; red points violate safety. **(b)** The mechanism: as warp amplitude grows, the median of the per-chart curvature falls increasingly far below the $Q_{0.90}$ and max that track the breakdown.

4.4 The boundary aggregate must be a high quantile

Proposition 3 says a larger aggregate is safer; it does not say how large is large enough. That is empirical, and it was decided on held-out data (all with $d_{\surd}$). The median aggregate is safe up to warp amplitude 0.20 on the discovery seeds and 0.15 on held-out seeds 4-6, but overclaims at 0.25 (seven held-out cells, ratios 1.017-1.053); the mechanism is that the median under-represents the outward curvature that actually drives the breakdown (Figure 6b). We pre-registered $Q_{0.90}$ on that mechanistic ground and validated on seeds 7-9 at amplitudes 0.20-0.35, data never used to make the choice. On those 72 fresh cells the median aggregate produces 49 violations (median ratio 1.057, max 1.353); $Q_{0.90}$ produces none (median 0.725, max 0.821); the max produces none and is looser (median 0.675) (Figure 6a). $Q_{0.90}$ is the tightest safe choice.

4.5 Transfer to real maps

Real 2D geometry. On LOT+TPS maps fitted to six tabular datasets the certificate transfers in the safe direction. Real maps are far rougher than the synthetic family (E_0 up to 1.15, $\|\mathrm{I\!I}\|$ up to 229 versus ≈ 3); at the standard chart radius all six discovery cells are safe. At the fine radius the band can hold as few as 27 charts and the sample $Q_{0.90}$ becomes biased low; the max below 100 charts is safe on every fresh seed (median 0.342, max 0.666), but one discovery cell (penguins, 1.134) and one cell each of diamonds and flights still overclaim (Appendix G).

Trained networks. Conditional flow-matching maps $\mathbb{R}^2 \to \mathbb{R}^{784}$ are much smoother than the TPS maps ($E_0 \approx 0.02$-0.13, $\|\mathrm{I\!I}\|_{q90} \approx 2.4$-$8.6$), and every certificate input remains in-support measurable.

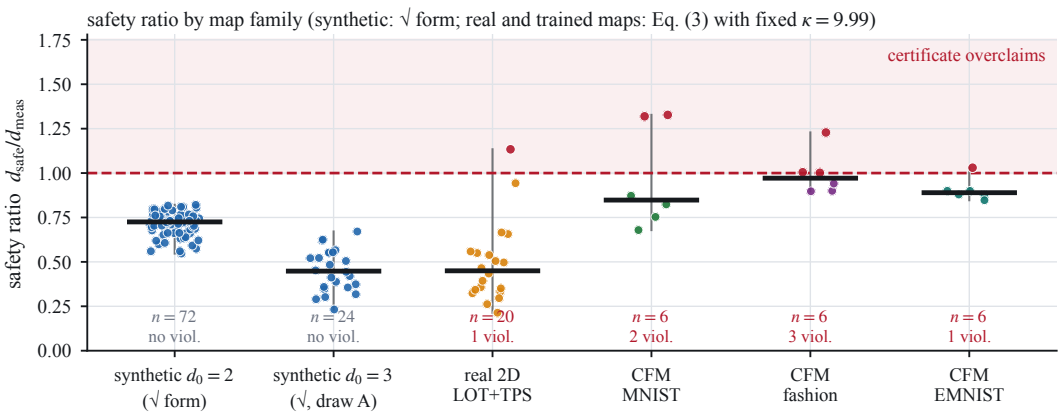


Figure 7: Safety ratio by map family. Each dot is one cell, bars are medians, red dots overclaim. Synthetic columns use $d_{\surd}$, the others equation 3 with $\kappa = 9.99$. Not shown: 16 further real 2D cells (2 overclaim, Table 4) and the fresh-seed runs of Tables 2 and 5.

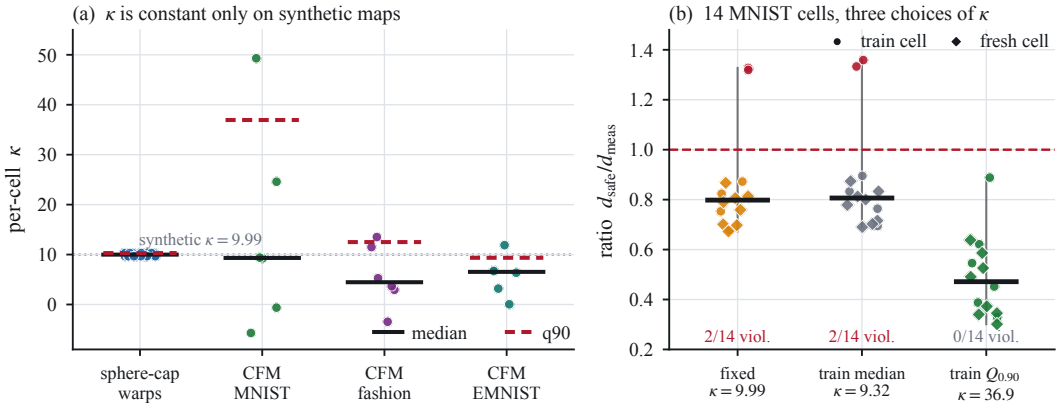


Figure 8: The κ heavy tail. (a) Per-cell κ by map class. On synthetic warps the values are exact constants; on trained networks they are dispersed, with a tail far above the median. **(b)** Fourteen MNIST cells under three choices of κ: the train median is no safer than 9.99; the train $Q_{0.90}$ is safe but loose.

With κ fixed at the synthetic value the predicted radius lands in the right range on all three image datasets, with median ratios 0.848 (MNIST), 0.971 (Fashion) and 0.890 (EMNIST), but the structure is right while the scale is not: $2/6$, $3/6$ and $1/6$ cells overclaim (Figure 7; Appendix H, Table 5). The failure is mostly in the linear term: five of the six overclaiming cells have $\kappa_{\text{cell}} > 9.99$ (§4.6).

4.6 κ IS NOT UNIVERSAL, AND ITS MEDIAN IS NOT THE RIGHT SUMMARY

The per-cell values $\kappa_{\text{cell}} = \alpha_{\text{curve}}/\alpha_{\text{res}}$ tell the story (Figure 8a). On sphere-cap warps they are two exact constants. On trained networks they are dispersed and heavy-tailed, spanning -5.7 to 49.3 on MNIST alone; Table 6 lists every per-cell value. The medians (9.3, 4.5, 6.5) sit well below the tail, and it is the tail that breaks safety: a cell with $\kappa_{\text{cell}} = 49$ is one whose true linear growth is five times what a median-calibrated certificate assumes. Negative values appear because α_{curve} is estimated from a short finite difference and is noisy at small d; they are a second reason the median is the wrong summary.

The median does not fix this. The MNIST train median, $\kappa = 9.32$, is *smaller* than 9.99, so by Proposition 3 it can only enlarge the radius; two train cells still overclaim, and the eight fresh cells are safe under both values only because they avoid the tail. The train $Q_{0.90}$, $\kappa = 36.9$, makes all fourteen MNIST cells safe at a fresh median ratio of 0.43, below our tightness target (Figure 8b). It is necessary but not sufficient: the per-class $Q_{0.90}$ of Fashion (12.5) and EMNIST (9.35) still leaves one overclaiming cell each, even in-sample. Turning the heavy tail into a concentration statement about the boundary atlas is the main theoretical question this paper leaves open.

5 LIMITATIONS AND SCOPE

It is a bound, not an equality. Tightness is a measured function of curvature: on the discovery grid, ratios run from 0.36-0.51 at the lowest warp amplitude to 0.79-0.96 at the highest. A user who needs the exact breakdown radius will not get it here; one who needs a radius that is never optimistic will. What is bounded is the atlas' extrapolation error, which is a proxy for perceptual or downstream degradation rather than identical to it.

κ is a per-class constant with a heavy tail. This is the certificate's weakest point (§4.6). A fixed κ is safe for the typical cell and unsafe on the tail; a high quantile over a handful of cells from the target class fixes MNIST but not every class we tried. We have no theory for the tail.

Validated source dimension ≤ 4. Safety holds at $d_0 = 2, 3, 4$ with zero violations throughout; tightness, the Law-A spread and the τ exponent vary with the warp draw, and one $d_0 = 3$ draw misses the tightness and spread targets (0.448, 0.056; Appendix F). Beyond $d_0 = 4$ is untested.

A measured scope boundary: stochastic and stiff samplers. Flow matching yields a deterministic ODE map by construction, which is exactly why the certificate applies to it; diffusion does not. We looked for the boundary there with a conditional diffusion transformer trained with the UniDB bridge objective (Zhu et al., 2025), certified through a 2D PCA slice of its sampler. The model converged (50k steps on $3{,}000$ images, final loss 5.1×10^{-6}), yet both sampler variants leave the certified class: the deterministic reverse ODE is stiff (outputs reach ± 78 in a $[0, 1]$ pixel space), and the fixed-seed reverse SDE is not locally affine-trackable ($E_0 \approx 94$-137 against $\tau \approx 188$-274, *no finite crossing* in any cell; Appendix I, Figure 16).

The scope is therefore *smooth deterministic maps*: flow-ODE maps, TPS maps, trained velocity-field Euler maps. The three exits we measured (Figure 17) are a sampler that is not a deterministic function of the latent, one that is not locally affine, and a boundary band too small for a quantile, which the max-guard handles but not perfectly. Knowing where a certificate stops is part of the certificate.

6 CONCLUSION

Bounding a generative map's out-of-distribution behaviour looks like it needs out-of-distribution data, and does not: three in-support measurements determine a radius within which the extrapolation error provably stays under tolerance whenever the error model dominates, as it does on every fresh synthetic cell and, with κ at a high per-class quantile, on trained MNIST networks. Explaining the heavy tail of κ, and extending the certificate past smooth deterministic samplers, are the natural next steps.

IMPACT STATEMENT

This work makes it possible to state, before deployment, how far outside its training support a generative model may be queried without its outputs degrading past a chosen tolerance. The intended effect is defensive: a conservative radius lets a practitioner refuse queries rather than serve silently degraded ones, and it removes the need to collect out-of-distribution data to establish that boundary. Two risks deserve naming. First, a certificate invites over-trust: ours bounds the deviation of a map from its own local linearisation, which is not the same as bounding harm, bias or perceptual quality. Second, the guarantee is conditional on assumptions a user can violate (smooth deterministic maps, a well-populated boundary band, and a κ calibrated on the right map class), and a certificate applied outside those conditions is worse than none, because it looks like evidence. We mitigate both by reporting the scope boundary as a measurement (§5), by publishing failure modes alongside successes, and by machine-checking the provable parts. The method needs no personal data and adds no capability.

AI USE STATEMENT

In this work, generative AI tools were used to (i) create or modify visual figures based on our experimental results, (ii) edit English grammar, phrasing, and vocabulary, and (iii) assist in producing Lean formalizations for checking selected mathematical equations. The tools were not used to generate data, formulate the research questions or theoretical claims, design or run experiments, analyze or interpret experimental results, or make scientific decisions. They were also not used to replace the authors' judgment about the scientific content.

We reviewed all AI-assisted work. The experimental data and results were checked against the original runs and stored result files; AI-assisted figures were checked against those results; and the Lean formalizations were compiled and checked by the authors to ensure that they corresponded to the equations stated in the paper. AI-assisted language edits were manually reviewed for meaning and accuracy. The authors take responsibility for the final content of this work, including all text, claims, figures, and formalizations produced with AI assistance.

REFERENCES

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.

Matthew Brand. Charting a manifold. In *Advances in Neural Information Processing Systems*, pp. 961–968, 2002.

Ya-Chien Chang, Nima Roohi, and Sicun Gao. Neural lyapunov control. In *Advances in Neural Information Processing Systems*, pp. 3240–3249, 2019.

Moustapha Cissé, Piotr Bojanowski, Edouard Grave, Yann N. Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pp. 854–863, 2017.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320, 2019.

Leonardo de Moura and Sebastian Ullrich. The lean 4 theorem prover and programming language. In *International Conference on Automated Deduction*, pp. 625–635, 2021.

Manfredo P. do Carmo. *Riemannian Geometry*. Birkhäuser, 1992.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J. Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 11423–11434, 2019.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pp. 6402–6413, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pp. 7167–7177, 2018.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A. Theodorou, Weili Nie, and Anima Anandkumar. I^2sb: Image-to-image schrödinger bridge. In *International Conference on Machine Learning*, pp. 22042–22062, 2023a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023b.

Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, pp. 4172–4182, 2023.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

The mathlib Community. The Lean mathematical library. In *ACM SIGPLAN International Conference on Certified Programs and Proofs*, pp. 367–381, 2020.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. In *International Conference on Learning Representations*, 2018.

Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S. Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. How neural networks extrapolate: From feedforward to graph neural networks. In *International Conference on Learning Representations*, 2021.

Kaizhen Zhu, Mokai Pan, Yuexin Ma, Yanwei Fu, Jingyi Yu, Jingya Wang, and Ye Shi. UniDB: A unified diffusion bridge framework via stochastic optimal control. In *International Conference on Machine Learning*, 2025.

A MACHINE-CHECKED STATEMENTS

The algebraic and analytic backbone of the certificate is formalised in Lean 4 against Mathlib (de Moura & Ullrich, 2021; The mathlib Community, 2020). The development is a single self-contained file, `DSafe/Basic.lean`, supplied with the code. It defines

$$\texttt{model}\ E_0\ b\ a\ d\ =\ E_0 + bd + ad^2, \qquad \texttt{dsafe}\ E_0\ b\ a\ \tau\ =\ \frac{-b + \sqrt{b^2 + 4a(\tau - E_0)}}{2a},$$

matching Equations 2-3 with $b = \kappa\alpha_{\mathrm{res}}$ and $a = \frac{1}{2}\|\mathrm{II}\|_{q90}$.

Table 1 lists every statement.

Table 1: The twenty-four machine-checked statements, grouped by what they license in the main text. All depend only on the three standard Mathlib axioms; none depends on `sorryAx`.

Lean name	Statement
Existence and uniqueness of the root	
`disc_pos`	the discriminant is positive when $\tau > E_0$, $a > 0$, $b \geq 0$
`dsafe_pos`	$d_{\mathrm{safe}} > 0$ in that regime
`model_dsafe`	$\widehat{E}(d_{\mathrm{safe}}) = \tau$
`model_strictMonoOn`	$\widehat{E}$ is strictly increasing on $[0, \infty)$
`model_lt_tau_of_lt_dsafe`	$d < d_{\mathrm{safe}} \Rightarrow \widehat{E}(d) < \tau$
`dsafe_unique_pos`	d_{safe} is the unique nonnegative root
Soundness	
`certificate_sound`	$E \leq \widehat{E}$ on $[0, d_{\mathrm{safe}}) \Rightarrow E(d) < \tau$ there
`dsafe_le_dmeas`	hence $d_{\mathrm{safe}} \leq d_{\mathrm{meas}}$: the safety ratio is at most 1
Monotonicity	
`dsafe_anti_a`	antitone in the curvature coefficient
`dsafe_anti_b`	antitone in the linear coefficient
`dsafe_anti_E0`	antitone in the in-support floor
`dsafe_mono_tau`	monotone in the tolerance
`dsafe_le_of_inputs_le`	upper-bounding every input can only shrink d_{safe}
Scaling	
`dsafe_pure_curvature`	$d_{\mathrm{safe}} = \sqrt{\tau/a}$ when $E_0 = b = 0$
`dsafe_half_power`	$\tau \mapsto c^2\tau$ scales d_{safe} by c
`log_dsafe_pure`	the log-log slope in τ is exactly $1/2$
Law A and the raw-Hessian convention (flat source)	
`fderiv_add_affine`	$D(Lz + v + q) = L + Dq$: an affine part shifts the first derivative by a constant
`secondDeriv_shrink_invariant`	$D^2(Lz + v + q) = D^2 q$ for continuous linear L
`secondDeriv_norm_shrink_invariant`	hence $\|\mathrm{II}\|$ is unchanged by the affine part
`rawNorm_const_in_tangent`	the raw norm is constant in the tangent map
`projected_not_invariant`	the projected norm is not: explicit counterexample
Quantile aggregation	
`quantile_mono`	order statistics are monotone in the level
`dsafe_quantile_anti`	d_{safe} is antitone in the quantile level
Taylor	
`taylor_quadratic_bound`	$\|g(t) - g(0) - t\,g'(0)\| \leq \frac{1}{2}Kt^2$ whenever $\|g'(s) - g'(0)\| \leq Ks$

What is *not* formalised. The Lean development says nothing about whether $E \leq \widehat{E}$ holds for any particular map: that hypothesis is exactly what the experiments test, and it is where a certificate can fail. Nor does it formalise the estimators (the chart fit, the quantile aggregate, the sampling of shells), which are numerical procedures. The Law-A statements are about a flat source; on a curved source the per-chart tangent-coordinate Hessian carries an extra term (Appendix D). The division of labour is deliberate: everything that can be a theorem is one, and everything else is reported as a measurement with its protocol.

The soundness picture. Figure 9 illustrates the three propositions that carry the argument: domination implies the ratio is at most one; the radius is antitone in every input; and therefore a larger boundary aggregate is always the safer choice.

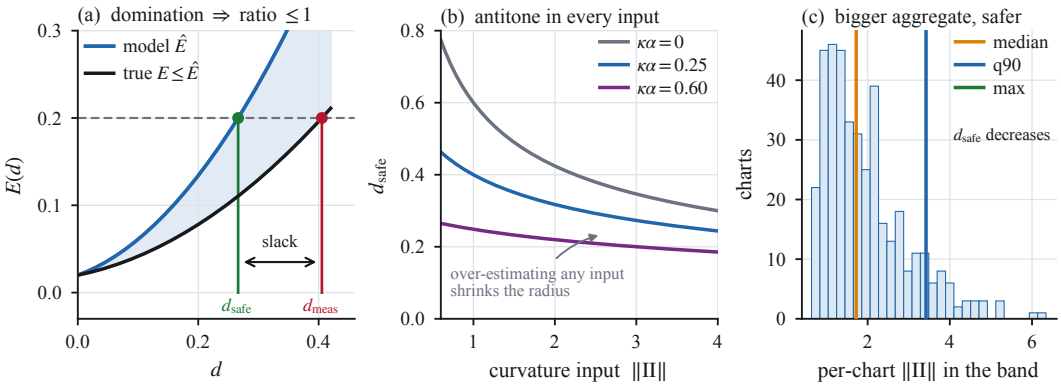


Figure 9: What the machine-checked propositions say. (a) Soundness: if the model dominates the true error curve, its crossing radius lies inside the true one and the slack is exactly the conservatism. **(b)** The radius is antitone in each input, so over-estimation is always safe. **(c)** Consequently a higher quantile of the per-chart curvature yields a smaller, safer certified radius.

B PROTOCOL AND PRE-REGISTRATION

Discipline. Every protocol change was written into the run script's docstring before the run, together with the outcome space and the decision rule. Three rules were fixed for the whole project and never relaxed: (i) κ and τ are fixed before any validation run; (ii) a correction is adopted only if it is justified mechanistically *and* validated on seeds never used to choose it; (iii) the offset $c = 0.6r$ in the earlier square-root certificate is never refit.

Definitions. R_{data} is the support radius (0.5 for sphere caps; 1 for the unit-disk source of the real 2D maps; the 99th-percentile latent radius, rescaled to 1, for trained maps). Chart centres come from farthest-point sampling at spacing $0.7r$; the boundary band is $|\,\|c\| - R_{\text{data}}\,| \le r$ on real and trained maps and $|\,\|c\| - R_{\text{data}}\,| \le 0.12$ at both radii on the synthetic family; a synthetic quadratic chart is valid if it holds at least 15 points. The atlas is the distance-blended affine one. $E(d)$ is the 99th-percentile discrepancy between atlas and map on the shell at $R_{\text{data}} + d$, evaluated on 16 shells spanning $[R_{\text{data}}, 1.25]$ with 900 samples each (synthetic), 16 shells spanning $d \in [0, 1]$ with 900 (real 2D) or 300 (MNIST) samples, and 10 shells spanning $d \in [0, 2]$ with 60 samples (Fashion-MNIST, EMNIST, where the 99th percentile is effectively the maximum). d_{meas} is the first crossing of τ, found by log-linear interpolation of E between shells; a cell with no crossing in range is reported as such and never imputed. On the synthetic family τ is a grid variable; on real and trained maps $\tau = \max(0.02, 2E_0)$.

Grids.

- *Discovery* (seeds 1-3): $r \in \{0.12, 0.08\}$, $a_w \in \{0.05, 0.10, 0.20\}$, shrink $\in \{0.0, 0.3, 0.6\}$, $\tau \in \{0.02, 0.05\}$.
- *Held-out* (seeds 4-6): $r \in \{0.12, 0.08\}$, $a_w \in \{0.03, 0.07, 0.15, 0.25\}$, shrink $\in \{0.15, 0.45\}$, $\tau \in \{0.015, 0.04, 0.08\}$; 48 cells, of which the 4 with $a_w = 0.03$, $\tau = 0.08$ have no crossing in range, leaving $n = 44$. (The per-cell log prints shrink to one decimal, as 0.1 and 0.5.)
- *Fresh validation* (seeds 7-9): $r \in \{0.12, 0.08\}$, $a_w \in \{0.20, 0.25, 0.30, 0.35\}$, shrink $\in \{0.0, 0.3, 0.6\}$, $\tau \in \{0.02, 0.06, 0.10\}$; 72 cells.
- *Equation 3 fresh sets* (seeds $\{7, 8\}$ and $\{9, 10\}$): the discovery grid, with the certificate inputs and κ taken from train seeds 1-3 and d_{meas} from the fresh seeds; 36 cells each.

Every synthetic round before the equation 3 sets used the square-root form $d_{\surd}$ of §3; Table 2 marks each row.

Criteria. PASS requires zero violations ($d_{\text{safe}} < d_{\text{meas}}$ on every cell) and a median ratio of at least 0.5. A violation is never tuned away; it is reported with the regime in which it occurred.

C FULL SYNTHETIC RESULTS

Table 2: Synthetic warp family, all rounds. "Discovery" cells were used to form hypotheses; "fresh" cells were never used to make the choice under test. Form: $\surd$ is the square-root form $d_{\surd}$ of §3; (3) is equation 3 with $\kappa = 9.99$ and every input in-support; $(3)^*$ fits the linear coefficient to the train seeds' OOD error curves and is a diagnostic, not a certificate.

Round	form	set	n	median	min	max	viol.
Discovery, r=0.12 (raw, median agg.)	$\surd$	discovery	18	0.589	0.355	0.860	0
Discovery, r=0.08 (raw, median agg.)	$\surd$	discovery	18	0.673	0.436	0.958	0
Held-out, median aggregate	$\surd$	held-out	44	0.721	0.257	1.053	7
Round 3, median aggregate	$\surd$	fresh	72	1.057	0.844	1.353	49
Round 3, $Q_{0.90}$ aggregate	$\surd$	fresh	72	**0.725**	0.547	**0.821**	**0**
Round 3, max aggregate	$\surd$	fresh	72	0.675	0.495	0.767	0
Eq. (3), seeds $\{7, 8\}$	(3)	fresh	36	**0.675**	0.507	0.845	**0**
Eq. (3), seeds $\{9, 10\}$	(3)	fresh	36	**0.675**	0.514	0.835	**0**
same cells, seeds $\{7, 8\}$	$\surd$	fresh	36	0.525	0.322	0.726	0
same cells, seeds $\{9, 10\}$	$\surd$	fresh	36	0.529	0.328	0.725	0
G2 root, fitted α	$(3)^*$	fresh	24	0.753	-	0.869	0
G2 square-root baseline	$\surd$	fresh	24	0.605	-	-	0
G1 rank-2 warp, Frobenius	$\surd$	fresh	24	0.618	0.346	0.801	0
G1 rank-2 warp, operator	$\surd$	fresh	24	0.627	0.347	0.807	0
G4 multi-frequency warps	$\surd$	fresh	24	0.570	0.331	0.752	0
G3 d_0=3, warp draw A	$\surd$	fresh	24	0.448	0.232	0.671	0
G3 d_0=3, warp draw B	$\surd$	fresh	24	0.589	0.289	0.731	0
G7 d_0=4	$\surd$	fresh	24	0.513	0.220	0.675	0

Warp draw A at d_0=3 is safe but misses the pre-registered tightness target (median $0.448 < 0.5$); we report it rather than replace it with draw B.

Figure 10 shows the round-3 grid cell by cell. The ratio rises with warp amplitude and with tolerance and is essentially flat in contraction, which is Law A read directly off the raw numbers.

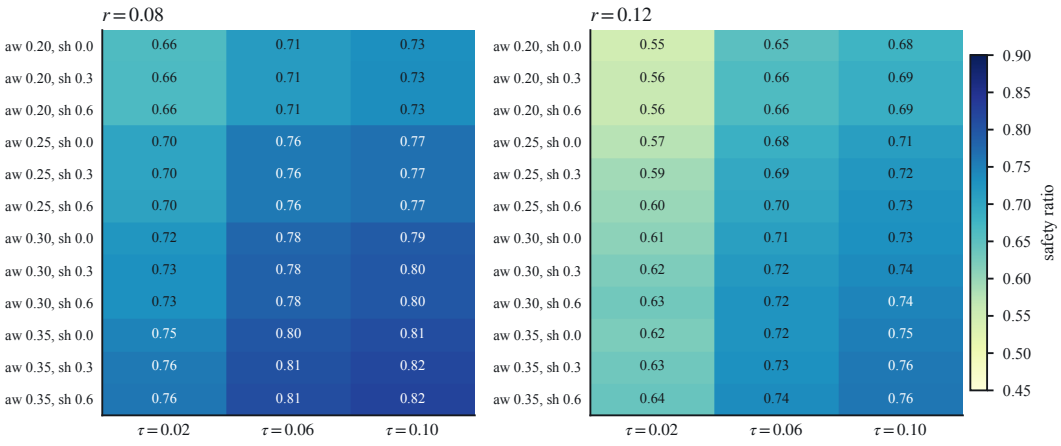


Figure 10: The full round-3 grid. Safety ratio under the $Q_{0.90}$ aggregate for every one of the 72 fresh cells, at both chart radii. No cell reaches 1.

The tolerance exponent. Proposition 4 is exact for the formula; the experiment asks whether the *measured* radius follows the same exponent. Figure 11 fits $\log(d_{\text{meas}} + 0.6r)$, the quantity $d_{\surd}$ predicts, against $\log \tau$. The slope is $+0.499$ against $+0.500$ on the sweep shown; across warp amplitudes 0.05, 0.10 and 0.20 it is $+0.496$, $+0.499$ and $+0.514$.

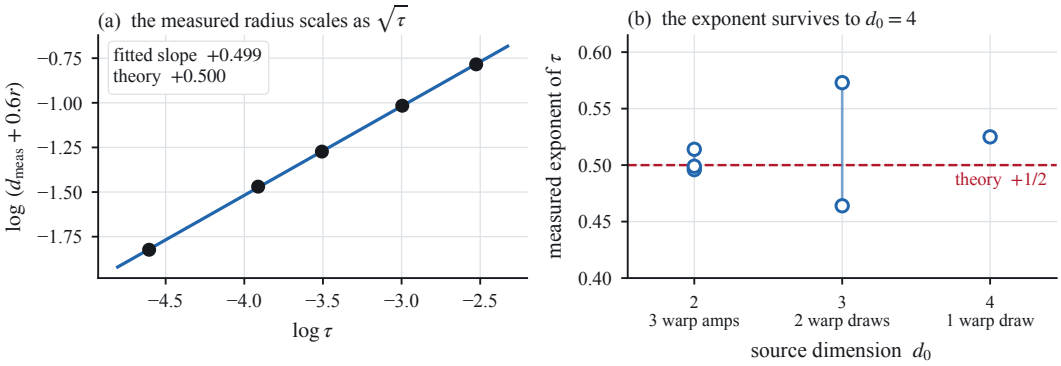


Figure 11: The $+1/2$ tolerance law. (a) One tolerance sweep, log-log, with the fitted slope. **(b)** The exponent across source dimension; at $d_0 = 3$ two warp draws span $+0.46$ to $+0.57$, so the exponent carries a warp-draw spread rather than a single value.

D THE CONTROLLED LAW-A EXPERIMENT

The discovery grid already reuses the same sample and chart centres at every contraction level, but it aggregates the per-chart Hessians by the band median. Here we hold the sample, the chart centres and the warp fixed, vary *only* the contraction $s \in \{0, 0.1, \ldots, 0.6\}$, and look at three objects separately: the error curve, the single charts, and the $Q_{0.90}$ aggregate (Table 3).

The error curve is exactly invariant. The atlas is affine in ambient coordinates, so the map's linear part contributes no extrapolation error: at $a_w = 0$, $E(d) \le 7.5 \times 10^{-15}$ on every shell, and for $a_w \in \{0, 0.05, 0.20\}$ the curves at $s = 0$ and $s = 0.6$ agree to within 3.4×10^{-15} (`lawA_per_chart.py`).

Single charts are not. The source is a sphere cap, and in a chart's tangent coordinates an ambient-linear map is not affine, so each raw per-chart Hessian carries a source-curvature term that depends on s. At $a_w = 0$, where the map is purely linear and the atlas error is zero, the raw norm is $\approx \sqrt{2}$ (1.42); from $s = 0$ to $s = 0.6$ single charts move by up to 35% (median 8%), and the band median falls from 1.42 to 1.30 ($1.53 \to 1.33$ at $a_w = 0.05$, the drift seen in the discovery grid). Proposition 5 does not cover this term: it is a statement about a flat source.

The $Q_{0.90}$ aggregate is flat. The top-decile charts are those where the contraction direction is nearly tangent, so the aggregate the certificate uses barely moves ($1.421 \to 1.417$ at $a_w = 0$), while the projected form moves several times more.

Table 3: Controlled Law-A test. Relative range of the boundary aggregate over the full contraction sweep, with everything else held fixed. At the standard chart radius the raw map-Hessian is invariant to within a fraction of a percent; the normal-projected form is not.

	$r = 0.12$ (42 charts)		$r = 0.08$ (62 charts)	
a_w	raw	projected	raw	projected
0.05	**0.41%**	2.61%	0.34%	1.99%
0.07	**0.35%**	2.95%	0.18%	1.87%
0.10	**0.29%**	3.16%	1.00%	1.69%
0.15	**0.21%**	2.49%	3.21%	1.64%
0.20	**0.16%**	1.00%	4.41%	1.58%

The asymmetry in Table 3 is worth stating plainly. At $r = 0.12$ the $Q_{0.90}$ aggregate is flat. At $r = 0.08$ the quadratic fit uses fewer points per chart, and at high warp amplitude the $Q_{0.90}$ over 62 noisier per-chart Hessians, which now include more of the source-curvature term, drifts more than the projected quantity does. This is a different small-sample effect from the one behind the max-guard of §4.5 (few points per chart rather than few charts), and it argues for choosing the chart radius large

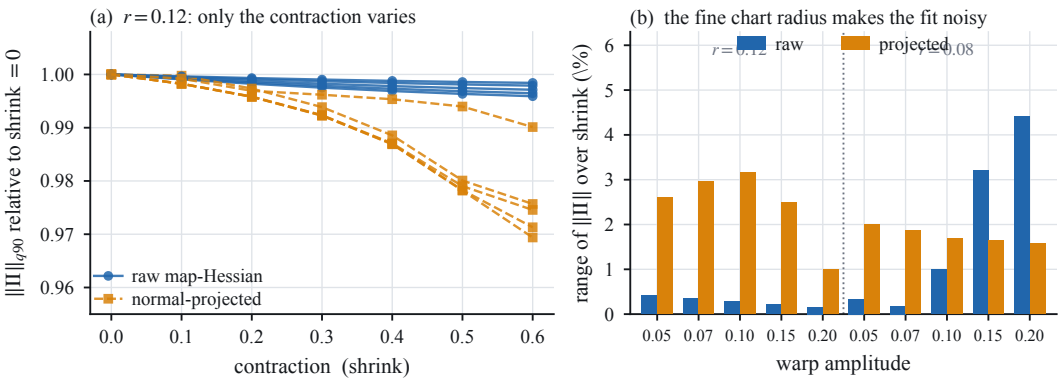


Figure 12: Only the contraction varies. (a) At $r = 0.12$ the raw map-Hessian is flat across the contraction sweep while the projected form drifts steadily downward. **(b)** At the fine chart radius each chart has fewer points, the per-chart quadratic fit is noisier, and at high warp amplitude the raw aggregate's spread exceeds the projected one: the source-curvature term is no longer screened out by the $Q_{0.90}$.

enough to fit a quadratic well. Removing the source-curvature term from the estimator, so that single charts obey Law A too, is a natural refinement we leave open.

E ATLAS AND NORM VARIANTS

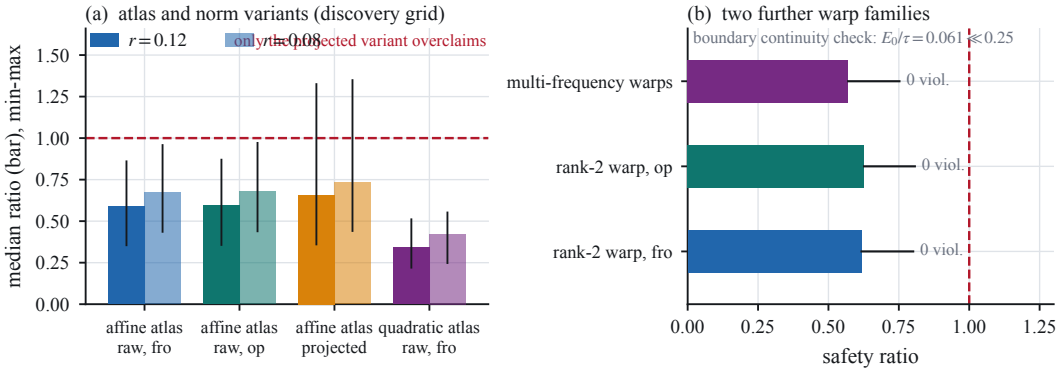


Figure 13: Variants. (a) Four combinations of atlas and norm on the discovery grid at both chart radii. The projected variant is the only one whose worst cell exceeds 1. The unblended nearest-chart *quadratic* atlas is roughly $2.7\times$ more accurate than the bound predicts, which is what makes the certificate a bound for affine-chart atlases rather than an equality. **(b)** Two further warp families, plus the boundary-continuity check.

Norm convention and fixed constants. The warp $a_w \sin(u \cdot w/\ell_w)v$ has rank-one curvature, so the Frobenius and operator norms of the fitted Hessian agree to within 1.5% and the question is untestable in that testbed. On a rank-two warp family both are safe and the operator norm is marginally tighter (median 0.627 versus 0.618). Every other reported run uses the Frobenius norm, which is never smaller than the operator norm and therefore conservative by Proposition 3. The offset $c = 0.6r$ in $d_\surd$ was fixed at the outset and never refit. Figure 13 compares the four atlas/norm combinations.

Boundary continuity. The in-support floor and the OOD error curve are one object measured in two regimes. If they were discontinuous at the boundary the model would be misspecified there. Measured across 24 cells, E_0/τ has median 0.061, far below the 0.25 threshold set in advance, confirming that the atlas error at the boundary is a small fraction of the tolerance and that the curve is continuous through $d = 0$.

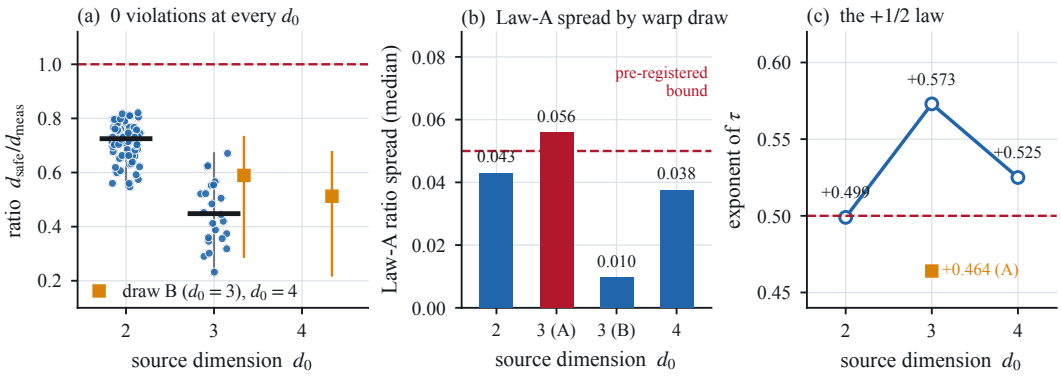


Figure 14: Scaling in source dimension ($d_{\surd}$ form). **(a)** Safety ratios at $d_0 = 2, 3, 4$; zero violations everywhere. Strips: round 3 at $d_0 = 2$ and warp draw A at $d_0 = 3$; squares: warp draw B at $d_0 = 3$ and the single $d_0 = 4$ draw. **(b)** Median Law-A contraction spread: warp draw A at $d_0 = 3$ exceeds the pre-registered bound of 0.05, the others stay under it. **(c)** The tolerance exponent stays near $+1/2$.

F SOURCE DIMENSION

Figure 14 collects the three panels. At $d_0 = 3$ two independent warp draws give tolerance exponents of $+0.464$ and $+0.573$; at $d_0 = 4$ a single draw gives $+0.525$. We report the exponent as $+0.5$ with a warp-draw spread of $-0.04/+0.07$ rather than as a precise constant. Safety is unaffected: zero violations at every d_0. Tightness and the Law-A spread are not: warp draw A at $d_0 = 3$ has median ratio 0.448 and median spread 0.056 (max 0.074), both just outside the pre-registered targets, while draw B has 0.589 and 0.010; at $d_0 = 2$ and 4 the median spreads are 0.043 and 0.038 (maxima 0.052 and 0.051).

G REAL 2D MAPS

Construction. For each tabular dataset we take two numeric columns, standardise them, and build a smooth map from the unit disk by pairing 60 farthest-point knots to the data cloud under a linear optimal-transport plan and interpolating with a thin-plate spline (Bookstein, 1989). The result is a genuinely rough real-geometry map that is still queryable everywhere, which is what the protocol needs.

Table 4: Real 2D maps, fixed $\kappa = 9.99$, max-guard active when the band holds fewer than 100 charts.

Dataset / set	n	**median**	**min**	**max**	**viol.**
penguins + mpg, discovery	12	0.501	0.296	1.134	1
penguins + mpg, fresh seeds	8	0.342	0.214	0.666	**0**
iris	4	0.489	0.265	0.711	0
tips	4	0.607	0.281	0.875	0
diamonds	4	0.402	0.320	1.544	1
flights	4	0.431	0.237	1.191	1

Real maps are much rougher than the synthetic family: E_0 reaches 1.15 and $\|\mathrm{I\!I}\|$ reaches 229, against $E_0 \approx 0.003$ and $\|\mathrm{I\!I}\| \approx 3$ for sphere-cap warps. That roughness is precisely why the small-sample quantile bias bites, and it is the regime in which a fixed κ is least reliable.

H TRAINED FLOW-MATCHING NETWORKS

Model. A conditional flow-matching map $F : \mathbb{R}^2 \to \mathbb{R}^{784}$: a 2D Gaussian latent, a fixed two-frame embedding, an MLP velocity field, and an Euler ODE solver. Trained on 8000 images for 3000 steps (MNIST) or on 6000 images for 1500 steps (Fashion-MNIST, EMNIST). The map is deterministic and queryable anywhere, which is what the certificate requires.

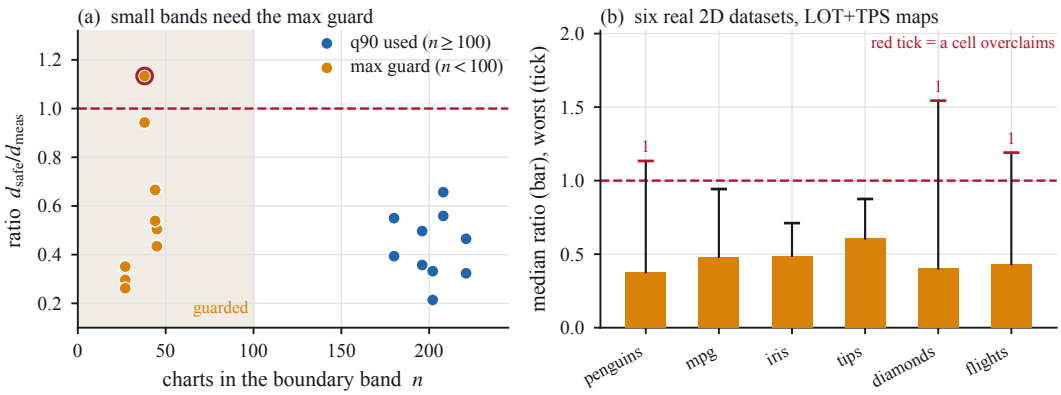


Figure 15: Real 2D transfer. (a) Safety ratio against the number of charts in the boundary band (penguins and mpg, all 20 cells, max guard active below 100 charts). The one overclaiming cell (penguins, 38 charts, 1.134) sits in the small-band regime and overclaims even under the guard. **(b)** Per-dataset medians and worst cells across the six datasets; three datasets have one overclaiming cell each.

Table 5: Trained flow-matching maps. The structure transfers: all three certificate inputs remain in-support measurable and the predicted radius is in the right range, but a fixed κ overclaims on part of each dataset. Rows marked † are recomputed on the logged cells from their logged inputs with the same root formula (`checks/recompute_kappa.py`); $Q_{0.90}$ is taken over the train cells, so its train and Fashion/EMNIST rows are in-sample.

Dataset, cells	n	E_0	$\|\mathbb{I}\|_{q90}$	**median**	**max**	**viol.**
MNIST, $\kappa = 9.99$, train	6	0.020-0.047	2.75	0.848	1.327	2
MNIST, $\kappa = 9.99$, fresh†	8	0.022-0.097	2.83	0.775	0.867	0
MNIST, $\kappa = 9.32$ (median), train†	6	0.020-0.047	2.75	0.864	1.359	2
MNIST, $\kappa = 9.32$ (median), fresh	8	0.022-0.097	2.83	0.789	0.874	0
MNIST, $\kappa = 36.9$ ($Q_{0.90}$), train†	6	0.020-0.047	2.75	0.499	0.888	**0**
MNIST, $\kappa = 36.9$ ($Q_{0.90}$), fresh†	8	0.022-0.097	2.83	0.432	0.638	**0**
Fashion-MNIST, $\kappa = 9.99$	6	0.061-0.125	8.55	0.971	1.229	3
Fashion-MNIST, $\kappa = 12.5$ ($Q_{0.90}$)†	6	0.061-0.125	8.55	0.899	1.167	1
EMNIST letters, $\kappa = 9.99$	6	0.019-0.040	2.38	0.890	1.030	1
EMNIST letters, $\kappa = 9.35$ ($Q_{0.90}$)†	6	0.019-0.040	2.38	0.905	1.044	1

Table 6: Per-cell $\kappa_{\text{cell}} = \alpha_{\text{curve}}/\alpha_{\text{res}}$ by map class. On synthetic warps these are two exact constants; on trained networks the distribution is dispersed and heavy-tailed, and its median lies well below the tail that breaks safety. The diffusion transformer is out of scope (Appendix I) and not listed.

Map class	**per-cell κ**	**median**	$Q_{0.90}$
sphere-cap warps	exactly 10.30 (r=0.12), 9.67 (r=0.08)	9.99	10.30
CFM MNIST	9.2, 24.6, −5.7, −0.6, 49.3, 9.4	9.32	36.9
CFM Fashion	13.5, 5.3, 2.9, 11.5, −3.5, 3.7	4.46	12.5
CFM EMNIST	6.8, 3.2, 6.4, 0.0, 6.7, 11.9	6.53	9.35

Why negative cells appear. α_{curve} is obtained from a two-point finite difference of the measured error curve at the smallest two shells. On a smooth map with a small linear term that difference is dominated by sampling noise and can come out negative, which is why κ_{cell} is occasionally negative. This is an estimator artefact of the *validation* quantity, not of the certificate inputs; α_{res}, E_0 and $\|\mathbb{I}\|$ are all stable. It is a second reason to calibrate κ from a high quantile over several cells rather than from a median, and it means a κ calibration should never be attempted from a single cell.

I THE DIFFUSION-TRANSFORMER ATTEMPT

What we built. A conditional diffusion transformer (hidden 192, depth 6, 6 heads, patch 4; 4.18M parameters) trained with the UniDB diffusion-bridge objective (Zhu et al., 2025) on EMNIST letters, and a certificate hookup that takes a 2D slice of its sampler: the class-mean prototype plus the top-two

PCA directions of the noise input, with the remaining noise held fixed, giving a map $\mathbb{R}^2 \to \mathbb{R}^{784}$ of exactly the form the certificate expects. The source support is the empirical PCA cloud and the OOD shells extend radially along real directions.

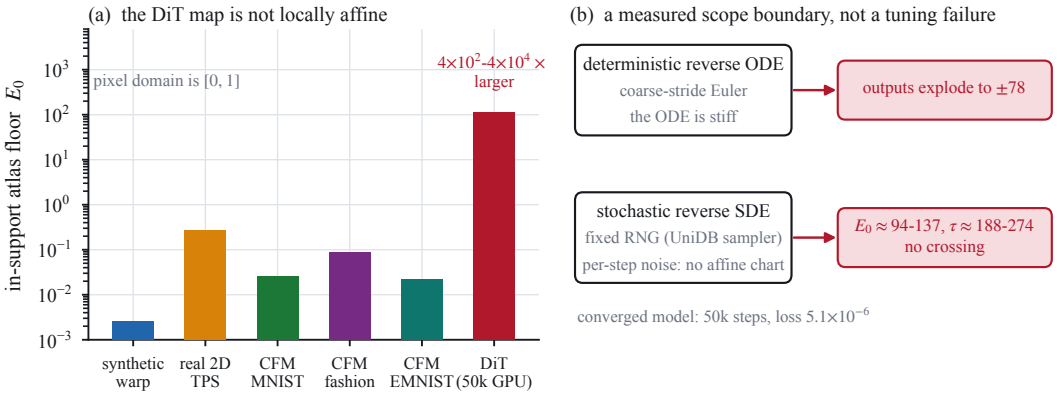


Figure 16: The scope boundary, measured. **(a)** Median in-support atlas floor E_0 by map class, log scale. The converged DiT sampler's floor is about $400\times$ that of the real 2D maps, 1.3-5 $\times$ 10^3 times that of the flow-matching maps and 4×10^4 times that of the synthetic warps; its local behaviour is not affine at any chart radius we can fit. **(b)** Both sampler variants, and the distinct reason each leaves the certified class.

Two runs. A first 2000-step CPU model, trained on 60,000 images, produced a reverse map that exploded to $[-71, +65]$ in a pixel space whose data domain is $[0, 1]$, which we attributed to undertraining (its certificate log was not retained). We therefore retrained on a GPU for 50000 steps on 3,000 images to a final loss of 5.1×10^{-6}. The converged model behaves the same way (outputs in $[-77.9, +69.9]$), which is what turns the result from a training problem into a scope statement.

Both sampler variants leave the class.

- *Deterministic reverse ODE* (coarse-stride Euler): outputs explode to ± 78 even at convergence. The ODE is stiff at the step budgets we could afford; the UniDB reference code relies on the stochastic reverse SDE, whose noise term regularises the trajectory.
- *Fixed-seed stochastic reverse SDE* (UniDB's own sampler, made deterministic by fixing the RNG): the map is deterministic but not locally affine-trackable, because per-step noise makes a boundary affine chart meaningless. Measured across four train cells: $E_0 \in [94.2, 136.9]$, $\tau = 2E_0 \in [188, 274]$, $\alpha_{\text{res}} \in [6.5, 12.0]$, and *no finite crossing* in any cell, train or fresh.

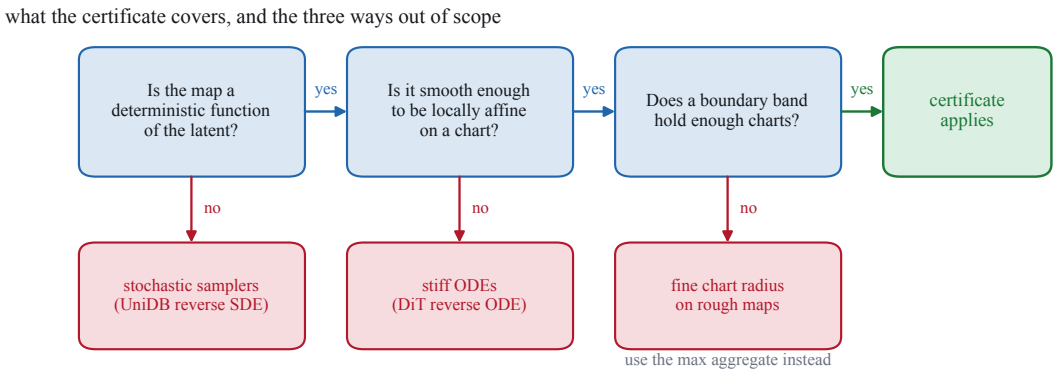


Figure 17: When the certificate applies. The three exits are the three failure modes we measured, not hypothetical ones.

What this is and is not. It is not a negative result about the certificate's correctness: the certificate's hypothesis (that the map is locally affine to within a small floor) is directly measurable, and here it is measurably false: E_0 is $400\times$ to $4 \times 10^4\times$ that of every map the certificate covers, and larger than

the diameter $\sqrt{784} = 28$ of the pixel domain itself. The useful output is a decision rule (Figure 17) and a diagnostic: if the in-support atlas floor E_0 is not small relative to the scale of the data, the certificate should not be used, and that check itself costs no OOD data.

J Extended related work

Detecting OOD versus bounding OOD. A large literature predicts *whether* a given input is out of distribution: softmax baselines (Hendrycks & Gimpel, 2017), temperature and input perturbation (Liang et al., 2018), Mahalanobis scores (Lee et al., 2018), and the observation that likelihoods from deep generative models can be actively misleading (Nalisnick et al., 2019). These methods classify points. We ask a different question: given a map and a tolerance, *how far* outside the support does the map remain accurate? The output is a radius, not a label, and it is computed without OOD samples.

Distribution-free uncertainty. Conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2023) gives finite-sample coverage guarantees, but its exchangeability assumption is precisely what fails under the shift we study; calibration and ensembling (Guo et al., 2017; Lakshminarayanan et al., 2017) similarly quantify uncertainty on the training distribution. Our guarantee is geometric rather than distributional: it bounds the deviation of one deterministic map from its own local linearisation.

Certified robustness. Lipschitz-based certificates (Szegedy et al., 2014; Cissé et al., 2017; Fazlyab et al., 2019; Weng et al., 2018) and randomized smoothing (Cohen et al., 2019) bound worst-case behaviour in a ball around an input. Global Lipschitz bounds for realistic generative networks are typically so loose as to be vacuous at the radii of interest. Our bound is local and quantile-based: it is not a worst case over the network, but a bound on the atlas' extrapolation error whose inputs are measured where data exist. It resembles neural certificates in control (Chang et al., 2019) in spirit (a certificate object fitted to data and then validated), but it is constructed rather than learned.

Local charts and the manifold hypothesis. Covering a manifold by local linear pieces is classical (Tenenbaum et al., 2000; Roweis & Saul, 2000; Belkin & Niyogi, 2003; Brand, 2002), and the manifold hypothesis itself is testable (Fefferman et al., 2016). We use a chart atlas not to embed data but as a *measuring instrument* for a map's local geometry, and specifically for its second-order behaviour (do Carmo, 1992). That neural networks extrapolate along essentially linear directions (Xu et al., 2021) is what makes an affine atlas the right reference object.

The maps we certify. Flow matching and rectified flows (Lipman et al., 2023; Liu et al., 2023b; Albergo & Vanden-Eijnden, 2023) produce deterministic ODE maps from latent to data space and are our main real target. Diffusion samplers (Ho et al., 2020; Song et al., 2021), including transformer-based ones (Peebles & Xie, 2023) and diffusion bridges (Liu et al., 2023a), are stochastic; §5 reports what happens when we try to certify one. Optimal-transport maps (Villani, 2009; Peyré & Cuturi, 2019) with thin-plate-spline interpolation (Bookstein, 1989) supply our real 2D targets.