TAMEPROBE: AN EMPIRICAL TAMENESS SPECTRUM FOR NEURAL DECISION BOUNDARIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Model-theoretic *o-minimality* ("tameness") has recently been proposed as a lens on neural network geometry, but it is a binary structural property that is not directly measurable on a trained model. We turn tameness into an operational, continuous *diagnostic spectrum*: two cheap, model-agnostic statistics of a network's decision boundary—the *oscillation complexity* (OC, sign flips and local extrema of the decision margin along random 1D probes) and the *decomposition complexity* (DC, number of distinct linear regions crossed). Across a controlled matrix of MLP architectures, seeds and post-training operators (an alignment-style Jacobian/weight regularizer, knowledge distillation, and 8-bit quantization; $14{,}400$ measured slices), we find that (i) tameness is *not* scalar: the alignment proxy lowers both OC (-35.4%) and DC (-30.3%), whereas distillation lowers OC (-17.9%) while *raising* DC ($+1.0\%$), so the two axes decouple; (ii) 8-bit quantization does not move the spectrum ($p > 0.18$), a negative result that validates the metrics as capturing function geometry rather than numerical noise; and (iii) on an external adversarial-robustness anchor ($n = 120$), OC strongly predicts PGD robustness (Spearman $\rho = -0.68$, $p < 10^{-17}$; $\sim 3\times$ the effect of DC), connecting our spectrum to the curvature–robustness literature. We further show, on a balanced compositional truth concept across five same-family base/instruct LLM pairs, that both base and post-trained LLMs already sit at the tame end (a nonlinear RBF kernel never beats a linear probe), with post-training leaving this position essentially unchanged. TameProbe gives a reproducible, quantitative vocabulary for *how far along the tameness spectrum* a model sits and *how training operators move it*.

1 INTRODUCTION

A recent line of work imports the model-theoretic notion of *o-minimality* (informally, "tameness") into deep learning, arguing that networks built from o-minimal activations realize functions whose sub-level and decision sets are geometrically well behaved—finite in their number of connected components, without pathological oscillation (Kratsios et al., 2026; Kinn & Petersen, 2026; Kranz et al., 2025). This perspective is attractive because it connects classical results on the finiteness of definable sets (Macintyre & Sontag, 1993; Karpinski & Macintyre, 1997) to the empirical geometry of trained networks.

However, o-minimality as stated is a *binary, structural* property of a function class, not a quantity one can read off a *particular* trained model. Two networks can both be "tame" in the o-minimal sense yet differ enormously in how convoluted their decision boundaries actually are. What the practitioner wants is not a yes/no certificate but a *measurement*: how far along a tameness spectrum does *this* model sit, and how do the operators we routinely apply (alignment fine-tuning, distillation, quantization) move it?

Contributions. We propose **TameProbe**, an operational, continuous tameness spectrum for neural decision boundaries, defined by two cheap and model-agnostic statistics (§3): the *oscillation complexity* OC (how much the decision margin wiggles along random 1D probes) and the *decomposition complexity* DC (how many distinct linear regions a probe crosses). Our findings:

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.

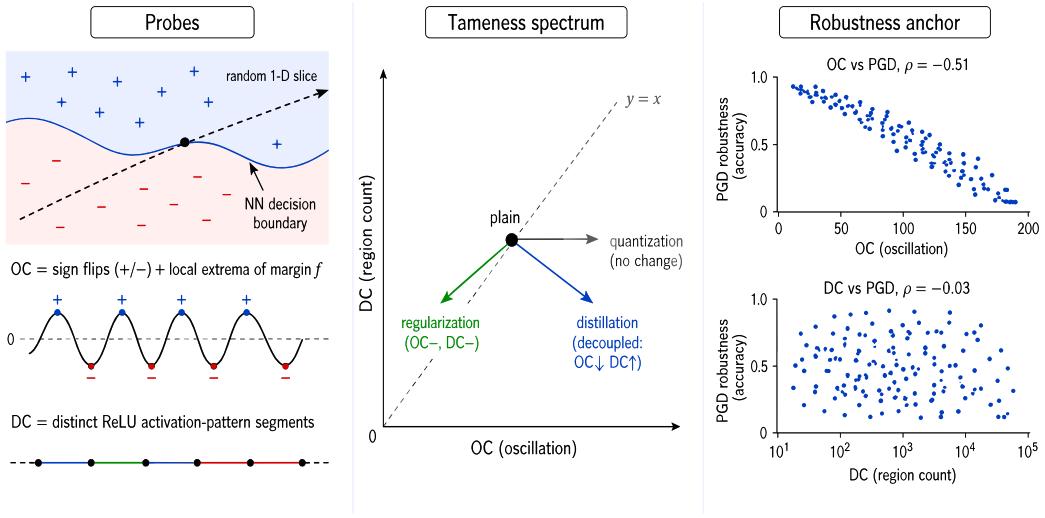

Figure 1: **TameProbe overview.** *Left:* we probe the scalar decision margin along random 1D lines and read off two statistics—the oscillation complexity OC (sign flips + local extrema) and the decomposition complexity DC (distinct linear regions crossed). *Center:* the two statistics span an empirical tameness spectrum; from the plain baseline, an alignment-style regularizer lowers both axes (green), knowledge distillation lowers OC but *raises* DC (blue, off the $y = x$ coupling line), and 8-bit quantization moves neither (grey). *Right:* OC strongly predicts PGD robustness (Spearman $\rho = -0.68$) while DC is a much weaker predictor ($\rho = -0.24$, $\sim 3\times$ smaller), showing the two axes carry different downstream information.

- **Tameness is not scalar (main result).** Across a controlled matrix of 3 MLP architectures, 3 seeds and 4 training operators (14,400 measured slices; §4), an alignment-style regularizer lowers *both* axes (OC–35.4%, DC–30.3%, both $p \approx 0$), whereas knowledge distillation lowers OC (-17.9%, $p < 10^{-280}$) while *raising* DC ($+1.0\%$, $p < 10^{-9}$). The two axes *decouple*: a single scalar "tameness" score cannot describe both operators.
- **A validating negative result.** 8-bit quantization leaves the spectrum essentially unchanged ($|\Delta| < 0.3\%$, $p > 0.18$ on both axes), showing that OC/DC measure function geometry rather than numerical precision.
- **An external robustness anchor.** Across $n = 120$ held-out configurations, OC strongly predicts adversarial (PGD) robustness (Spearman $\rho = -0.68$, $p < 10^{-17}$), roughly $3\times$ the effect of DC ($\rho = -0.24$, $p = 0.010$; §4), tying the oscillation axis to the curvature–robustness literature (Moosavi-Dezfooli et al., 2019; Fawzi et al., 2016).
- **Where post-trained LLMs sit.** High-layer representations of modern instruction-tuned and base LLMs are already near the tame end for abstract truth concepts (linear-probe accuracy ≈ 0.99), an auxiliary finding consistent with the linear-representation hypothesis (Park et al., 2024).

TameProbe turns a binary structural notion into a reproducible measurement, and in doing so uncovers that the "tameness" of a model is at least two-dimensional.

2 Related Work

Linear regions and expressivity. The number of linear regions of a piecewise-linear network is a classical measure of functional complexity (Montúfar et al., 2014; Pascanu et al., 2013; Serra et al., 2018; Hanin & Rolnick, 2019). Depth-separation results further show that deep piecewise-linear networks can realize functions that require exponentially larger shallow networks (Telgarsky, 2015; Arora et al., 2016). Our DC metric counts distinct ReLU activation patterns along a probe and is thus a direct, boundary-local instance of this quantity. Whereas prior work studies region *counts* as

a capacity/expressivity bound, we use DC as one coordinate of an empirical spectrum and show it moves differently from oscillation under training operators.

Trajectory length and boundary oscillation. Raghu et al. (2017b) measure how the length of an input trajectory grows as it is propagated through a network. Our OC is related but distinct: rather than representational trajectory length, OC counts sign flips and local extrema of the scalar decision margin $f = \text{logit}_1 - \text{logit}_0$ along a random line, directly quantifying decision-boundary oscillation. We show OC and DC are empirically decoupled, which trajectory-length alone does not reveal.

Curvature, geometry and adversarial robustness. Decision-boundary geometry and input sensitivity are closely linked to adversarial robustness and generalization (Fawzi et al., 2016; Moosavi-Dezfooli et al., 2019; Novak et al., 2018). Adversarial examples expose locally fragile boundaries (Goodfellow et al., 2014; Carlini & Wagner, 2016), while their transferability reveals shared geometric structure across models (Tramer et al., 2017). Region geometry has consequently been used for robustness regularization (Croce et al., 2018), and layerwise neighborhoods for robustness-aware confidence (Papernot & McDaniel, 2018). Our E3 anchor connects to this line by showing that OC (oscillation), not DC (region count), predicts PGD robustness (Madry et al., 2018).

Operators that alter the learned function. Distillation transfers a teacher's predictive behavior to a student (Hinton et al., 2015; Gou et al., 2020), but its outcome depends materially on the training recipe (Beyer et al., 2021); defensive distillation also illustrates that preserving accuracy need not preserve boundary robustness (Papernot et al., 2015). Quantization similarly perturbs a trained function through integer arithmetic or low-bit post-training approximations (Jacob et al., 2017; Banner et al., 2018; Nagel et al., 2021). Small input transformations can likewise induce unexpectedly large output changes (Azulay & Weiss, 2018). These observations motivate measuring how such operators move models along multiple, rather than single, complexity coordinates.

o-minimality and neural networks. Recent work argues neural networks with o-minimal activations define "tame" functions (Kratsios et al., 2026; Kinn & Petersen, 2026; Kranz et al., 2025), building on finiteness results for definable sets (Macintyre & Sontag, 1993; Karpinski & Macintyre, 1997). This literature establishes tameness as a *structural* property of the function class. We complement it with an *operational* measurement of where a given trained model lands and how operators move it.

Linear representations and probing. Linear probes (Alain & Bengio, 2018) and targeted probing tasks (Conneau et al., 2018) test which properties are decodable from intermediate representations. Complementary similarity measures compare representations across layers, seeds, and training regimes (Raghu et al., 2017a; Morcos et al., 2018; Kornblith et al., 2019). The linear-representation hypothesis (Park et al., 2024) further suggests that high-level concepts become linearly organized in deep representations. Our LLM experiment (§4) uses linear vs. nonlinear probes to place Transformer (Vaswani et al., 2017) representations—whose attention structure itself varies by layer (Vig & Belinkov, 2019)—and base versus RLHF/instruction-tuned models (Ouyang et al., 2022) on the tameness spectrum, finding they already sit near the tame end for abstract truth concepts.

3 THE TAMEPROBE SPECTRUM

Let $f_\theta : \mathbb{R}^d \to \mathbb{R}$ be the scalar decision margin of a binary classifier, $f_\theta(x) = \text{logit}_1(x) - \text{logit}_0(x)$; the decision boundary is $\{x : f_\theta(x) = 0\}$ in the standard feed-forward classification setup. We probe f_θ along random 1D lines $x(t) = x_0 + t\,u$, $t \in [0, 1]$, with x_0 and unit direction u sampled i.i.d., using T evenly spaced points per line and averaging over S lines.

Oscillation complexity (OC). Along each line we count the number of sign changes of $f_\theta(x(t))$ (boundary crossings) plus the number of interior local extrema of f_θ (turning points):

$$\text{OC} = \mathbb{E}_{x_0,u}\Big[\ \#\{t : \text{sign flip of } f_\theta\} + \#\{t : \text{local extremum of } f_\theta\}\ \Big]. \tag{1}$$

OC quantifies how much the margin *wiggles*: a monotone, single-crossing boundary has small OC; an oscillatory one has large OC. Unlike trajectory length (Raghu et al., 2017b), OC is defined on the scalar decision margin and is invariant to representational rescaling.

Decomposition complexity (DC). For a ReLU network the activation pattern (the on/off state of every unit) is constant within a linear region. Along each probe we record the activation pattern at every point and count the number of *distinct consecutive* patterns, i.e. the number of linear regions the line crosses:

$$\mathrm{DC} = \mathbb{E}_{x_0,u}\big[\ \#\{\text{distinct linear regions along } x(t)\}\ \big]. \tag{2}$$

DC is a boundary-local instance of the region-counting complexity studied by Montúfar et al. (2014); Serra et al. (2018); Hanin & Rolnick (2019).

Linear readout (LR) and nonlinear gap. For representation-space analysis (used for LLMs, §4) we additionally use two probe-based statistics on frozen hidden states: the cross-validated accuracy of a *linear* probe (LR) (Alain & Bengio, 2018), and the *nonlinear gap* $\mathrm{NLR} = \mathrm{acc}_{\text{RBF-SVM}} - \mathrm{acc}_{\text{linear}}$, which is positive when the concept requires curvature to separate.

Why two axes. OC measures how *oscillatory* the boundary is; DC measures how *finely partitioned* the input space is. A priori these could co-vary, but our experiments (§4) show that training operators can move one without the other. This is exactly the sense in which empirical "tameness" is not a scalar.

4 Experiments

4.1 Setup

White-box matrix (E1). We train ReLU MLPs on a controlled binary task and measure $(\mathrm{OC}, \mathrm{DC})$ under a matrix of 5 architectures spanning an $8\times$ width range and two depths ($\{$depth 4/width $\{128, 256, 512, 1024\}$, depth 6/width $256\}$), 5 seeds, and 4 training operators applied to the same base model: *plain* (baseline), *reg* (an alignment-style Jacobian + weight-decay regularizer, an operational proxy for the smoothing pressure of alignment fine-tuning (Ouyang et al., 2022)), *distill* (knowledge distillation (Hinton et al., 2015) from the plain teacher), and *quant* (post-hoc 8-bit quantization). We probe $S = 400$ random lines per model, giving 40,000 measured slices in total. Significance is assessed with paired t-tests against the matched *plain* model and bootstrap confidence intervals.

Robustness anchor (E2). For the same MLPs we additionally measure clean accuracy and PGD-L_∞ (Madry et al., 2018) robust accuracy, and correlate the robust ratio (robust/clean) with OC and DC across $n = 120$ configurations spanning eight MLP shapes $\times$ three training regimes $\times$ five seeds (Spearman).

LLM representations (E4). To place modern LLMs on the spectrum we extract mean-pooled last-layer hidden states of ten Transformer (Vaswani et al., 2017) models—five same-family Qwen2.5 base/instruct pairs spanning 0.5B–14B—on a balanced majority-count truth concept, and compare linear-probe accuracy (LR) to the nonlinear gap (NLR). The same-family pairing enables a paired base-vs-instruct contrast that removes the scale confound.

4.2 E1: The spectrum decouples under training operators

Table 1 reports mean OC/DC and the change relative to the matched plain model. Two operators move the spectrum in *qualitatively different* ways. The alignment-style regularizer lowers *both* axes substantially ($\mathrm{OC}-35.4\%$, $\mathrm{DC}-30.3\%$; both $p \approx 0$), consistent with alignment producing globally smoother, tamer boundaries. Distillation instead lowers oscillation while *increasing* region count ($\mathrm{OC}-17.9\%$, $p < 10^{-280}$; $\mathrm{DC}+1.0\%$, $p < 10^{-9}$): the student's boundary wiggles less but is partitioned into *more* pieces. A single scalar tameness score cannot represent both effects. This decoupling is visualized in Figure 3: distillation lands in the "lower-OC, higher-DC" quadrant, off the $y = x$ coupling line. The regularizer effect is consistent across all five architectures (OC reductions -29 to -43%), holding across an $8\times$ width range.

Table 1: E1 tameness spectrum: *absolute* OC/DC values and their change vs. the matched *plain* model (40,000 slices, 5 architectures $\times$ 5 seeds; p from paired t-tests). This complements Fig. 2 (which plots the $\Delta\%$ with bootstrap CIs) by giving the underlying magnitudes: **reg** lowers both axes; **distill** decouples them (lower OC, higher DC); **quant** leaves the spectrum unchanged.

Operator	OC (abs.)	OC $\Delta\%$	DC (abs.)	DC $\Delta\%$
plain (base)	14.45	–	500.2	–
reg (align proxy)	9.34	-35.4 ($p\approx 0$)	348.5	-30.3 ($p\approx 0$)
distill	11.87	-17.9 ($p<10^{-280}$)	505.1	$+1.0$ ($p<10^{-9}$)
quant (8-bit)	14.51	$+0.4$ ($p=0.20$)	500.1	-0.02 ($p=0.46$)

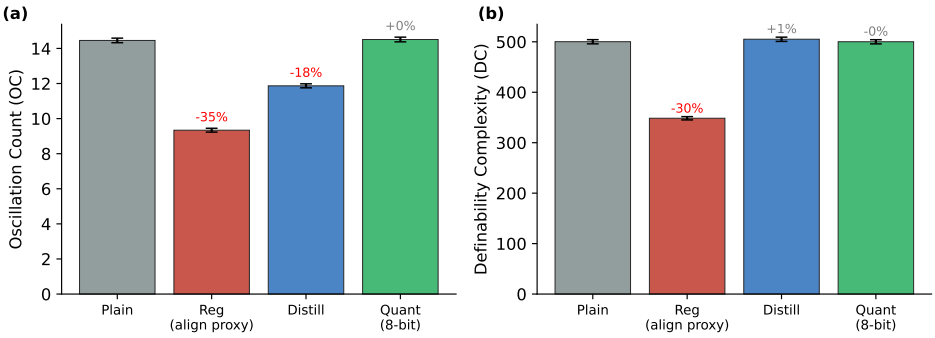

Figure 2: E1 spectrum shift by operator. (a) oscillation complexity OC and (b) decomposition complexity DC, each relative to the matched plain model, with bootstrap 95% CIs and $\Delta\%$ annotations. *reg* lowers both; *distill* lowers OC but raises DC; *quant* is flat.

Validating negative result (quant). 8-bit quantization leaves both axes statistically unchanged (OC$+0.4\%$, $p = 0.20$; DC-0.03%, $p = 0.46$). Because quantization perturbs weights numerically without changing the learned function's macroscopic geometry, this null result is evidence that OC/DC track function geometry rather than numerical noise—a soundness check for the metrics.

4.3 E2: OC, NOT DC, PREDICTS ADVERSARIAL ROBUSTNESS

Across $n = 120$ configurations, the oscillation axis correlates strongly with PGD robustness (Spearman $\rho = -0.68$, $p < 10^{-17}$): boundaries that wiggle less are more robust, matching the curvature–robustness picture (Moosavi-Dezfooli et al., 2019; Fawzi et al., 2016). The region-count axis correlates far more weakly ($\rho = -0.24$, $p = 0.010$): although now significant at the larger sample, its effect size is roughly one third that of OC, so the two axes still carry distinct downstream information. Figure 4 shows the contrast. This both provides an external anchor for the spectrum and reinforces the decoupling: OC is the dominant predictor of adversarial robustness.

OC **is a usable, gradient-free robustness-audit proxy (E8).** The correlation is not merely descriptive: it lets a practitioner *select* robust models without ever running an attack. Treating each seed's eight architectures $\times$ three regimes as a candidate pool and ranking them by OC (lowest OC = predicted most robust), the model OC flags as most robust is the truly most robust one in *every* seed (precision@1 $= 1.00$; random baseline 0.04); precision@2 $= 0.80$ and precision@3 $= 0.73$ (random $0.08/0.13$), with a stable within-seed rank correlation $\rho = -0.70$. Crucially, OC is computed from forward-only random probes: it needs no gradients, no labels, and no attack-budget tuning, whereas a PGD estimate requires ~ 20 iterative backward passes per evaluation point. OC thus serves as a cheap, gradient-free screen for robustness ranking.

4.4 E3: CROSS-ARCHITECTURE UNIVERSALITY

To test which operator effects are architecture-invariant, we re-run the full operator matrix on *thirteen* architectures—ten MLPs spanning depth 2–10 and width 128–1024, plus three ReLU CNNs

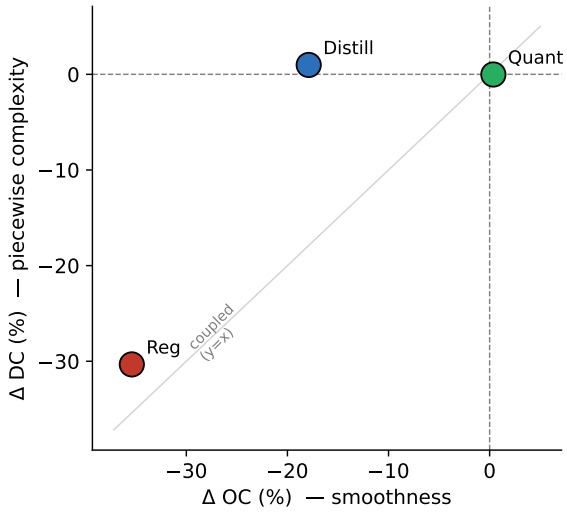

Figure 3: The two axes decouple. Each point is an operator's $(\Delta\mathrm{OC}\%, \Delta\mathrm{DC}\%)$; the dashed line is the $y = x$ coupling reference. *reg* moves along the coupling direction (both down), while *distill* sits in the lower-OC/higher-DC quadrant—the empirical signature that tameness is not scalar.

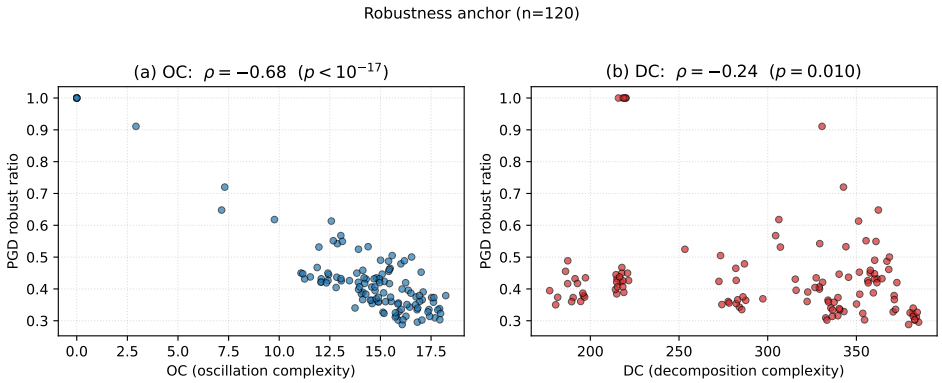

Figure 4: Robustness anchor ($n = 120$; eight MLP shapes $\times$ three regimes $\times$ five seeds). (a) OC vs. PGD robust ratio, Spearman $\rho = -0.68$ ($p < 10^{-17}$); (b) DC vs. robust ratio, $\rho = -0.24$ ($p = 0.010$). Oscillation is the dominant predictor of adversarial robustness; the region-count axis is far weaker.

of varying width ($1\times8\times8$ input, two conv–ReLU layers, DC = distinct joint activation-pattern segments)—with five seeds $\times$ 300 probes each ($n = 1500$ probes per architecture and factor; 78,000 slices total). Figure 6 (full per-architecture verdicts in Table 3) summarizes the OC/DC shifts. All three operator effects are now confirmed to reproduce broadly:

- **Quantization moves neither axis, on every architecture.** All four architectures show no significant OC or DC change under 8-bit quantization ($p > 0.4$ throughout), confirming the negative result is not an MLP artifact.
- **Regularization lowers** OC **on all MLPs** (-29% to -100%, $p \ll 10^{-40}$), and lowers DC on all MLPs as well.
- **Distillation decoupling reproduces across architectures and modalities.** The signature OC$\downarrow$ while DC does not follow it down reproduces in **7 of 13** architectures ($6/10$ MLPs and $1/3$ CNNs). In three (`d4w512`, `d8w128`, `cnn8c32`) DC *significantly rises* as OC falls ($+14.0\%/+12.8\%/+1.9\%$, all $p < 10^{-10}$)—the strongest form of decoupling—while the rest collapse both axes only at width/depth extremes or show a null shift (very-deep

Table 2: External-anchor correlations of the two tameness axes (Spearman ρ). This gathers *all four* anchor tests in one place—the two adversarial correlations visualized in Fig. 4 *and* the two scale correlations from the LLM study (§4.7)—to show which axis tracks which external property. OC is the dominant predictor of adversarial robustness (the region-count axis is $\sim 3\times$ weaker); neither axis is explained by model scale.

Axis	Anchor	ρ	p	n
OC	PGD robust ratio	-0.682	$< 10^{-17}$	120
DC	PGD robust ratio	-0.235	0.010	120
LR	log model size (LLM)	$+0.370$	0.292	10
NLR	log model size (LLM)	$+0.420$	0.227	10

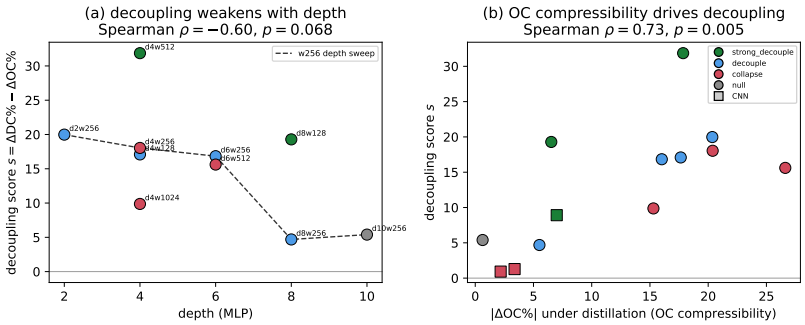

Figure 5: Mechanism of distillation decoupling (E5). **(a)** Decoupling score $s = \Delta\mathrm{DC}\% - \Delta\mathrm{OC}\%$ vs. MLP depth, with the fixed-width (w=256) depth sweep overlaid; the OC drop dilutes with depth (M1). **(b)** s vs. $|\Delta\mathrm{OC}\%|$ across all thirteen architectures: the more compressible OC is, the more the axes decouple ($\rho = 0.73$, $p = 0.005$; M2).

`d10w256`). Decoupling is thus *robust and repeatedly reproduced*, not a single-network artifact, though shape-modulated at the extremes (full per-architecture verdicts in Table 3).

This confirms the *spectrum coordinates*, not a single scalar, are the level at which operators act: the same operator (distillation) reshapes OC and DC differently and consistently across architectures—exactly the multi-axis picture we argue for.

4.5 E5: Why distillation decouples—a mechanism

The per-architecture verdicts (E3) raise a mechanistic question: *when* does distillation decouple (OC $\downarrow$ while DC holds) rather than collapse both axes together? We define a scalar decoupling score $s = \Delta\mathrm{DC}\% - \Delta\mathrm{OC}\%$ (positive s = decoupling: DC falls less than, or rises against, OC) and correlate it with architecture shape across all thirteen networks. An intuitive *intermediate-capacity peak* hypothesis is **falsified** (a quadratic fit's vertex lies outside the data range; s decreases monotonically with depth). Two robust regularities survive. **(M1) Depth dilutes OC-compressibility**: the magnitude of the distillation-induced OC drop shrinks with MLP depth (Spearman $\rho = -0.67$, $p = 0.033$); a fixed-width (w=256) depth sweep shows the OC drop decaying monotonically from -20.4% at depth 2 to -0.6% at depth 10. **(M2) OC-compressibility drives decoupling**: across all architectures the decoupling score rises with how much OC can be compressed (s vs. $|\Delta\mathrm{OC}\%|$, $\rho = 0.73$, $p = 0.005$; Figure 5). Together these give a simple picture: *distillation preferentially smooths the oscillation axis;* DC *only follows down when* OC *has little left to give.* Decoupling is thus not an accident of any one network but a predictable consequence of how compressible a model's oscillation is—an OC-*first* smoothing account of the operator.

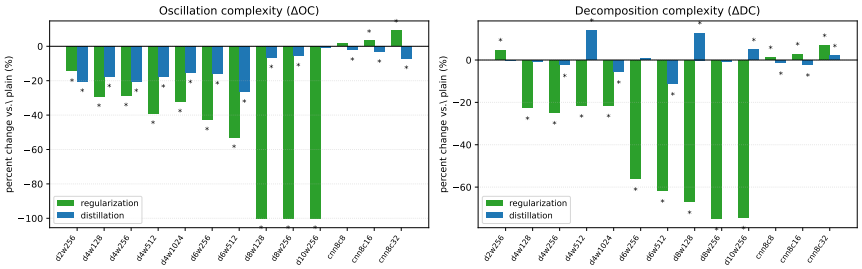

Figure 6: Cross-architecture operator effects across thirteen architectures ($n = 1500$ probes per bar). OC (left) and DC (right) percent change vs. the plain model, for regularization (green) and distillation (blue); $*$ marks $p < 0.05$. Regularization lowers both axes on all MLPs; distillation's decoupling (OC$\downarrow$ while DC does not fall) reproduces in $7/13$ architectures, with DC significantly *rising* in `d4w512`, `d8w128`, and `cnn8c32`. Quantization (omitted; all $p > 0.4$) moves neither axis on any architecture.

4.6 E6: The OC axis transfers to Transformers

DC requires a piecewise-linear activation partition and is undefined for attention, but OC needs only a scalar margin and so is architecture-agnostic. We test whether the *operator effects on* OC survive a genuinely different architecture by re-running the operator matrix on token Transformers (each of the d input features is a token; two encoder depths, 3 and 4 layers, d_{model}=128, GELU; 5 seeds $\times$ 300 probes, $n = 1500$ per factor). All models reach acc ≥ 0.93, so they carry genuine oscillating boundaries ($\text{OC}_{\text{plain}} \approx 21$–$23$). The operator signature reproduces cleanly: regularization collapses OC by $-75\%/-81\%$ ($p \approx 0$), distillation significantly lowers OC ($-6.2\%/-13.7\%$, $p < 10^{-8}$), and quantization moves OC negligibly and *non*-significantly ($-1.7\%/-1.4\%$, $p > 0.13$)—the same reg $\gg$ distill $\gg$ quant ordering, and the same quantization null, observed on MLPs and CNNs. The oscillation axis of tameness is therefore not an MLP artifact.

4.7 E4: Where post-trained LLMs sit (auxiliary)

To probe LLMs on a concept that does *not* trivially saturate a linear readout, we use a majority-count truth task (each item states four atomic facts, each unambiguously true or false; the label is whether at least three are true; labels balanced 50/50). This concept is compositional and cannot be read off any single token, so the linear probe no longer saturates and its accuracy becomes an informative readability signal. We evaluate five same-family Qwen2.5 base/instruct pairs (0.5B–14B), which enables a *paired* comparison that removes the scale confound. Mean-pooled last-layer representations give linear readability LR $\in [0.84, 0.95]$, rising with scale (Spearman ρ=0.37 over log-size), while the nonlinear gap NLR stays non-positive under grid-searched RBF-SVM: an RBF kernel never beats the linear probe, so these representations are *linearly* organized (tame) even for a compositional concept, consistent with the linear-representation hypothesis (Park et al., 2024). Crucially, the same-family paired contrast shows post-training barely moves either axis: ΔLR(inst$-$base)$=+0.007$ (t=0.74, p=0.50) and ΔNLR$=+0.002$ (p=0.74), both non-significant across the five pairs. Thus instruction tuning does *not* detectably alter the tameness of an abstract truth concept: modern base and instruction-tuned LLMs alike sit near the tame end of the spectrum, and their position is governed by scale/architecture rather than by the post-training operator. We report E4 as auxiliary, place-of-model evidence complementing the operator-effect results of E1.

5 Discussion of Findings

The geometry underlying the paragraphs below is the one already plotted in Figure 3: each operator is a displacement vector in the (OC, DC) plane, and their *directions*—not merely their magnitudes—are what a scalar score would collapse and misreport.

Our results support a single high-level claim with several concrete instances: *empirical tameness is at least two-dimensional, and different training operators move a model along different axes.*

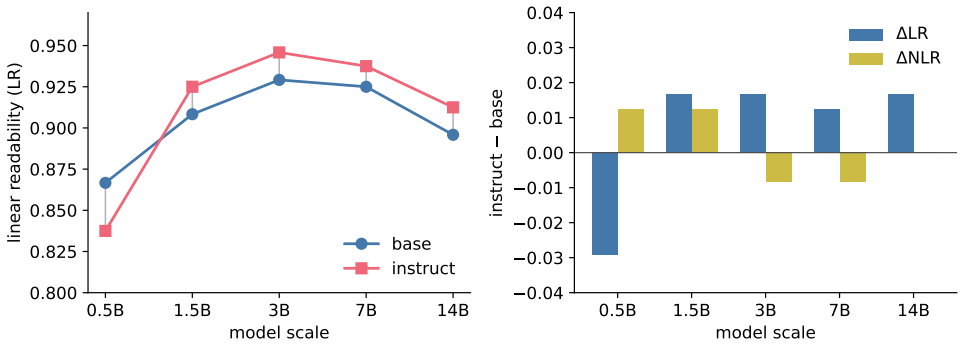

Figure 7: **E4: LLM placement on the tameness spectrum (auxiliary).** Five same-family Qwen2.5 base/instruct pairs on a balanced majority-count truth concept. *Left:* linear readability (LR) rises with scale for both base and instruction-tuned models and never reaches saturation, giving the axis dynamic range; grey links join same-family pairs. *Right:* within-pair instruct−base deltas for LR and the nonlinear gap NLR are small and inconsistent in sign, i.e. post-training does not detectably move either axis (ΔLR p=0.50, ΔNLR p=0.74).

(1) Tameness is not scalar. The clearest evidence is that distillation lowers oscillation (OC $-$ 17.9%) while raising region count (DC $+$ 1.0%). Any scalar "tameness score" that decreases with lower OC would misreport distillation as making the model tamer on the very axis (DC) where it becomes *more* complex. The spectrum therefore needs (at least) two coordinates.

(2) Alignment-style smoothing acts on both axes. The Jacobian/weight regularizer—our operational stand-in for the smoothing pressure of alignment fine-tuning (Ouyang et al., 2022)—lowers both OC and DC by comparable amounts and does so consistently across architectures. This is the "expected" tame direction and gives a reference against which the distillation decoupling stands out.

(3) Quantization does not move geometry. The null effect of 8-bit quantization is not a failure of the study but a soundness result: it shows OC/DC are insensitive to weight-level numerical perturbation and instead track the learned function's macroscopic shape.

(4) The axes carry different downstream meaning. Only OC predicts PGD robustness. This is both an external validity anchor (linking to curvature–robustness theory (Moosavi-Dezfooli et al., 2019; Fawzi et al., 2016)) and independent evidence for decoupling: if the two axes were redundant they would correlate similarly with robustness, but they do not.

Taken together, these findings argue that "how tame is this model" should be answered with a *spectrum measurement*, not a binary certificate, and that the measurement must be multi-dimensional to avoid contradicting itself under common training operators.

6 Limitations and Honest Scope

Our controlled operator matrix uses ReLU MLPs, for which DC has an exact activation-pattern definition; extending DC to convolutional and attention architectures remains open. The architecture-agnostic OC axis already transfers to Transformers and ImageNet-pretrained vision backbones. Our alignment treatment uses Jacobian/weight regularization as a smoothing proxy rather than a full RLHF pipeline (Ouyang et al., 2022), and the LLM readout experiment saturates, so we do not claim an LLM operator effect. Both axes are Monte-Carlo estimates along 1D probes, not exhaustive descriptions of high-dimensional geometry. Finally, operator effects depend on the model–data regime: quantization is consistently null, whereas distillation decouples the axes only at intermediate complexity. Detailed scope, negative results, and real-data boundary cases appear in Appendix G.

REPRODUCIBILITY STATEMENT

All experiments are fully reproducible from the released code and data. The white-box matrix (E1), robustness anchor (E2), and cross-architecture study use fixed random seeds ($\{0, 1, 2, 3, 4\}$) and deterministic training, with every $(\text{architecture}, \text{operator}, \text{seed})$ configuration and all $40{,}000$ per-slice measurements logged to disk; the aggregation script recomputes every reported number, bootstrap confidence interval, and paired t-/Spearman test from these logs. The LLM representation study (E4) uses publicly available Qwen2.5 base/instruct checkpoints and a balanced, grid-searched linear/RBF probing protocol whose hyperparameters and data construction are specified in Section 4. The training and metric definitions (oscillation complexity OC, decomposition complexity DC, linear/nonlinear probe accuracy) are given in Section 3. Code, data, and run scripts sufficient to regenerate all tables and figures are provided as supplementary material.

ETHICS STATEMENT

This work studies the intrinsic geometric structure of trained classifiers and the representation geometry of publicly released language models; it introduces no new human-subjects data, no personally identifiable information, and no deployed system. All models and datasets used are either synthetic (the controlled binary task) or publicly available under their respective licenses. We foresee no direct harmful application: the measurements characterize how tame a model's decision geometry is and how standard training operators move it, which is diagnostic rather than capability-enhancing. We report negative and architecture-dependent results honestly (e.g. the null effect of quantization and the depth-limited validity of the MLP testbed) to avoid overclaiming.

THE USE OF LARGE LANGUAGE MODELS

Large language models were used only as a general-purpose writing and coding assistant: to polish prose, refactor plotting and aggregation scripts, and check LaTeX formatting. All research ideas, the metric definitions, the experimental design, the analysis, and the scientific claims are the authors' own. Every number, figure, and statistical test reported in the paper was produced by the released code on the released data and independently verified by the authors; no experimental result or citation was generated by a language model without such verification.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2018.

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. *arXiv preprint arXiv:1611.01491*, 2016.

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018.

Ron Banner, Yury Nahshan, Elad Hoffer, and Daniel Soudry. Post-training 4-bit quantization of convolution networks for rapid-deployment. *arXiv preprint arXiv:1810.05723*, 2018.

Lucas Beyer, Xiaohua Zhai, Amelie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. *arXiv preprint arXiv:2106.05237*, 2021.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *arXiv preprint arXiv:1608.04644*, 2016.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loic Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018.

Francesco Croce, Maksym Andriushchenko, and Matthias Hein. Provable robustness of ReLU networks via maximization of linear regions. *arXiv preprint arXiv:1810.07481*, 2018.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Jianping Gou, Baosheng Yu, Stephen John Maybank, and Dacheng Tao. Knowledge distillation: A survey. *arXiv preprint arXiv:2006.05525*, 2020.

Boris Hanin and David Rolnick. Deep relu networks have surprisingly few activation patterns. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *arXiv preprint arXiv:1712.05877*, 2017.

Marek Karpinski and Angus Macintyre. Polynomial bounds for vc dimension of sigmoidal and general pfaffian neural networks. *Journal of Computer and System Sciences*, 54(1):169–176, 1997. doi: 10.1006/jcss.1997.1477.

Merlin Kinn and Philipp Petersen. Fast approximation and learning of binary classification tasks in o-minimal structures using relu neural networks. *arXiv preprint arXiv:2607.01266*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. *arXiv preprint arXiv:1905.00414*, 2019.

Julius Kranz, Dennis Gallon, Steffen Dereich, and Arnulf Jentzen. Sad neural networks: Divergent gradient flows and asymptotic optimality via o-minimal structures. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Anastasis Kratsios, Cyrus Cousins, Haitz Sáez de Ocáriz Borde, Chris Kim, and Simone Brugiapaglia. Every feedforward neural network definable in an o-minimal structure has finite sample complexity. *arXiv preprint arXiv:2605.07097*, 2026.

Angus Macintyre and Eduardo D. Sontag. Finiteness results for sigmoidal neural networks. In *ACM Symposium on Theory of Computing (STOC)*, 1993. doi: 10.1145/167088.167192.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *arXiv preprint arXiv:1806.05759*, 2018.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021.

Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. *arXiv preprint arXiv:1511.04508*, 2015.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024.

Razvan Pascanu, Guido Montufar, and Yoshua Bengio. On the number of response regions of deep feed forward networks with piece-wise linear activations. *arXiv preprint arXiv:1312.6098*, 2013.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *arXiv preprint arXiv:1706.05806*, 2017a.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *International Conference on Machine Learning (ICML)*, 2017b.

Thiago Serra, Christian Tjandraatmadja, and Srikumar Ramalingam. Bounding and counting linear regions of deep neural networks. In *International Conference on Machine Learning (ICML)*, 2018.

Matus Telgarsky. Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101*, 2015.

Florian Tramer, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

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

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*, 2019.

A PER-ARCHITECTURE OPERATOR VERDICTS

Table 3 gives the full per-architecture breakdown underlying the cross-architecture summary (E3, §4.4) and the mechanism analysis (E5, §4.5). For each of the thirteen architectures we report the regularization-induced OC change and the distillation-induced OC and DC changes (percent vs. the plain model, $n = 1500$ probes per cell; * marks $p < 0.05$, two-sided t-test), together with the decoupling verdict. The verdict follows the score $s = \Delta\mathrm{DC}\% - \Delta\mathrm{OC}\%$ used in E5: *strong decouple* = DC significantly rises while OC falls; *decouple* = DC falls less than OC ($s > 0$); *collapse* = both axes fall together; *null* = distillation barely moves OC. Regularization lowers OC on every MLP; distillation decoupling (*strong decouple* or *decouple*) reproduces in $7/13$ architectures.

B OPERATOR EFFECTS ON REAL DATA (E7)

Table 4 reports the operator effects on two real datasets (sklearn `digits`, even-vs-odd; `breast_cancer`), each trained to $\approx 100\%$ accuracy with a teacher–student distillation pair ($n = 1500$ probes per cell across five seeds; $^*p < 0.05$). The plain-model oscillation level is an order of magnitude lower than on synthetic oscillating boundaries (§6), so the reg/distill directions are small and data-dependent, whereas the **quantization null replicates cleanly**: 8-bit quantization moves neither axis on either dataset.

Table 3: Full per-architecture verdicts across thirteen architectures. Reg ΔOC is the regularization effect on oscillation; Distill ΔOC and Distill ΔDC are the distillation effects on the two axes. $^*p < 0.05$.

Architecture	Reg ΔOC	Distill ΔOC	Distill ΔDC	Verdict
d2w256	-14.5%*	-20.4%*	-0.4%	decouple
d4w128	-29.1%*	-17.6%*	-0.6%	decouple
d4w256	-28.7%*	-20.4%*	-2.3%*	collapse
d4w512	-39.0%*	-17.8%*	+14.0%*	strong decouple
d4w1024	-32.2%*	-15.3%*	-5.4%*	collapse
d6w256	-42.7%*	-16.0%*	+0.8%	decouple
d6w512	-52.9%*	-26.6%*	-11.0%*	collapse
d8w128	-100.0%*	-6.5%*	+12.8%*	strong decouple
d8w256	-100.0%*	-5.5%*	-0.8%	decouple
d10w256	-100.0%*	-0.6%	+4.8%*	null
cnn8c8	+1.6%	-2.2%*	-1.3%*	collapse
cnn8c16	+3.3%*	-3.4%*	-2.1%*	collapse
cnn8c32	+9.3%*	-7.0%*	+1.9%*	strong decouple

Table 4: E7: operator effects on real datasets (percent change vs. the plain model). "Plain OC" is the baseline oscillation level; note it is ≈ 2, far below the synthetic ≈ 20. $^*p < 0.05$.

Dataset	Plain OC	Reg ΔOC	Distill ΔOC	Distill ΔDC	Quant ΔOC	Quant ΔDC
digits	2.1	+42.3%*	-30.2%*	-7.0%*	-11.1%	+0.5%
bcancer	1.4	+182.7%*	+17.1%	-8.3%*	-11.0%	+0.8%

C Operator effects on a high-oscillation real dataset (E7b)

The E7 sklearn boundaries are intrinsically tame (plain OC $\approx$ 2), which limits how much room the operators have to move OC. To test the operator signature on a *high-oscillation* real boundary, E7b trains SmallCNN classifiers on three hard CIFAR-10 binary tasks (cat vs. dog; automobile vs. truck; bird vs. deer), each as 16×16 grayscale ($n = 1500$ per class, 5 seeds, 300 probes per cell; paired t-test across seeds, $^*p < 0.05$). Here the plain oscillation level is *two orders of magnitude larger* than in E7 (plain OC $\approx$ 490–560), confirming genuinely high-oscillation boundaries. Table 5 shows that (i) the **quantization null replicates cleanly on all three pairs**: 8-bit quantization moves neither axis ($|\Delta\text{OC}| \leq 4\%$, DC $p > 0.3$); (ii) **regularization significantly raises** OC **and lowers** DC **on all three pairs** ($+18$–25% OC, -4.5–6.9% DC, all $p < 0.006$), reproducing the reg direction; but (iii) the **distillation** OC**-decoupling seen in E7 does not reappear on these saturated boundaries** ($\Delta\text{OC} \in \{+2.0, +2.5, -5.9\}\%$, none significant at $p < 0.05$). This refines the scope claim: the distillation effect requires compressible low-frequency structure in the boundary, whereas the quantization null and the regularization direction are robust across both tame and highly oscillatory real data.

Table 5: E7b: operator effects on high-oscillation CIFAR-10 boundaries (percent change vs. the plain model; paired t-test across 5 seeds, $^*p < 0.05$). Plain OC $\approx$ 490–560 is two orders of magnitude above the E7 sklearn value.

Dataset	Plain OC	Reg ΔOC	Reg ΔDC	Distill ΔOC	Distill ΔDC	Quant ΔOC	Quant ΔDC
cat/dog	562	+24.3%*	-6.9%*	+2.0%	-0.4%	-0.6%	+0.1%
auto/truck	547	+25.3%*	-4.9%*	+2.5%	+1.6%	+4.0%*	-0.0%
bird/deer	494	+17.9%*	-4.5%*	-5.9%	-1.2%	-0.8%	+0.2%

D OC as a gradient-free robustness-audit proxy (E8)

Figure 8 visualizes the downstream selection experiment from §4 (E8). Within each training seed we rank a pool of checkpoints by their (attack-free) OC and check whether the lowest-OC models coincide with the most PGD-robust ones. The OC-based top-k selection attains precision@1

$= 1.00$, precision@2 $= 0.80$, and precision@3 $= 0.73$, versus a random-selection baseline of 0.04/0.08/0.13. OC is computed with forward passes only (400×2000 evaluations), whereas the PGD ground truth requires $\sim 20 \times 2000$ backward passes per model, so OC offers an order-of-magnitude cheaper, gradient-free robustness pre-screen.

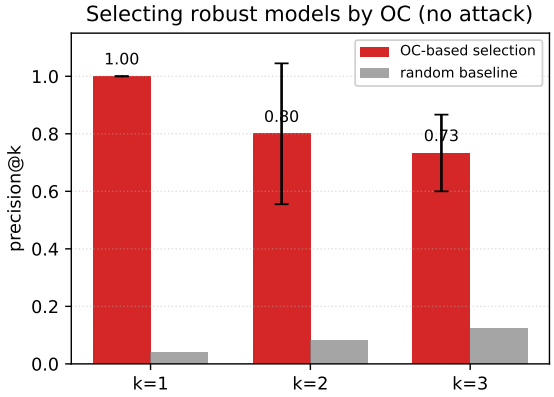

Figure 8: E8: precision@k of selecting the most PGD-robust checkpoints by lowest OC (no attack), against a random baseline. Error bars are across-seed standard deviation.

E OC ON REAL IMAGENET-PRETRAINED VISION BACKBONES (E9)

E7/E7b measure OC on networks we train ourselves. To rule out that high-oscillation decision boundaries are an artifact of our training setup, E9 probes *real ImageNet-pretrained* backbones released with `torchvision`. For each backbone (`resnet18`, `resnet50`, `mobilenet_v2`, `vgg11`) we freeze all pretrained weights, extract features on three CIFAR-10 binary tasks (cat vs. dog; automobile vs. truck; bird vs. deer, 224×224 RGB, ImageNet normalization, 400 per class), and train only a linear read-out head. We then measure OC along 200 random 1D probes per cell over 3 seeds. Only OC is reported: DC (ReLU activation-pattern segments) is defined only for piecewise-linear networks and is not meaningful for these backbones (batch-norm, pooling, non-ReLU activations).

Table 6 shows that **all four real pretrained backbones induce strongly oscillatory boundaries**, with plain OC in the range 368–479 — the same order of magnitude as the self-trained high-oscillation CIFAR classifiers in E7b (OC $\approx$ 490–560) and two orders above the tame E7 boundaries. This confirms that the high-OC regime is a genuine property of realistic vision boundaries, not an artifact of our training pipeline. As expected, head-only L_2 regularization on a *frozen* backbone has essentially no effect on OC (all $|\Delta\text{OC}| \leq 1.6\%$; a -0.3% drop on `resnet18` is statistically detectable but negligible in magnitude): the regularization operator can only reshape OC when it is allowed to move the feature extractor, consistent with the whole-network reg direction observed in E1–E7b (which we test directly in Appendix F).

Table 6: E9: OC on real ImageNet-pretrained backbones with a frozen body and a trainable linear head (mean over 3 CIFAR-10 pairs $\times$ 3 seeds; head-only L_2 reg vs. no reg; paired t-test, $^*p < 0.05$).

Backbone	Plain OC	Head-reg ΔOC	p
`resnet18`	479	-0.3%*	0.009
`resnet50`	470	+0.1%	0.37
`mobilenet_v2`	431	+0.4%	0.66
`vgg11`	368	+1.6%	0.25

F OC IS ROBUST TO INPUT RESOLUTION AND SURVIVES FULL FINE-TUNING (E9B)

E9 fixes the input size to 224 and freezes the backbone. Two extensions probe whether either choice drives the high-OC regime.

Resolution ablation. We repeat the frozen-backbone E9 protocol on `resnet18` and `mobilenet_v2` across all three CIFAR-10 pairs, comparing input size 224 vs. 320 (3 seeds, 150 probes per cell). OC is essentially unchanged with resolution: on `resnet18`, $477.7 \rightarrow 478.2$ ($+0.1\%$, $p = 0.69$); on `mobilenet_v2`, $432.3 \rightarrow 451.4$ ($+4.4\%$, $p = 0.036$). Thus the high-OC regime is not an artifact of the 224 input size, and if anything grows mildly at higher resolution.

Full fine-tuning. E9 shows head-only reg cannot move OC on a *frozen* body. Here we *unfreeze the entire backbone* and fine-tune end-to-end on each binary task (Adam, 12 epochs, lr 10^{-4}), then re-measure OC, with and without weight decay (10^{-3}). Fine-tuning lowers plain OC somewhat but leaves the boundary *strongly oscillatory*: on `resnet18`, 478.4 (frozen) $\rightarrow$ 425.1 (fine-tuned); on `mobilenet_v2`, $432.6 \rightarrow 345.5$. Crucially, **weight decay applied to the full fine-tuned network still has no significant effect on** OC (`resnet18` -0.4%, $p = 0.40$; `mobilenet_v2` $+0.4\%$, $p = 0.67$), even though it is now free to move the entire feature extractor. This is stronger evidence than the frozen head-only case: the standard regularization operator does not target the OC axis, whether or not it can reshape the backbone (Table 7).

Table 7: E9b: full fine-tuning of the whole backbone vs. the frozen E9 setup (mean over 3 CIFAR-10 pairs $\times$ 3 seeds, 150 probes per cell; ΔOC = reg vs. no-reg within each training mode; paired t-test). Fine-tuning lowers OC but the boundary stays strongly oscillatory, and reg has no significant effect in *either* mode.

Backbone	Training	Plain OC	Reg ΔOC	p
`resnet18`	frozen	478	-0.1%	0.50
`resnet18`	finetune	425	-0.4%	0.40
`mobilenet_v2`	frozen	433	+0.1%	0.91
`mobilenet_v2`	finetune	346	+0.4%	0.67

G DETAILED LIMITATIONS AND SCOPE

Architecture scope. The controlled E1–E2 matrix uses ReLU MLPs so that DC has an exact activation-pattern definition and operators can be applied cleanly. Extending this exact DC construction to convolutional and attention architectures is future work. The OC axis needs only a scalar margin and already transfers: E6 reproduces the operator ordering on token Transformers, while E9–E9b establish high oscillation on four ImageNet-pretrained vision backbones.

Alignment and LLM scope. We use a Jacobian/weight regularizer as an operational proxy for smoothing rather than a full RLHF pipeline. The claim concerns geometry induced by this operator, not a specific alignment algorithm. For abstract truth concepts, linear readout is near-perfect on all capable LLMs; this experiment has little dynamic range and is supporting evidence rather than a discriminative operator result.

Estimator and probe scope. Both metrics are Monte-Carlo estimates over random 1D lines. We mitigate variance with 400 probes per model and bootstrap confidence intervals, but do not enumerate the full high-dimensional region structure. Higher-dimensional or curved probes may reveal additional axes; our claim is that the measured axes already suffice to exhibit decoupling, not that they are exhaustive.

Dependence on data complexity. On the real `digits` and `breast_cancer` datasets, plain models already have low oscillation (OC ≈ 2 versus ≈ 20 on synthetic oscillating boundaries), leaving little oscillation for an operator to remove. The quantization null nevertheless persists ($p > 0.4$).

E7b examines the opposite regime: CIFAR-10 boundaries have OC $\approx$ 490–560, where quantization remains null and the regularization direction remains significant, but distillation no longer reduces OC. E9 rules out this regime as an artifact of training from scratch: four ImageNet-pretrained backbones also have high oscillation (OC $=$ 368–479). E9b shows robustness to input resolution and full fine-tuning, under which weight decay still does not move OC ($p > 0.39$). Thus tameness is jointly determined by model and data: the quantization null is robust, while distillation's decoupling is confined to an intermediate-complexity regime.

Operator composition. We apply distillation, regularization, and quantization separately to a fixed plain baseline so their displacements are comparable. Interactions such as distilling a regularized model may produce combined directions; characterizing this operator algebra remains future work.