Row-Attention Extracts, Column-Attention Projects: How ConTextTab Solves In-Context Linear Classification

Anonymous Authors[1]

Abstract

Tabular Foundation Models (TFMs) achieve strong zero-shot classification via in-context learning, but their internal computations remain poorly understood. We present, to our knowledge, the first causal intervention study of a TFM. ConTextTab is uniquely suited to mechanistic analysis because its per-cell tokenisation preserves the feature axis as a directly observable computational dimension, unlike TFMs that bundle features into row-level tokens. We study it on the Rotated Linear Threshold task in $\mathbb{R}^2$, whose Bayes-optimal algorithm is known exactly, and address its post-LayerNorm architecture by restricting activation patching to sub-block boundaries – a clean causal baseline that does not require auxiliary modeling choices. We localise a linear plug-in classifier $\mathrm{LD}(x^*) \approx f\big(x^* \cdot \hat{n}(\alpha) - \hat{\theta}\big)$ to the two architectural axes: row attention extracts the boundary normal $\hat{n}(\alpha)$ and threshold $\hat{\theta}$ in a single sub-block (L0), distributed across all twelve heads with none individually dispensable; column-attention then progressively projects the query onto the boundary, with the final three of twelve layers carrying 0.588 of the causal weight.

1. Introduction

Tabular Foundation Models (TFMs) are Transformer-based models pretrained at scale for tabular prediction, performing supervised learning on new data via in-context learning (ICL) without parameter updates (Hollmann et al., 2023; 2025; Qu et al., 2025; Spinaci et al., 2025). Despite strong empirical performance, the internal computations that enable this zero-shot generalisation remain poorly understood.

Mechanistic interpretability provides a toolkit for reverse-engineering how neural networks implement algorithms, but applying it to TFMs poses two structural challenges absent in language models. First, TFMs process a two-dimensional grid of features (columns) and data points (rows) with column-wise and row-wise attention mechanisms rather than a one-dimensional token sequence. Second, the post-LayerNorm architecture (Xiong et al., 2020) entangles pre-LN activations from distinct computational paths, invalidating the additive decompositions on which standard causal tools rely.

We adapt the standard toolkit to this setting and present, to our knowledge, the first causal intervention study of a TFM. We focus on ConTextTab (Spinaci et al., 2025) specifically because its per-cell tokenisation—each (row, feature) cell embedded as its own token in an $N \times C$ grid—preserves the column axis as a directly observable computational dimension. TFMs that bundle features into a single token, such as TabPFN (Hollmann et al., 2025) (one token per row) or TabICL (Qu et al., 2025) (one token per row, with features fused into the row representation), foreclose this granularity of mechanistic analysis. We evaluate on the Rotated Linear Threshold (RLT) task in $\mathbb{R}^2$, a parametric synthetic classification problem whose ground-truth inference algorithm is known exactly, providing a controlled setting for formulating and testing mechanistic hypotheses.

Contributions.

1. We adapt the mechanistic interpretability toolkit – logit lens, linear probing, and activation patching – to a post-LayerNorm tabular foundation model with axial attention. We show that restricting activation patching to post-LayerNorm checkpoints provides a baseline at which the intervention is a well-posed causal operation without auxiliary statistics, and we map out the consequences of this restriction.

2. We map the layer-wise information flow in ConTextTab on the RLT task. row attention is the locus of cross-example aggregation and carries near-ceiling probe quality for both α and θ from L0 onwards; column-attention is the dominant attention pathway transforming the residual into the scalar logit difference, with the causal effect back-loaded into the final three layers.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Scope. We restrict the present analysis to RLT in $\mathbb{R}^2$, where the optimal algorithm is most cleanly observable. Section 6 discusses generalisation; preliminary experiments at $d \in \{4, 8, 16\}$ are reported in Section A and indicate that the column-attention dominance and the OV-versus-QK dissociation persist, with MLP sub-blocks contributing non-negligibly at $d \geq 4$.

2. Related Work

Mechanistic interpretability of in-context learning. The mech-interp toolkit we use originates in the language-model literature: causal patching (Vig et al., 2020), the logit lens (nostalgebraist, 2020), attribution patching (Nanda, 2022), and linear probing (Alain & Bengio, 2017; Park et al., 2023), on the conceptual foundation laid by Olah et al. (2020) and Elhage et al. (2021). These tools have been used to identify ICL-relevant circuits in language models, notably the induction-head circuit (Olsson et al., 2022) and the indirect-object identification circuit (Wang et al., 2022). Our work transfers this methodology to a non-language ICL setting in which the ground-truth inference algorithm is known exactly, allowing us to evaluate the localisation against a known target rather than against a behavioural proxy.

Transformers as in-context learners of algorithms. A complementary line of work asks whether transformers implement classical learning algorithms in-context. Garg et al. (2022) show that transformers can in-context-learn linear functions, decision trees and small neural networks; Von Oswald et al. (2023) and aky argue that transformers trained on linear regression implement gradient-descent-like or ridge-regression-like updates internally, with subsequent work providing mechanistic and theoretical accounts of this construction. Our finding that ConTextTab realises a plug-in linear classifier $\mathrm{LD}(x^*) \approx f\big(x^* \cdot \hat{n}(\alpha) - \hat{\theta}\big)$ on RLT sits in this conceptual landscape but in the tabular-classification regime, and our contribution is at the level of identifying the components that realise the algorithm, rather than function-class equivalence.

Interpretability of tabular foundation models. Pretraining data varies across TFMs: TabPFN (Hollmann et al., 2023; 2025) and TabICL (Qu et al., 2025; 2026) use fully synthetic data generated by structural causal models, whereas ConTextTab (Spinaci et al., 2025) trains on real-world datasets from T4 (Gardner et al., 2024). Interpretability work on TFMs has so far focused on post-hoc explanation methods adapted to the in-context setting: Rundel et al. (2024) provide TabPFN-specific implementations of Shapley values, leave-one-covariate-out, and data-valuation estimators that exploit ICL to avoid retraining, and Sena & Azevedo (2026) integrate Shapley value regression directly into the architecture for real-time feature attributions in a single forward pass. These methods treat the model as an input–output system to be explained behaviourally. Mech-interp-style analysis of the internal computations, in contrast, remains nascent: Gupta et al. (2026) use linear probing to show that on linear regression datasets, quantities such as linear coefficients are linearly decodable from intermediate representations of TabPFN, but stop short of causal analysis. To our knowledge, ours is the first study to combine probing with causal interventions on a TFM.

3. Background

3.1. Tabular In-Context Learning

A tabular foundation model is a neural network q_θ trained to approximate the posterior predictive $p(y^* \mid \vec{x}^*, \mathcal{D}_{\mathrm{ctx}})$, where $\mathcal{D}_{\mathrm{ctx}} = \{(\vec{x}_i, y_i)\}_{i=1}^n$ is the labelled context and $(\vec{x}^*, y^*)$ is the query. At test time a single forward pass produces predictions with no gradient updates.

3.2. ConTextTab Architecture

ConTextTab maps the combined context and query set into a token grid $x^{(0)} \in \mathbb{R}^{N \times C \times d_{\mathrm{model}}}$ (N rows, C column tokens per row, $d_{\mathrm{model}} = 768$). Numeric values are embedded via quantile-binned lookup tables, dates via summed positional embeddings, and text/categorical values via a frozen MiniLM-L6-v2 encoder (Wang et al., 2020), all projected to d_{model}. The target value is embedded into the last column token only; for query rows this token carries a learnable mask embedding.

Axial attention. The grid passes through $L = 12$ post-LayerNorm transformer layers ($H = 12$ heads, $d_h = 64$, $d_{\mathrm{mlp}} = 3072$). Each layer applies four sub-blocks in sequence—column-attention (over the C columns, independently per row), a column-MLP, row attention (over the N rows, independently per column), and a row-MLP—each following the post-LN update $x \leftarrow \mathrm{LN}(x + f(x))$:

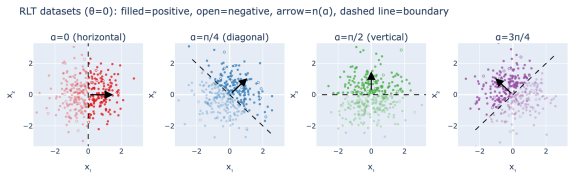
$$x^{(\ell)}_{\mathrm{col,mid}} = \mathrm{LN}\Big(x^{(\ell)} + \mathrm{MHA}^{(\ell)}_{\mathrm{col}}(x^{(\ell)})\Big), \tag{1}$$

$$x^{(\ell)}_{\mathrm{col,post}} = \mathrm{LN}\Big(x^{(\ell)}_{\mathrm{col,mid}} + \mathrm{MLP}^{(\ell)}_{\mathrm{col}}(x^{(\ell)}_{\mathrm{col,mid}})\Big), \tag{2}$$

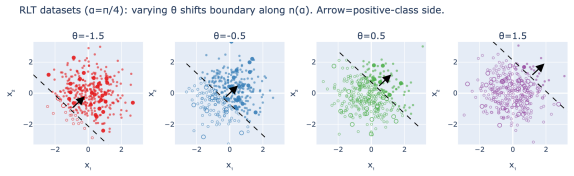
$$x^{(\ell)}_{\mathrm{row,mid}} = \mathrm{LN}\Big(x^{(\ell)}_{\mathrm{col,post}} + \mathrm{MHA}^{(\ell)}_{\mathrm{row}}(x^{(\ell)}_{\mathrm{col,post}})\Big), \tag{3}$$

$$x^{(\ell+1)} = x^{(\ell)}_{\mathrm{row,post}} = \mathrm{LN}\Big(x^{(\ell)}_{\mathrm{row,mid}} + \mathrm{MLP}^{(\ell)}_{\mathrm{row}}(x^{(\ell)}_{\mathrm{row,mid}})\Big). \tag{4}$$

The four post-LN checkpoints $x^{(\ell)}_{\mathrm{col,mid}}$, $x^{(\ell)}_{\mathrm{col,post}}$, $x^{(\ell)}_{\mathrm{row,mid}}$, and $x^{(\ell+1)}$ are the intervention sites used throughout Section 4.2.



(a) Four datasets with $\theta = 0$ and $\alpha \in \{0,\ \pi/4,\ \pi/2,\ 3\pi/4\}$.



(b) Four datasets with $\alpha = \pi/4$ and $\theta \in \{-1.5,\ -0.5,\ +0.5,\ +1.5\}$.

Figure 1. The RLT task. Filled/open markers denote positive/negative labels. The arrow shows $n(\alpha)$ and the dashed line the boundary $x \cdot n(\alpha) = \theta$.

Output head. The final residual state $\mathbf{h}^{(L)} \in \mathbb{R}^{d_{\text{model}}}$ at each query row is decoded by a two-layer MLP head

$$f(\mathbf{h}) = W_2\ \mathrm{GELU}(W_1\mathbf{h} + \mathbf{b}_1) + \mathbf{b}_2\ \in\ \mathbb{R}^K, \quad (5)$$

with $K = 64$ classes; trailing logits are discarded at inference. In our binary RLT setting the model outputs a two-class probability $P(y = 1)$ for each query row.

3.3. The Rotated Linear Threshold Task

The RLT task is binary classification over $\mathbb{R}^2$. With $n(\alpha) = (\cos\alpha, \sin\alpha)^\top$ the unit boundary direction, the label rule is

$$y = \mathbf{1}\big[x \cdot n(\alpha) > \theta\big], \qquad x \sim \mathcal{N}(0, I_2). \quad (6)$$

Latent parameters are drawn independently per dataset: $\alpha \sim \mathrm{Uniform}[0, \pi)$ and $\theta \sim \mathcal{N}(0, 1)$ truncated to $[-2, 2]$. Both are unknown to the model and must be inferred from the labelled context.

Ideal algorithm. The Bayes-optimal algorithm computes $\hat{n}(\alpha) \propto \sum_{i \in \text{ctx}} (2y_i - 1)x_i$ and classifies by $\mathrm{sign}(x_q \cdot \hat{n}(\alpha) - \hat{\theta})$. Under Gaussian x, this label-weighted mean coincides up to a scalar with the LDA plug-in direction.

4. Methods

4.1. Logit Lens for Tabular Foundation Models

In a pre-LN architecture the sub-block update $x' = x + f(\mathrm{LN}(x))$ yields a residual stream that decomposes into independent per-layer contributions, $\mathbf{h}^{(L)} = \mathbf{h}^{(0)} + \sum_\ell f_\ell(\mathrm{LN}(\mathbf{h}^{(\ell-1)}))$. ConTextTab's post-LN update (Equations (1) to (4)) instead applies LayerNorm to the sum of the current state and the sub-layer output, entangling the two through the shared mean and variance: $\mathrm{LN}(x + f(x)) \neq \mathrm{LN}(x) + \mathrm{LN}(f(x))$. The residual stream therefore does not decompose into clean additive per-layer contributions, and the standard tools that rely on such a decomposition lose their theoretical grounding.

We apply the logit lens by passing each post-LN checkpoint state $\mathbf{h}^{(\ell)} \in \mathbb{R}^{d_{\text{model}}}$ through the trained nonlinear MLP head f of Equation (5). These are the actual vectors produced by the forward pass and seen by all downstream computation. The trace should be interpreted as an empirical diagnostic of what the output head would predict if it read from a given checkpoint, not as a theoretically grounded decomposition of the residual stream.

4.2. Activation Patching

Let $a_s(D)$ denote the activation at intervention site s produced by running the model on a clean dataset D, and $\mathrm{LD}(D) = f_1(D) - f_0(D)$ the logit difference at the query row. The patched forward pass

$$M\big(D';\ \mathrm{do}\big(a_s \leftarrow a_s(D)\big)\big) \quad (7)$$

runs the model on a corrupt dataset D' but replaces the activation at s with the value cached from the clean run.

Restriction to post-LN checkpoints. At a pre-LN site, the LayerNorm of the patched forward pass computes its σ over a vector that mixes the patched contribution with corrupt residuals. Concretely, consider the row attention sub-block (Equation (3)), and write $\mathrm{LN}(v) = \gamma \odot \frac{v - \mu(v)}{\sigma(v)} + \beta$ with $\mu(v), \sigma(v)$ computed over the full pre-LN vector. Writing $r = x^{(\ell)}_{\text{col,post}}$ and $\mathrm{MHA}^{(\ell)}_{\text{row}}(r) = \sum_{h=1}^{H} a^{(\ell)}_h$, suppose we patch head h with the clean value $\tilde{a}^{(\ell)}_h$ while holding the remaining heads at their corrupt values. The patched pre-LN vector is

$$v_{\text{patch}} = r' + \tilde{a}^{(\ell)}_h + \sum_{h' \neq h} a^{(\ell)\prime}_{h'}, \quad (8)$$

and the clean contribution $\tilde{a}^{(\ell)}_h$ is rescaled in the post-LN state by $\sigma(v_{\text{patch}})$, which depends on r' and $\{a^{(\ell)\prime}_{h'}\}_{h' \neq h}$. The estimated causal effect is therefore confounded by the corrupt context through the normalisation denominator. The same argument applies to any pre-LN intermediate. No such coupling arises at post-LN checkpoints, which are the actual vectors produced by the forward pass; we therefore restrict all patching to post-LN sites. We adopt the post-LN restriction as a clean baseline. A consequence is that pre-LN intermediates such as individual head outputs cannot be patched cross-run; we therefore work at sub-block granularity for all causal claims, recovering head-level resolution at one site (L0 row attention) via within-run mean ablation (Section 5.3.2).

Row-stratified interventions. We apply Equation (7) separately to each row partition. A *query-row patch* replaces only the query-row slice at site s and measures how much of the prediction is determined by the query representation alone; a *context-row patch* replaces all context-row slices and measures how much is determined by the context.

Paired datasets. We construct two interchange families that share the feature matrix X and differ in exactly one parameter. *α-pairs* keep θ fixed and set $\alpha' = \alpha + \pi/2$, rotating the boundary by 90 while leaving the class prior unchanged. *θ-pairs* keep α fixed and contrast $\theta' = -\theta$, shifting the class balance without altering the boundary direction. We discard pairs with fewer than 10 minority-class context examples or with output gap $|\mathrm{LD}(D) - \mathrm{LD}(D')| < \delta_{\min} = 0.3$.

4.3. Patching Metric

The *normalised indirect effect* at site s for pair i is

$$\tau_s^{(i)} = \frac{\mathrm{LD}_{\mathrm{patch}}^{(s)} - \mathrm{LD}(D')}{\mathrm{LD}(D) - \mathrm{LD}(D')}, \tag{9}$$

the fraction of the clean – corrupt output gap causally mediated through s. $\tau = 0$ means no recovery, $\tau = 1$ means full recovery. The sub-block attribution $\Delta\hat{\tau}_{\ell,b} = \hat{\tau}_{s_{\text{after}}} - \hat{\tau}_{s_{\text{before}}}$ measures the causal increment contributed by sub-block b in layer ℓ.

4.4. Linear Probing

We probe four targets at the (query row, target column) cell of each post-LN checkpoint across $N_{\mathrm{DS}} = 400$ held-out datasets, using five-fold cross-validation: the boundary angle α (Ridge regression onto $(\cos\alpha, \sin\alpha) \in S^1$, reported as $\bar{R}^2_\alpha$ and mean angular error $\bar{\varepsilon}$); the sign of the threshold θ (binary LDA, chance 0.5); the angle of the empirical context sufficient statistic $\hat{\mathbf{s}}$ (same as α); and the scalar query projection $p^* = x^* \cdot n(\alpha)$ (Ridge, reported as R^2). Component-level probes use the full post-LN state $\in \mathbb{R}^{d_{\mathrm{model}}}$; head-level probes use the value-weighted pre-projection output $z_{l,h} \in \mathbb{R}^{d_{\mathrm{head}}}$ separately for column- and row attention.

For circular targets (α and the angle of $\hat{\mathbf{s}}$) we regress onto the unit-circle embedding $(\cos\alpha, \sin\alpha) \in S^1 \subset \mathbb{R}^2$. For a head output $z_{l,h} \in \mathbb{R}^{d_{\mathrm{head}}}$, we fit two independent Ridge regressors and report

$$\begin{aligned} \bar{R}^2_\alpha &= \tfrac{1}{2}(R^2_{\cos} + R^2_{\sin}), \\ \bar{\varepsilon} &= \frac{1}{N}\sum_{i=1}^{N} \arccos\left(\frac{\hat{y}^{(i)}_{\cos}\cos\alpha_i + \hat{y}^{(i)}_{\sin}\sin\alpha_i}{\|\hat{y}^{(i)}\|}\right). \end{aligned} \tag{10}$$

A random predictor yields $\bar{\varepsilon} = \pi/2 \approx 90°$.

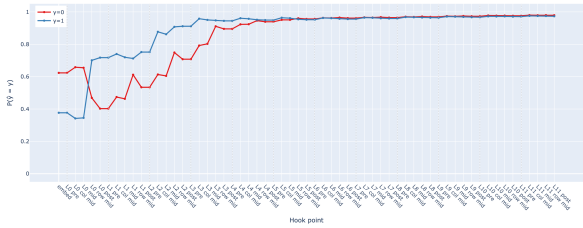


Figure 2. Logit-lens trace. $P(y = \text{correct label})$ at each sub-layer boundary, averaged across query rows and coloured by ground-truth label.

5. Experiments

5.1. Experimental Setup

All experiments use the publicly released ConTextTab checkpoint.[1] RLT datasets have $d = 2$ features and $N_{\mathrm{ctx}} = 200$ context examples plus one query point. The model solves the task in this setting (mean accuracy 0.981, AUC-ROC 0.99 across 1000 held-out datasets), confirming that the task is learnable via in-context learning at this context size. For patching, α- and θ-pairs are constructed as in Section 4.2 with $\delta_{\min} = 0.3$. For probing, we draw $N_{\mathrm{DS}} = 400$ independent RLT datasets spanning all combinations of (α, θ).

5.2. The logit lens reveals two phases

Figure 2 shows $P(\hat{y} = y)$ obtained by applying the output head to the residual stream at each sub-layer boundary for the query row. At the final layer the model achieves $P(\text{correct}) = 0.989$ for y=0 and 0.976 for y=1 queries. The trace exhibits two qualitatively distinct phases. In layers 0–3 the two class means rise to $P \approx 0.7$–0.8 with a persistent gap of approximately 0.25 in favour of y=1, consistent with a class-balance shortcut $\bar{y} \approx \Phi(-\theta)$ rather than a query-dependent decision. From layer 4 onward the gap closes and both class means rise together monotonically, the signature one would expect of a class-agnostic projection $x^* \cdot \hat{n}(\alpha)$ being completed in column-attention. The remaining sections of Sections 5.3 and 5.4 characterise the mechanisms underlying these two phases.

5.3. Phase 1: Distributed Direction Extraction in row attention

5.3.1. All twelve L0 row attention heads encode direction

Figure 3 shows $\bar{R}^2_\alpha$ and θ-LDA accuracy across all 12×12 head-layer combinations for both attention types. The two attention mechanisms display qualitatively opposite patterns. Column-attention shows layer-specific peaked

[1] https://huggingface.co/SAP/sap-rpt-1-oss

signals that emerge at L1–L2 and decay with depth. Figure 4 shows the geometric structure at the peak column attention heads: at `L2H6`, α is encoded as a circular variable on S^1 rather than as an arbitrary scalar (**??**); at `L1H2`, a single linear projection separates $\operatorname{sign}\theta$ at LDA $= 0.989$ (**??**). row attention by contrast carries high-quality representations of both variables from the embedding layer onward: at L0 it already achieves $\bar{R}^2_\alpha \in [0.80, 0.87]$ and θ-LDA above 0.97 across all twelve heads, exceeding the peak column-attention performance at any layer. The signal is distributed across heads with no single head standing out as a dedicated direction head, and degrades only slowly with depth.

5.3.2. Head-level ablation

The probing result invites a natural causal question: is one of the twelve L0 heads dominantly responsible for the direction signal while the others write redundant copies, or is the computation distributed?

We test this with two interventions on the L0 row attention output, evaluated by the angular error of the L1 query-row query-vector relative to a Ridge regressor trained on clean activations. *Single-head ablation* replaces the output of head h with its mean over the dataset, holding all other heads at their clean values. The clean angular error is $10.3°$. Ablating any single head sends the error to a value in the range 84–$93°$, near the $90°$ chance level for circular regression (Figure 5). The probe-leading head `H6`, despite the highest individual probe R^2 at L0, is the *least* disruptive ($+80.5°$); the most disruptive is `H0` ($+89.6°$). The probe leader is therefore not the direction head in any causal sense.

This is consistent with both a partial-sum decomposition (each head contributes a distinct component of the same statistic) and a redundant-copy decomposition combined with a magnitude effect (under which ablation reduces the summed magnitude by $11/12$ and could degrade the L1 probe through magnitude rather than direction). To separate them we add a *rescaling control*: mean-ablate head h and additionally multiply the surviving eleven head outputs by $12/11$, which preserves the expected magnitude of the summed L0 row attention output. Under a strict redundant-copy decomposition this should restore the L1 representation; under a partial-sum decomposition it should not. The rescaled and unscaled ablations produce angular-error increases that coincide within evaluation noise. The control therefore rules out the simplest redundant-copy hypothesis. We note that it does not distinguish a strict partial-sum decomposition (orthogonal subspaces) from intermediate cases (overlapping but not identical subspaces); both predict that rescaling-after-ablation fails to restore the probe.

5.3.3. Patching: L0 row attention is sufficient

The head-level ablation establishes that no single head can be removed. We now ask the dual question, whether the L0 row attention sub-block alone is sufficient to install the direction signal at L1.

We use α-pairs and inject the post-LN output of the L0 row attention sub-block from the clean run into the corrupt run, leaving all other activations corrupt. We then probe the L1 query-row Q-vector for $\hat{n}(\alpha)$. In the corrupt baseline the L1 Q-vector encodes $\hat{n}(\alpha + \pi/2)$, so the regressor trained on $\hat{n}(\alpha)$ predicts the wrong direction (mean angular error $176°$, $R^2 = -2.97$). With the L0 row attention output patched, the L1 Q-vector recovers to within $3.7°$ of clean ($R^2 = 0.991$). Combined with Section 5.3.2, this characterises L0 row attention as a distributed sub-block that is jointly necessary at the head level and individually sufficient at the sub-block level for the L1 direction representation.

5.4. Phase 2: Column-Attention Encodes the Query Projection

Phase 1 writes $\hat{n}(\alpha)$ into the residual stream at L1. Phase 2 must produce the scalar $p^* = x^* \cdot \hat{n}(\alpha)$, the quantity directly consumed by the output head.

5.4.1. Column-attention dominates the attention pathway

Figures 6 and 7 show the patching profile and per-sub-block attribution. Under context-row patching, $\hat{\tau}$ saturates at approximately 1.000 already at the embedding layer for both interchange families: the raw embeddings of the 200 labelled context examples therefore contain all the task-relevant information that the model uses. Under query-row patching, $\hat{\tau}$ rises smoothly from 0 at the embedding to 1 at the final layer, with no sharp transition. Column-attention carries the dominant share of the causal weight at every layer and the effect is strongly back-loaded: $\Delta\hat{\tau} = 0.130, 0.192, 0.266$ at L9, L10, L11 for α-interchange, so the three final column attention sub-blocks together account for 0.588 of the total unit effect. The remaining ≈ 0.41 of the causal weight is distributed across L1–L8 column-attention with no individual sub-block dominant; row attention and both MLP sub-blocks contribute negligibly throughout. Under context-row patching the picture inverts, with early row attention and row-MLP carrying the bulk of the effect.

5.4.2. Late column-attention heads encode the query projection

We probe each column attention head for the scalar query projection $p^* = x^* \cdot \hat{n}(\alpha)$. The top six heads by $R^2(p^*)$ all reside in layers 9–11 and achieve $R^2 \in [0.818, 0.835]$,

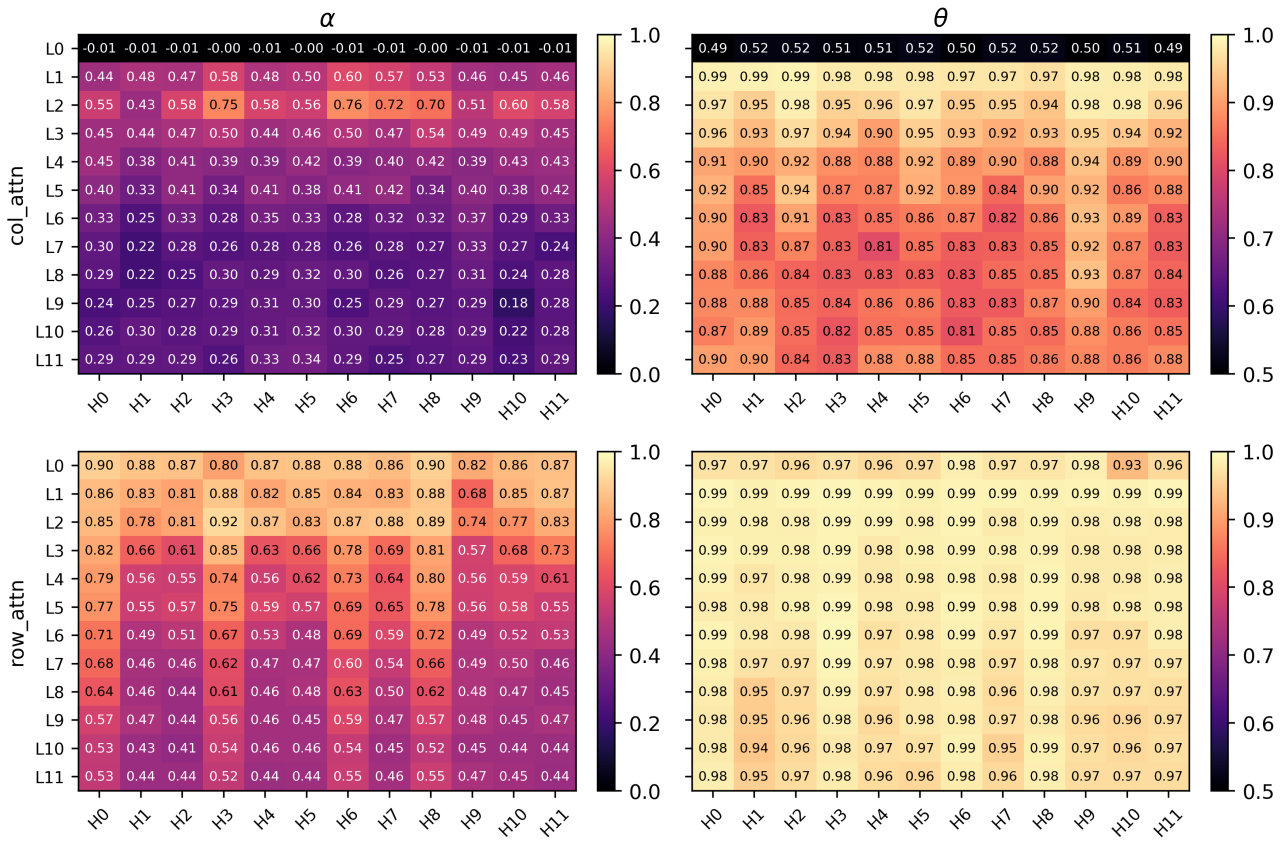


Figure 3. Probe R^2 heatmaps across 12 layers and 12 heads, for $\bar{R}^2_{\cos\alpha}$, $\bar{R}^2_{\sin\alpha}$, and θ-LDA. Top: `col_attn`; bottom: `row_attn`. row attention is uniformly high from L0; column-attention is layer-specific.

with class-indiscriminate scatter aligned along the identity diagonal (Figure 8). What is encoded is therefore the class-agnostic scalar projection rather than a class label. Earlier column attention layers carry partial α-information (Figure 3, peak $\bar{R}^2_\alpha = 0.76$ at `L2H6`) but progressively lose direction decodability as it is consumed into the scalar. The co-location of high $R^2(p^*)$ with high causal weight at L9–L11 is joint representational and causal evidence that the dot product is read out at the late column attention sub-blocks.

5.5. Synthesis

The two phases assemble into a plug-in classifier of the form

$$\mathrm{LD}(x^*) \approx f\Big(x^* \cdot \hat{n}(\alpha) - \hat{\theta}\Big), \tag{11}$$

with $\hat{n}(\alpha)$ assembled by the twelve L0 row attention heads collectively and $x^* \cdot \hat{n}(\alpha)$ realised across L1–L11 column-attention via the OV pathway, with the final three sub-blocks carrying 0.588 of the causal weight.

Axis-to-stage assignment. Axial attention forces a structural split. Row attention operates per column, attending across all N rows, and is the only sub-block type whose receptive field admits aggregation across context examples; column-attention operates per row, attending across the C feature columns, and is the only one that can combine within-sample features with information already written into the residual stream. The empirical question is therefore not *whether* the axes correspond to stages but *how cleanly*, and what each axis carries. We find that the cross-example aggregation is fully completed in a single sub-block (L0 row attention)—sharper than architectural necessity demands—and that the within-sample evaluation is back-loaded into L9–L11 column-attention via the OV pathway.

Reconciling the logit-lens and patching profiles. The early class-balance gap in the logit-lens trace (Figure 2, layers 0–3) and the late causal saturation in the patching profile (Figure 6, layers 9–11) measure different things and are consistent under the decomposition above. The early logit-lens jump reflects the broadcast of $\hat{\theta}$ across all twelve column attention heads at L1 and the resulting class-balance shortcut, which is a partial predictor whose accuracy plateaus around $P = 0.7$–0.8. The late restoration of $\hat{\tau}$ reflects the completion of the class-agnostic projection $x^* \cdot \hat{n}(\alpha)$, which only becomes causally relevant to the query-row prediction

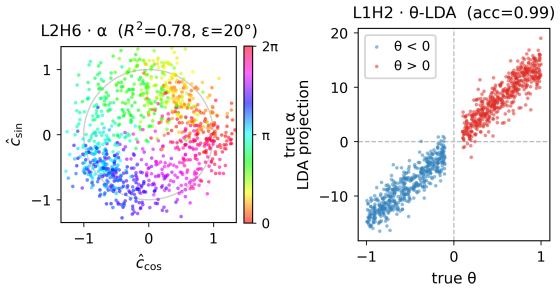


Figure 4. Geometric structure of the column-attention direction encoding at peak heads. **(a)** Col-attn α-probe at `L2H6` ($\bar{R}^2_\alpha = 0.78$, $\bar{\varepsilon} = 20.1°$): predicted $(\hat{y}_{\cos\alpha}, \hat{y}_{\sin\alpha})$ coloured by true α; dashed circle marks S^1. α is represented on the unit circle—the rainbow ordering preserves angular structure rather than collapsing α to a scalar. **(b)** Col-attn θ-probe at `L1H2` (LDA $= 0.989$): projection $z \cdot \mathbf{w}_{\text{LDA}}$ vs. true θ (red: $\theta > 0$, blue: $\theta < 0$). sign θ is read out near-perfectly by a single linear projection. Per-layer scatter grids for all heads are deferred to Section B.

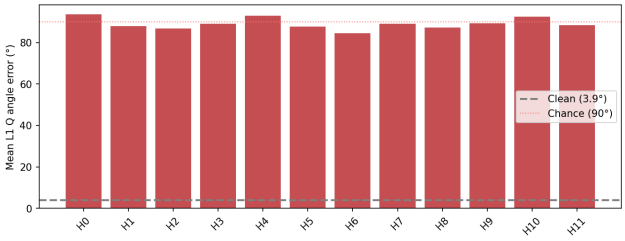


Figure 5. L0 row attention head ablation. Bars show the L1 query-row Q-vector angular error after mean-replacement of each head. Clean: $10.3°$. Ablation of any single head: 84–$93°$. Dashed line marks chance ($90°$).

once it has been written as a scalar into the residual stream by the final column attention sub-blocks.

6. Limitations and Open Questions

Sub-block patching granularity. The post-LN restriction precludes head-level patching across runs. We recover head-level resolution at one site (L0 row attention) via within-run mean ablation, but the same trick does not extend cleanly to mid-stream heads. Whether late-layer column-attention heads (e.g. those leading on $R^2(p^*)$) are individually necessary is open.

Independence of the α and θ channels. $\hat{n}(\alpha)$ and $\hat{\theta}$ are both extracted in L0 row attention but propagate through apparently distinct sub-pathways to the final logit difference, suggesting but not establishing additive contributions. A double-patching experiment testing $\hat{\tau}_{\alpha+\theta} \approx \hat{\tau}_\alpha + \hat{\tau}_\theta$ would settle whether the contributions are indeed additive.

Anisotropic input distributions. All experiments assume isotropic Gaussian inputs, where the LDA and label-

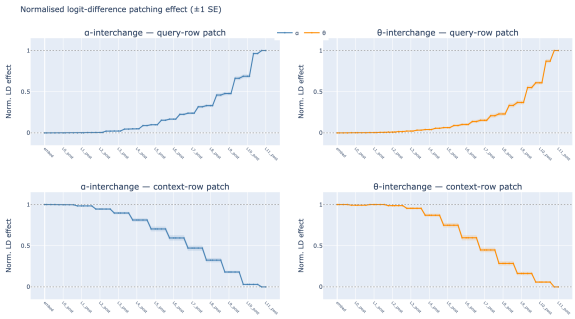


Figure 6. Logit-difference restoration $\hat{\tau}_s$ at each post-LN checkpoint. Top: query-row patch. Bottom: context-row patch. Left: θ-interchange. Right: α-interchange.

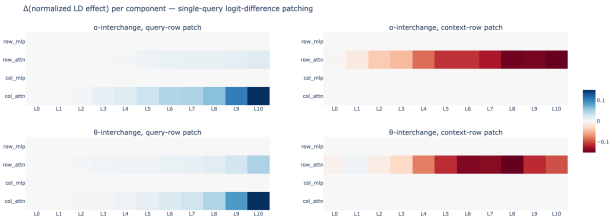


Figure 7. Per-sub-block attribution $\Delta\hat{\tau}_{\ell,b}$ for query-row (left) and context-row (right) patches. Column-attention dominates the query-row mediation; row attention and MLPs are negligible.

weighted-mean plug-in classifiers coincide up to a scalar. Whether the mechanism extends to arbitrary covariance is open. α-interchange patching shares Σ within each pair and therefore does not interrogate the covariance pathway.

Generalisation beyond RLT and $d = 2$. The RLT task is deliberately minimal and we restrict to $d = 2$ here. Preliminary experiments at $d \in \{4, 8, 16\}$ (Section A) indicate that the column-attention dominance, the back-loading, and the OV-versus-QK dissociation persist; MLP sub-blocks contribute non-negligibly at $d \geq 4$ and a complete characterisation of their role requires further work. Whether the same decomposition extends to heterogeneous real-world tabular data, or to TFMs trained on synthetic priors (TabPFN, TabICL), is open.

Architectural choices for mechanistic analysis. Our methodological constraints stem from two specific choices in ConTextTab: post-LayerNorm placement and per-cell tokenisation. A pre-LN TFM would lift the post-LN restriction and enable head-level cross-run patching with the standard mech-interp toolkit. Among existing models, TabICL (Qu et al., 2025) uses a pre-LN architecture, but its row-level tokenisation forecloses per-feature analysis along the column axis. A pre-LN TFM that retained ConTextTab-style per-cell tokenisation would combine both desirable

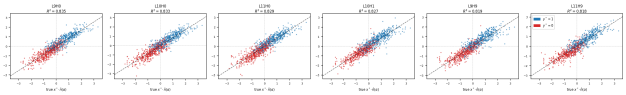


Figure 8. Top six column attention heads by $R^2(p^*)$, all from layers 9–11. Ridge prediction $z_{\cdot\mathrm{reg}}$ versus oracle $p^* = x^* \cdot \hat{n}(\alpha)$. Blue: $y^* = 1$; red: $y^* = 0$. The class-indiscriminate diagonals confirm encoding of the class-agnostic scalar projection rather than a class label.

properties for mechanistic interpretability.

7. Conclusion

We have presented the first causal intervention study of a tabular foundation model. Using ConTextTab on the Rotated Linear Threshold task in $\mathbb{R}^2$, we combined logit-lens analysis, linear probing, within-run head ablation, and activation patching at post-LN sub-block boundaries to characterise the information flow and component roles in a controlled setting.

The evidence converges on an axial decomposition that realises a linear plug-in classifier $\mathrm{LD}(x^*) \approx f(x^* \cdot \hat{n}(\alpha) - \hat{\theta})$ along the axial structure: the boundary parameters $\hat{n}(\alpha)$ and $\hat{\theta}$ are extracted by L0 row attention as a distributed computation across all twelve heads, and the query projection $x^* \cdot \hat{n}(\alpha)$ is written into the residual stream across L1–L11 column-attention via the OV pathway, with the final three sub-blocks carrying 0.588 of the causal weight.

Two points seem to us of broader relevance. First, post-LayerNorm imposes a hard constraint on causal-intervention granularity that any mechanistic study of a post-LN TFM must respect; post-LN patching gives a clean baseline that does not require auxiliary modeling choices, and within-run mean ablation can recover head-level resolution at single sites. A pre-LN TFM that retained ConTextTab-style per-cell tokenisation would combine cross-run head-level patching with feature-axis analysability; we are not aware of an existing model with both properties. Second, the axial attention structure provides a natural correspondence between architectural axes and algorithmic stages, namely row attention for cross-example statistics and column-attention for within-sample evaluation.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Gardner, J. P., Perdomo, J. C., and Schmidt, L. Large scale transfer learning for tabular data via language modeling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=WH5blx5tZ1.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c529dba08a146ea8d6cf715ae8930cbe-Paper-Conference.pdf.

Gupta, A., Sethi, A., and Kumar, D. TabPFN Through The Looking Glass: An interpretability study of TabPFN and its internal representations, January 2026.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second, September 2023.

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, January 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08328-6.

Nanda, N. Attribution Patching: Activation Patching At Industrial Scale — Neel Nanda — neelnanda.io. https://www.neelnanda.io/mechanistic-interpretability/attribution-patching, 2022. [Accessed 06-05-2026].

nostalgebraist. interpreting GPT: the logit lens — LessWrong. August 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023. URL https://openreview.net/forum?id=T0PoOJg8cK.

Qu, J., Holzmüller, D., Varoquaux, G., and Morvan, M. L. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data. In *Forty-Second International Conference on Machine Learning*, June 2025.

Qu, J., Holzmüller, D., Varoquaux, G., and Morvan, M. L. TabICLv2: A better, faster, scalable, and open tabular foundation model, 2026. URL https://arxiv.org/abs/2602.11139.

Rundel, D., Kobialka, J., von Crailsheim, C., Feurer, M., Nagler, T., and Rügamer, D. Interpretable machine learning for tabpfn. In *World Conference on Explainable Artificial Intelligence*, pp. 465–476. Springer, 2024.

Sena, L. B. T. R. and Azevedo, F. G. [short] real-time explanations for tabular foundation models. In *ICLR 2026 Workshop on Foundation Models for Science: Real-World Impact and Science-First Design*, 2026. URL https://openreview.net/forum?id=StSMBSZqxx.

Spinaci, M., Polewczyk, M., Schambach, M., and Thelin, S. ConTextTab: A Semantics-Aware Tabular In-Context Learner. 2025. URL https://openreview.net/forum?id=kGMRb4jbTP.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in language models using causal mediation analysis. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *NeurIPS ML Safety Workshop*, 2022. URL https://openreview.net/forum?id=rvi3Wa768B-.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33: 5776–5788, 2020.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Zhang, H., Lan, Y., Wang, L., and Liu, T.-Y. On layer normalization in the transformer architecture, 2020. URL https://openreview.net/forum?id=B1x8anVFPr.

A. Higher-Dimensional RLT Experiments

We report preliminary results at $d \in \{4, 8, 16\}$ at the same context size $N_{\text{ctx}} = 200$. Zero-shot RLT accuracy degrades gracefully across this range, from 98.4% at $d = 2$ to 92.2% at $d = 16$. We summarise the findings.

Phase 1: per-column locality. Per-column locality is empirically clean from $d = 4$ onward (**??**). At $d = 4$ the diagonal mean of the column-restricted L0 row attention probe is 0.913 with off-diagonal mean 0.169; at $d = 8$, 0.844 versus -0.138; at $d = 16$, 0.670 versus -0.206. The diagonal R^2 degrades gracefully and the off-diagonal R^2 collapses to the regression-noise floor at every higher-dimensional setting.

Phase 1: head-level ablation across dimensions. Repeating the head-ablation protocol of Section 5.3.2, $12/12$ heads cross the $\Delta \geq 2^\circ$ threshold for causal consequence at $d \in \{4, 8\}$ but only $6/12$ cross it at $d = 16$. The clean angular error of the pooled L1 Q-vector probe is 86.7° at $d = 16$, near the 90° chance baseline; the probe is at floor before any ablation. The per-column locality matrix indicates that direction information is nevertheless present at the L0 row attention output at $d = 16$ ($R^2 \approx 0.67$), so the pooled-probe failure is consistent with a probe-design limitation rather than a model failure, but disentangling this requires a stronger probe.

Phase 2: column-attention dominance and OV-versus-QK persist. Under α-interchange query-row patching, the cumulative column-attention weight at L11 is 0.675, 0.690, 0.679 at $d \in \{4, 8, 16\}$, comparable to the 0.588 value at $d = 2$. The causal effect crosses 0.5 at L11 in all cases. The OV-versus-QK dissociation also persists: at L11 the OV-to-p^* R^2 stays high (0.800, 0.835, 0.745).

Phase 2: MLP contribution at $d \geq 4$. Summed across all layers and both MLP sub-blocks, the MLP causal weight under α-interchange is 0.467, 0.481, 0.417 at $d \in \{4, 8, 16\}$, materially larger than near-zero at $d = 2$. At $d \geq 4$ the MLPs do measurable causal work alongside column-attention. Whether they implement a generalisation-related computation that becomes necessary as d grows beyond the per-column receptive-field capacity, perform book-keeping that redistributes the same computation across more sub-blocks, or implement the multiplicative step that we have not localised at $d = 2$, is left to future work.

B. LDA Projections of Attention Heads

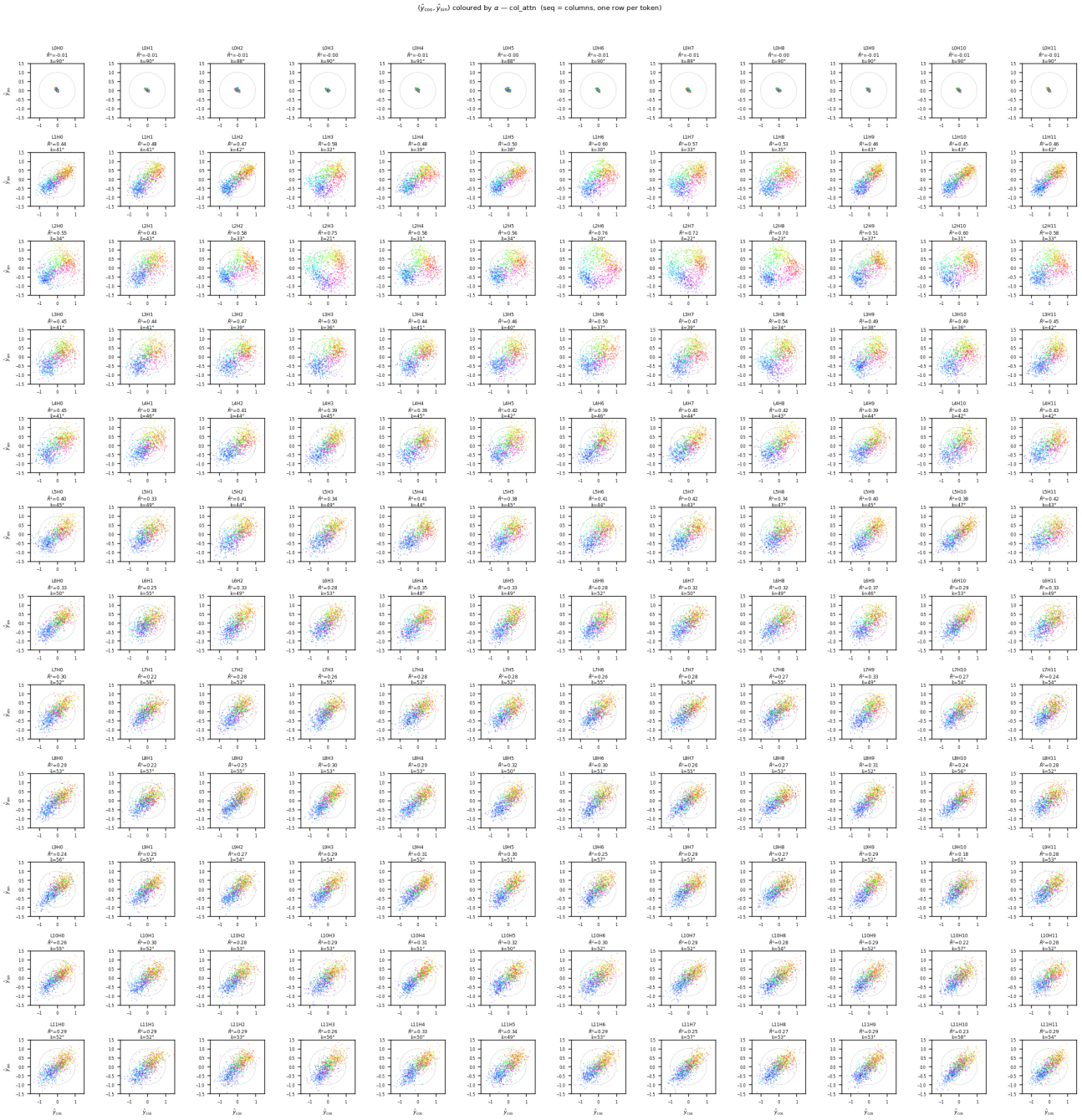


Figure 9. Column-attention α-probe: polar scatter of cross-validated predictions $(\hat{y}_{\cos}, \hat{y}_{\sin})$ for all 12 layers $\times$ 12 heads, $N = 1{,}000$ points per panel. Points coloured by true α; dashed circle marks the unit circle. Layer 0 shows a point mass at the origin; layer 2 heads L2H3, L2H6, L2H7, L2H8 show the clearest rainbow rings (peak $\bar{R}^2_\alpha = 0.76$, $\bar{\varepsilon} = 20°$).

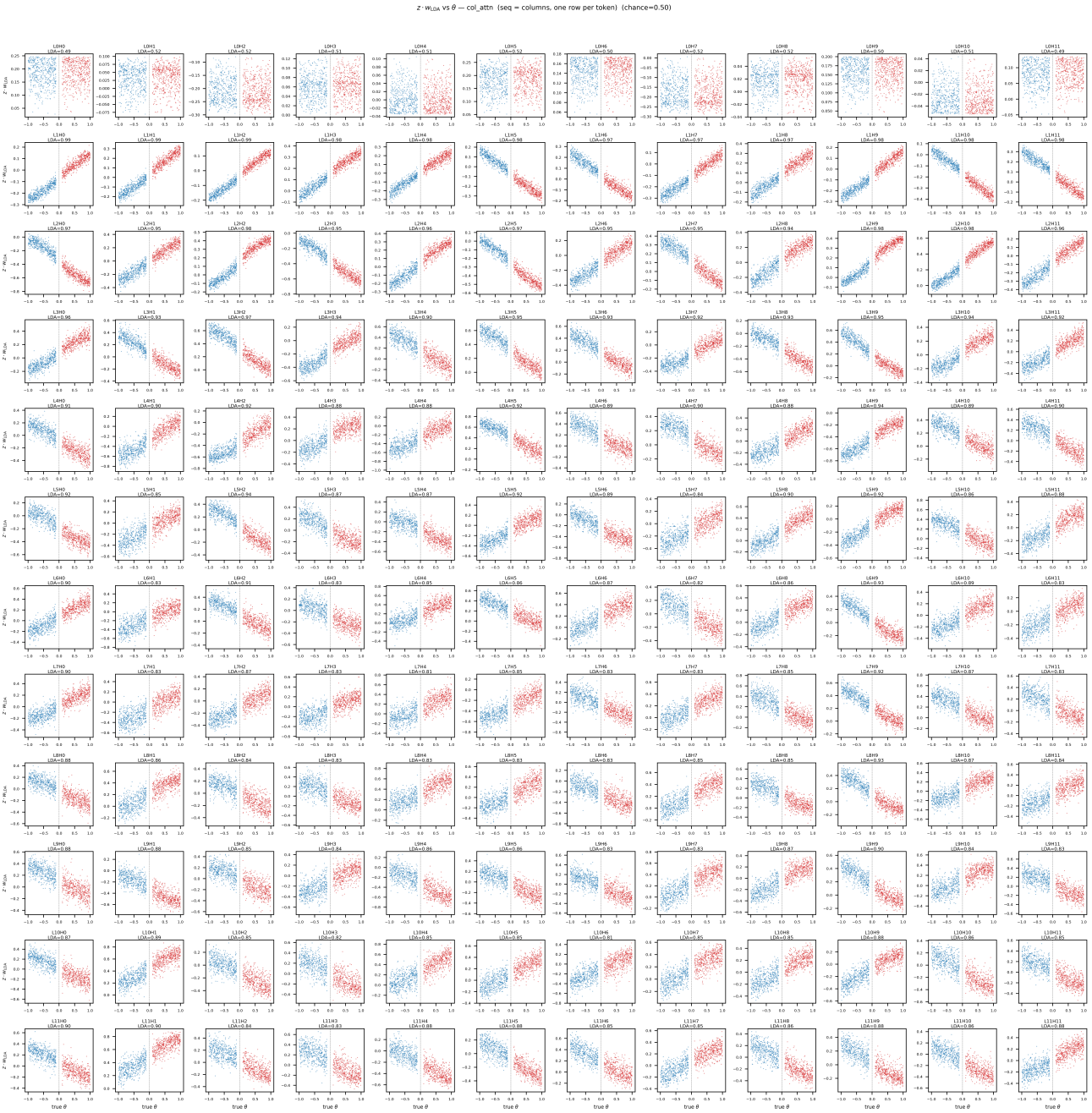


Figure 10. Column-attention θ-probe: binary LDA projection scatter for all 12 layers $\times$ 12 heads. Layer 0 at chance; all twelve layer-1 heads jump simultaneously to LDA ≥ 0.97; separation degrades gradually from layer 3 onward.

Figure 11. row attention α-probe. Complete rainbow rings present from layer 0 ($\bar{R}^2_\alpha \approx 0.80$–$0.87$, $\bar{\varepsilon} \approx 7$–$15°$).

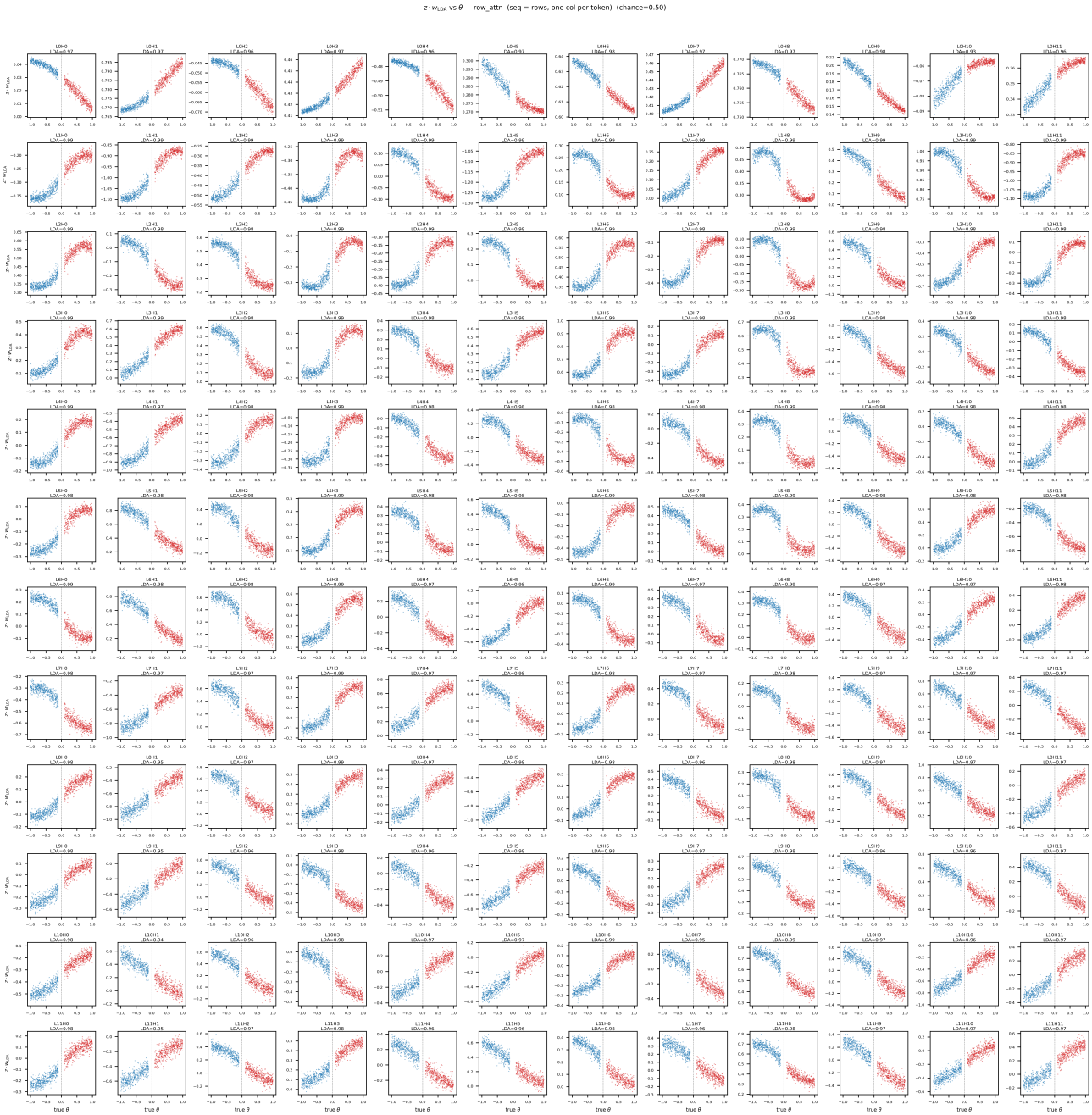


Figure 12. row attention θ-probe. Near-perfect separation (LDA ≈ 0.97) from layer 0, sustained across all twelve layers.