When to Trust a PFN: Detecting Harmful Shift in Tabular Foundation Models

Viet Nguyen
University of Toronto
viet@cs.toronto.edu

Herman Bergström
Chalmers University of Technology
hermanb@chalmers.se

Stephan Rabanser
Princeton University
rabanser@princeton.edu

Rahul G. Krishnan
University of Toronto
rahulgk@cs.toronto.edu

Abstract

Tabular prior-fitted networks (PFNs) work well across tabular classification problems but can fail silently at deployment when distributions shift. We propose **CALD** (Calibrated Anchor-Loss Disagreement), a lightweight unsupervised monitor that triggers an alarm to warn when a PFN might fail. By finetuning the decoder to maximize disagreement with the test-set prediction while maintaining accuracy on an ID anchor set, **CALD** identifies harmful OOD batches through the resulting anchor loss, requiring no labels on test batches. We prove that the post-finetune anchor loss has a strictly lower expectation under sufficiently severe shift than under ID eval, yielding a one-sided test. Across 14 datasets covering tabular and structured distribution shift, on both TabPFN and TabICL, **CALD** matches or outperforms baselines on harmful shifts and is markedly more resilient to benign-shift false alarms.

1 Introduction

Several machine learning (ML) areas underwent a paradigm shift in the last decade, transitioning from training custom models on task-specific data toward leveraging in-context learning (ICL) to obtain predictions without any parameter updates [7, 20, 66, 61, 1]. The tabular modality enjoyed its own such paradigm shift after the development of amortized pre-trained models known as prior-fitted networks (PFNs) [38, 21, 22, 46, 47], which provide classification and regression predictions on novel datasets in a single forward pass, outperforming the status quo of gradient boosted trees [9, 26, 45, 15]. PFNs for tabular data, such as TabPFN [21, 22, 16] and TabICL [46, 47], are trained on millions of synthetic datasets to ingest a training set (the "support set") and test features (the "query set") as in-context prompts and return accurate predictions, capturing rich and complex inductive biases which make up the tabular modality.

Still, PFNs largely inherit the same fundamental limitations as classical ML, remaining brittle under distribution shift [48, 51, 18, 27, 59, 37, 11, 68]. Because their synthetic priors construct the support and query sets as independent draws from the same generative process, PFNs expect by design that the support and the query follow the same distribution. Consequently, the model may experience silent performance deterioration when deployment data drifts substantially from the support set, while it is unclear whether existing out-of-distribution (OOD) detectors designed for the classical paradigm [19, 31, 28, 34, 52, 57, 64, 23, 67] are suitable for the in-context regime. Reliable deterioration detection is crucial if we are to leverage PFNs' strong performance in high-stakes applications, such as healthcare [2, 43, 10].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

A partial solution is to compute certain distributional statistics between training and deployment data to flag deviations in the covariate distribution [14, 35, 32, 49, 33, 69]. However, these approaches are inherently label-agnostic and therefore sensitive to *benign* shifts, where changes in the input space do not necessarily degrade model performance. More recently, methods that employ model disagreement have emerged as a more robust alternative [40, 25, 12, 60, 54, 41], leveraging the differences in a trained model's ability to disagree with its own predictions in- and out-of-distribution as an indicator of deterioration. However, these methods rely on perturbing the weights of a model that has been fine-tuned for a specific task, whereas an amortized PFN's weights remain fixed across tasks, and retraining is prohibitive and can degrade performance. Exploring how to apply disagreement-based detection in this regime is therefore a natural extension.

We introduce **CALD** (Calibrated Anchor-Loss Disagreement), a deployment deterioration monitoring mechanism for PFNs split into two phases (Figure 1). In the **Calibration** phase (2), **CALD** samples a small labeled *anchor* set from an in-distribution (ID) validation batch and treats the remainder as a held-out ID *eval* set. The decoder is then briefly fine-tuned under a dual objective consisting of an agreement term on the anchor and a disagreement term on the eval set with respect to the PFN's own pseudolabels. Repeating this procedure across many validation draws yields a distribution of anchor losses in the ID setting. During the **Deployment** phase (3), we run the same fine-tuning procedure, substituting a (possibly OOD) unlabeled deployment batch for the eval set. We compare the resulting anchor loss against the losses in the ID distribution, and flag it if it falls in the lower tail. This is motivated by the idea that a lower loss is achievable when the evaluated samples occupy a different region of the PFN's internal class-conditioned representations compared to the anchor samples.

Our contributions are as follows:

1. **A detector for tabular ICL.** We propose **CALD**, to our knowledge the first deployment-time deterioration detector for PFNs that exploits their specific structure at inference.
2. **Supporting theory.** We derive conditions under which the post-finetune anchor loss is provably lower in expectation under sufficiently severe shift than under ID eval, with the gap controlled by the alignment of the anchor-agreement and eval-disagreement gradient fields.
3. **Benchmark evaluation.** We evaluate **CALD** on 14 datasets spanning covariate, domain, and structured covariate shift, across four shift severities, for both TabPFN and TabICL. **CALD** matches or outperforms 15 post-hoc baselines on harmful shifts, and on benign shifts achieves the most favorable trade-off among all methods evaluated (Table 1).
4. **Ablations and mechanism.** We show that the detection signal scales with the harm a shift inflicts rather than with its magnitude, and we verify the gradient-interference mechanism behind the theory directly on the deployed models, by measuring the proof's premise along the finetuning trajectories. We further provide a large suite of ablations, among which are the finetuning budget and its deployment cost, the disagreement weight, a frozen deployment configuration covered by the false-alarm guarantee, robustness to anchor quality, and detectability across shift types.

2 Related Work

Post-hoc OOD detection. A long line of work scores each input under a frozen classifier by inspecting the softmax shape [19, 31, 44, 34], the distance from the training data in the model's feature space [28, 52, 57, 64], and auxiliary-loss gradients [23]. Yang et al. [67] survey the space. Label-agnostic two-sample tests [35, 33, 69, 49] target distributional change directly and can therefore lead to false positives on benign covariate shifts. Every detector in this family relies on a single forward pass through a frozen model and we adopt the strongest of these as baselines.

Disagreement-based monitoring. The observation that independently trained models often agree on ID samples but disagree on OOD samples [40, 25] has been leveraged in several lines of work. Detectron [12] trains a small ensemble under a constrained-disagreement objective and derives its verdict from the deployment-batch disagreement rate. Rosenfeld and Garg [54] add to this by introducing *disagreement discrepancy* and provable error bounds. Finally, D3M [41] constructs a harm-calibrated Bayesian three-stage test with explicit FPR/TPR guarantees. All three assume a fixed converged classifier whose weights are perturbed externally, which is prohibitive in the PFN regime,

where training a second copy is not an option. **CALD** instead perturbs the decoder of a model and reads its signal from the *anchor loss* rather than from the disputed prediction itself.

Adaptation under shift. A separate body of work tries to account for, rather than detect, distributional shifts. Methods include test-time adaptation [63, 42, 24, 58, 65], source-free DA [30], and feature alignment [56, 62]. This line of work is distinct from detection, answering the question of *how should we change the model?* as opposed to *should this prediction be trusted?*, and treating one as a substitute for the other is risky when the shift falls outside the family the adapter was designed for. The in-context translation of this idea is Drift-Resilient TabPFN [17], which retrains the PFN prior on synthetic data sampled from *evolving* structural causal models. On our 14-dataset benchmark it does *not* close the deterioration gap: mean $\delta\mathrm{OOD} = -1.4$ pp across all datasets and -2.2 pp on the harmful subset, with PACS, CIFAR-10, and Waterbirds made measurably *worse* by the adapted prior, and every harmful dataset remains harmful by the same Hoeffding criterion. Full numbers are in Appendix C.5. If anything, the result argues for placing a detector ahead of the adapter, including to flag when the adapter is itself the source of the problem.

3 Background and Method

3.1 Prior-Fitted Networks

Prior-fitted networks (PFNs) [38] are tabular foundation models pretrained to perform classification in-context, without any task-specific gradient updates at deployment. During pretraining, a PFN is exposed to millions of synthetic supervised tasks drawn from a designer-specified prior $p(\theta)$ over data-generating processes. Each task contributes a labeled **support set** $\mathcal{D}^{\mathrm{sup}} = (X_{\mathrm{sup}}, y_{\mathrm{sup}})$ of n examples that serve as the in-context prompt, together with a query batch with known labels. The model is optimized by minimizing the in-context cross-entropy between its predictions on the query and the ground-truth labels from the same generative process. After pretraining, a PFN approximates the posterior predictive distribution (PPD):

$$p(y \mid x,\, \mathcal{D}^{\mathrm{sup}}) = \int p(y \mid x, \theta)\, p(\theta \mid \mathcal{D}^{\mathrm{sup}})\, \mathrm{d}\theta, \tag{1}$$

amortizing Bayesian inference into a single forward pass over the sequence $(X_{\mathrm{sup}}, y_{\mathrm{sup}}, x)$.

We view TabPFN and TabICL as PFN instantiations specialized to tabular classification and regression, with priors built from structural causal models (SCMs), Bayesian neural networks, and (for TabICL) tree-like inductive biases via XGBoost-fitted SCM edges. Both represent the current frontier in tabular supervised learning and are the empirical subjects of this work, though we note that the following exposition applies to the broader PFN family [36, 3, 53].

3.2 Failure Monitoring of PFNs for Classification

PFN priors are largely ID: for each pretraining task, support and query sets are sampled from the same generative model. Thus, they do not account for queries whose covariates are OOD with respect to the support, which can produce post-hoc performance gaps beyond acceptable ranges. Detecting this requires a computable signal that distinguishes ID from OOD batches, from which we may build a hypothesis test. Per the model disagreement literature [12, 41, 60, 54], it is easier to force a model to disagree with its own predictions on OOD than on ID data; classical-paradigm methods such as DETECTRON [12] and D3M [41] exploit this by perturbing the weights of a fixed converged classifier. In the following, we show how forcing model disagreement can be done for PFNs and extract a relevant signal that helps distinguish ID and OOD, building the foundations of **CALD**.

3.3 In-Context Model Disagreement Optimization

Let us consider how disagreement could be maximized in the ICL paradigm with PFNs. Denote the ID distribution by p_0 and let $\mathcal{D}^{\mathrm{sup}} = (X_{\mathrm{sup}}, y_{\mathrm{sup}}) \sim p_0^n$ be a size-n ID support. We partition the queries into two disjoint roles:

Anchor $(X_{\mathrm{anc}}, y_{\mathrm{anc}}) \sim p_0$ with $X_{\mathrm{anc}} \cap X_{\mathrm{sup}} = \emptyset$. A held-out set of *labeled* ID samples. Their true labels y_{anc} are known and serve as a reference ID signal.

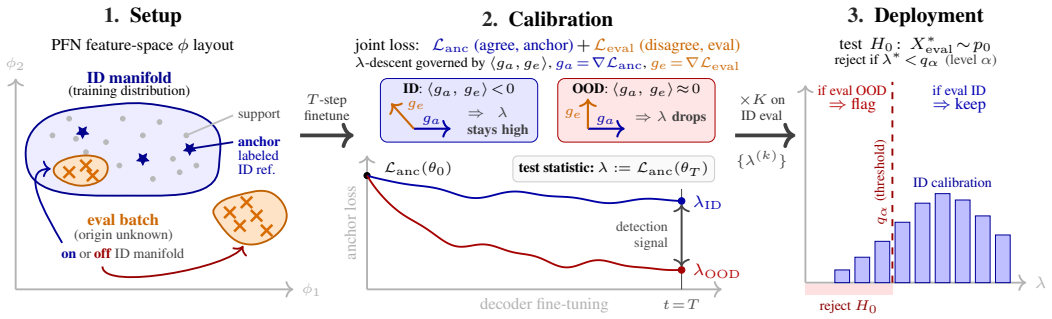


Figure 1: **CALD in three steps. (1) Setup.** In the PFN's feature space, a labeled *anchor* (★) sits inside the *ID manifold*; the unlabeled *eval* batch (✕) arrives with unknown origin — it may sit **on** or **off** the ID manifold. **(2) Calibration.** $\mathcal{L}_{\text{anc}}$ provides a stable ID reference; $\mathcal{L}_{\text{eval}}$ pulls the decoder off its eval pseudolabels. A short T-step finetune optimizes their sum, and the angle between g_a, g_e reveals the eval's origin: on ID, the gradients fight ($\langle g_a, g_e \rangle < 0$; λ stays high); off ID, they are nearly orthogonal (≈ 0; λ drops). The post-finetune anchor loss $\lambda := \mathcal{L}_{\text{anc}}(\theta_T)$ is the test statistic. **(3) Deployment.** Repeating step 2 across K known-ID eval batches yields a calibration record; reject H_0: $X^*_{\text{eval}} \sim p_0$ (flag OOD) when λ^* falls below the α-quantile q_α (red region).

Eval $(X_{\text{eval}}) \sim p_1 \stackrel{?}{=} p_0$. A batch of *unlabeled* query points whose distributional origin is unknown at test time, the target of our disagreement forcing.

Let q_θ denote a PFN. Conditional on $\mathcal{D}^{\text{sup}}$, its forward pass returns logits and predictions:

$$z \in \mathbb{R}^C, \quad q_\theta(\cdot \mid x, \mathcal{D}^{\text{sup}}) := \text{softmax}(z) \in \Delta^{C-1}, \quad \hat{y} = \text{argmax}_y \, q_\theta(y \mid x, \mathcal{D}^{\text{sup}}) \tag{2}$$

for each $x \in X_{\text{anc}} \cup X_{\text{eval}}$, where C is the number of classes. Write $\hat{y}_{\text{eval}}$ for predictions (pseudolabels) the model assigns to X_{eval}.

To force q_θ to disagree with itself on X_{eval} while remaining in agreement with the true labels from the anchor, we compute the cross-entropy (CE) loss rewarding correct predictions for the anchor:

$$\mathcal{L}_{\text{anc}}(\theta) = \frac{1}{|X_{\text{anc}}|} \sum_{(x,y) \in (X_{\text{anc}}, y_{\text{anc}})} - \log q_\theta(y|x, \mathcal{D}^{\text{sup}}) \tag{3}$$

and the Disagreement Cross-Entropy (DCE) loss [12] for the eval set, rewarding the model for placing low probability on its own initial pseudolabels:

$$\mathcal{L}_{\text{eval}}(\theta) = \frac{1}{|X_{\text{eval}}|} \sum_{(x,y) \in (X_{\text{eval}}, \hat{y}_{\text{eval}})} \left[\frac{1}{C-1} \sum_{c \neq y} - \log q_\theta(c \mid x, \mathcal{D}^{\text{sup}}) \right] \tag{4}$$

If minimizing the CE loss bumps logits for the correct class up, minimizing the DCE loss pushes logits for the pseudolabel class down. Performing gradient descent on the joint objective

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{anc}}(\theta) + \beta \, \mathcal{L}_{\text{eval}}(\theta) \tag{5}$$

with disagreement weight $\beta = 1$ throughout unless otherwise stated (Appendix D.3 ablates it), thus has the effect of forcing q_θ to disagree on the eval set while remaining performant on ID samples.

3.4 CALD: Calibrated Anchor-Loss Disagreement

Finetuning q_θ under (5) exerts two competing pressures on θ: $\mathcal{L}_{\text{anc}}$ pulls the decoder toward the true labels on the anchor, while $\mathcal{L}_{\text{eval}}$ pushes it away from its own pretrained pseudolabels on the eval batch. The driving observation behind **CALD** is the way these pressures resolve.

The *post-finetune anchor loss* $\lambda := \mathcal{L}_{\text{anc}}(\theta_T)$ is systematically *lower* when X_{eval} is OOD than when X_{eval} is ID (Figure 1, Panel 2). **CALD** turns this one-dimensional signal into a calibrated one-sided test by collecting a null record of λ values under known-ID conditions and rejecting at deployment when the observed λ^* falls into the lower tail of that record. We defer justification to Section 4.

Given hyperparameters T (finetune steps), K (calibration rounds), η (learning rate), and α (test level), **CALD** proceeds in two phases (full pseudocode in Appendix A.1):

1. **Calibration phase.** For each round $k = 1, \ldots, K$, draw an ID eval batch $X_{\text{eval}}^{(k)} \sim p_0$, freeze its pseudolabels $\hat{y}_{\text{eval}}^{(k)}$ assigned by q_{θ_0} to $X_{\text{eval}}^{(k)}$, and reset $\theta \leftarrow \theta_0$. At each of T gradient steps, the PFN forward-passes the joint sequence $(X_{\text{sup}}, y_{\text{sup}}, X_{\text{anc}}, X_{\text{eval}}^{(k)})$: the anchor logits supply $\mathcal{L}_{\text{anc}}(\theta)$ against the true labels y_{anc}, the eval logits supply $\mathcal{L}_{\text{eval}}(\theta)$ against the frozen pseudolabels $\hat{y}_{\text{eval}}^{(k)}$, and the joint loss is descended via $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta)$ with only the decoder trainable. Record the post-finetune anchor loss $\lambda^{(k)} := \mathcal{L}_{\text{anc}}(\theta_T)$. After K rounds, we have an empirical null record $\{\lambda^{(k)}\}_{k=1}^{K}$ under known-ID conditions.
2. **Deployment phase.** On a deployment batch X_{eval}^* of unknown origin, freeze its pseudolabels $\hat{y}_{\text{eval}}^*$, reset $\theta \leftarrow \theta_0$, and run the same T-step finetune, forward-passing $(X_{\text{sup}}, y_{\text{sup}}, X_{\text{anc}}, X_{\text{eval}}^*)$ at each step and recording the deployment statistic $\lambda^* := \mathcal{L}_{\text{anc}}(\theta_T)$. The verdict is a one-sided test against the calibration record:

$$\lambda^* < \text{Quantile}_\alpha\left(\lambda^{(1)}, \ldots, \lambda^{(K)}\right) \implies \text{declare OOD.} \tag{6}$$

The support set $\mathcal{D}^{\text{sup}}$ and anchor $(X_{\text{anc}}, y_{\text{anc}})$ are held fixed across both phases.

4 Theory

This section justifies the headline of §3.4: that $\lambda_{\text{OOD}} < \lambda_{\text{ID}}$ in expectation, and that the resulting one-sided test has controlled Type-I error. §4.1 gives an informal mechanistic argument, §4.2 states the formal results, and §4.3 flags one related claim (monotone scaling of the signal with harm severity) that we support empirically but do not prove.

4.1 Intuitive argument

The two terms of $\mathcal{L}(\theta)$ impose complementary forces on the shared decoder, and the way they resolve depends on whether the pretrained PFN is confident on its eval pseudolabels. When $X_{\text{eval}} \sim p_0$, the PFN places high probability on its own (correct) pseudolabels, so $\nabla\mathcal{L}_{\text{eval}}$ is large in magnitude and points in a direction that conflicts with $\nabla\mathcal{L}_{\text{anc}}$; forcing disagreement on in-distribution points actively interferes with maintaining anchor agreement, and the anchor loss stays elevated after T steps. When X_{eval} is shifted, however, the PFN is uncertain on the same pseudolabels: $\nabla\mathcal{L}_{\text{eval}}$ is small in magnitude and approximately orthogonal to $\nabla\mathcal{L}_{\text{anc}}$, so the disagreement pressure does not impede anchor descent and the anchor loss drops below its ID counterpart. The post-finetune $\lambda := \mathcal{L}_{\text{anc}}(\theta_T)$ therefore summarizes in a single scalar whether the eval-side optimization pressure can co-exist with anchor agreement, with $\mathbb{E}[\lambda_{\text{OOD}}] < \mathbb{E}[\lambda_{\text{ID}}]$.

4.2 Theoretical guarantees

We summarize the two main results that justify **CALD**; full statements, assumptions, and proofs are deferred to Appendix A.

The first result establishes that the post-finetune anchor loss is, in expectation, smaller under harmful shift than under in-distribution eval. It rests on a gradient-geometry analysis of the joint objective (5): the disagreement gradient on ID eval interferes with anchor descent (driving the anchor loss up by a structural amount $\gamma > 0$), while on far-OOD eval it becomes approximately orthogonal to the anchor gradient (controlled by a parameter ϵ that vanishes as $\|p_1 - p_0\|_{\text{TV}} \to 1$).

Theorem 4.1 (Main separation, simplified)**.** *Let $\mathbb{E}[\lambda_{\text{ID}}]$ and $\mathbb{E}[\lambda_{\text{OOD}}]$ denote the expected post-finetune anchor losses when the eval batch is drawn from p_0 and p_1 respectively. Under the assumptions stated in Appendix A, there exist constants $\gamma > 0$ and $\epsilon \geq 0$ (the ID interference strength and the OOD feature-orthogonality parameters, respectively) such that*

$$\mathbb{E}[\lambda_{\text{ID}}] - \mathbb{E}[\lambda_{\text{OOD}}] \;\geq\; T\eta\,(\gamma - \epsilon\, B) \;-\; O(T^2\eta^2), \tag{7}$$

where B collects bounded architecture- and data-dependent constants. Whenever $\epsilon < \gamma/B$ and $T\eta$ is in the small-step regime, the right-hand side is strictly positive and $\mathbb{E}[\lambda_{\text{OOD}}] < \mathbb{E}[\lambda_{\text{ID}}]$.

The separation gap grows linearly in the finetune budget $T\eta$ (within the small-step regime), is driven by a structural ID quantity γ that depends only on the pretrained classifier and p_0, and admits its severity dependence through ϵ. The full version (Theorem A.9) makes the constants explicit.

It is important to acknowledge that Theorem 4.1 certifies detection in the *far-OOD regime only*. At finite shift severity its condition $\epsilon < \gamma/B$ is not satisfied on any of the ten benchmarks on which we estimated it. Outside the far-OOD regime we therefore make no formal claim and instead measure the proof's operative premise directly: the time-averaged gradient-interference gap $\Delta_{\mathrm{traj}} := \mathbb{E}_{\mathrm{OOD}}[\langle \bar{g}_a, \bar{g}_e \rangle] - \mathbb{E}_{\mathrm{ID}}[\langle \bar{g}_a, \bar{g}_e \rangle]$ is positive on nine of the ten datasets, and the two cells where it is weakest or negative are exactly the two with near-zero λ-separation (§5.4, Figure 3(b); Table 18). These measurements support the theorem's mechanism at moderate shift severity and help predict where the detector will struggle. A useful extension for future work would be to derive a finite-ϵ bound using constants estimated from data instead of worst-case values.

The second result certifies the practical decision rule. The detector observes only K calibration anchor losses and a single deployment statistic, and rejects ID when the deployment value falls below the empirical α-quantile of the calibration set. A standard exchangeability argument controls the false-alarm rate.

Proposition 4.2 (Finite-sample Type-I error control, simplified)**.** *Suppose the anchor set, support set, and finetune hyperparameters are fixed across rounds, the eval batches are i.i.d. draws, and the inner finetune loop is deterministic given its inputs. Under the null $H_0 : X^*_{\mathrm{eval}} \sim p_0$, the one-sided test in* (6) *satisfies*

$$\mathbb{P}_{H_0}(\textit{reject } H_0) \;\leq\; \alpha + \frac{1}{K+1}. \tag{8}$$

The slack $1/(K+1)$ is the cost of using a finite calibration set; it vanishes as $K \longrightarrow \infty$. Combining Theorem 4.1 with Proposition 4.2 gives a detector with controlled Type-I error and a non-vanishing power gap whenever the OOD shift is severe enough to satisfy the feature-orthogonality regime; we expand on the corresponding power analysis in Appendix A.11.

Proposition 4.2 requires the calibration and deployment batches to be exchangeable under the null hypothesis. Here, "in-distribution" means that the batches come from the same covariate distribution; no geometric boundary between ID and OOD data is needed (Definition A.10, Appendix A.11). The hyperparameters must also be fixed in advance so that calibration and deployment use the same scoring procedure. The per-cell learning-rate selection in Table 1 (Appendix C.4) uses both ID and known OOD trajectories, so the guarantee does not cover that procedure. A fixed learning rate avoids this dependence on known OOD data and appears practical: it gives similar average detection performance, while observed no-shift false-alarm rates remain close to the theoretical bound across the tested settings (see Appendix D.5).

4.3 Calibration to harm severity is empirical

Theorem 4.1 delivers a separation in the far-OOD regime, but it does not predict that the detection signal scales *monotonically* with the magnitude of the shift or with the model's accuracy drop. We observe this calibration-to-harm property empirically (Figure 2): on a controlled 2D Gaussian sweep and on a Waterbirds severity sweep, AUROC rises smoothly from 0.5 at zero shift to 1.0 at catastrophic shift, while small benign translations remain near chance; large benign translations do fire, as Lemma A.12 predicts for any feature-space monitor. We treat this as a desirable structural property of the gradient cross-term $\langle g_a, g_e \rangle$ tracking shift severity; we explore it in §5.3.

5 Experiments

We evaluate **CALD** against post-hoc OOD-detection baselines on 14 datasets and the two PFN architectures (TabPFN and TabICL). Our experiments are designed to support three claims: (i) **CALD** reliably flags harmful distribution shift, (ii) it is markedly more resilient to benign-shift false alarms than label-blind baselines, though not immune to them, and (iii) the gradient mechanism of §4 is the one operating in the deployed models. Extended experimental details, dataset descriptions, and

baselines are deferred to Appendix C; hyperparameter sweeps, deployment cost, design choices, the frozen deployment configuration, anchor robustness, feature collapse, a two-signal harm detector, and detectability across shift types are ablated in Appendix D. The code repository for **CALD** is available at `https://github.com/teivng/cald`.

5.1 Setup

Datasets: we evaluate on 14 datasets covering two regimes. (i) *Raw tabular shift* (7 datasets): UCI Heart Disease, Diabetes Readmission, Adult Income, Bank Marketing, Credit Default, Covertype, and Folktables ACS Income; and (ii) *structured shift via foundation-model representations* (7 datasets): Camelyon17, CIFAR-10$\rightarrow$10.1, PACS, Waterbirds, CelebA, CivilComments, and Amazon Reviews, where image and text inputs are encoded by pretrained foundation models (DINOv2, ResNet-18, Jina-v3) and the resulting representations are passed to TabPFN/TabICL as ordinary tabular features. Per-dataset descriptions, ID/OOD splits, and a survey of the resulting accuracy deteriorations across the severity spectrum (catastrophic, moderate, negligible, benign) are given in Appendices C.1 and C.3. **Baselines:** we compare against 15 post-hoc OOD detection signals computed without finetuning: MSP, Entropy, Energy, Mahalanobis distance, Relative Mahalanobis (RMD), KNN distance, ViM, BBSD-softmax, BBSD-embedding, MMD, H-divergence, GradNorm, GradOrth, GAIA-Z, and WeiPer. Detailed descriptions are in Appendix C.2.

PFN models. We apply **CALD** to two SoTA PFN models: **TabPFN v2.5** [16] and **TabICL v2** [47].

Evaluation metrics. We report **TPR@5% FPR** (primary), the true-positive rate at a fixed 5% false-positive operating point set by the empirical $\alpha = 0.05$ quantile of the calibration anchor-loss distribution, and **AUROC** (secondary).

5.2 Main Results: Deterioration Detection

Table 1 reports TPR@5%FPR for **CALD** and the 15 post-hoc baselines across 14 datasets and both PFN architectures. All results use 10 seeds, 50 calibration rounds, $T = 50$ finetune steps, and eval batches of 128 points; the full table with standard deviations and the learning-rate selection protocol are deferred to Appendix C.4.

CALD reliably flags harmful shift. Across the seven harmful columns, the better **CALD** variant scores TPR ≥ 0.99 on CivilComments, Camelyon, Waterbirds, CIFAR, PACS, and Covertype, whereas for UCI Heart, both variants reach only 0.71 TPR at the chosen calibration size.

CALD is more resilient to benign shifts than baselines, but not immune. On four of seven benign datasets the better **CALD** variant achieves correct-silence rate ≥ 0.92 (Adult, Amazon, Folktables-TabPFN, CelebA-TabICL); on Credit Default, TabICL recovers 0.76 silence where TabPFN false-alarms at 0.00. The honest weaknesses are Bank (≤ 0.02), Diabetes Readmission (≤ 0.08), and Folktables under TabICL (0.42). All three have small accuracy gaps ($|\Delta| \leq 4.4\%$) close to the Hoeffding floor, and Appendix B dissects the mechanism: the post-finetune anchor loss is a feature-space quantity, sensitive to any reweighting of the input domain, whereas the harm label is determined entirely by reweighting on the model's mistake region (Lemma A.12). When a benign shift moves substantial mass on the complement of the mistake region, the gradient geometry registers a real distributional difference and **CALD** fires.

TabICL outperforms TabPFN, especially on subpopulation shift. TabICL strictly dominates TabPFN on four datasets whose shift is defined at the subgroup level: Camelyon ($0.19 \rightarrow 1.00$), Waterbirds ($0.46 \rightarrow 0.99$), CelebA ($0.27 \rightarrow 0.97$), and Credit Default ($0.00 \rightarrow 0.76$). On eight of the other ten datasets the two architectures track within 0.05; Folktables ($0.98 \rightarrow 0.42$) is the one cell where TabPFN dominates. The likely culprit is decoder capacity: TabICL's trainable decoder (Linear $512 \rightarrow 1024 \rightarrow C$) is roughly an order of magnitude larger than TabPFN's ($192 \rightarrow 384 \rightarrow C$), and on subtle group-level shifts the disagree-finetune dynamics need enough decoder expressivity to separate ID from OOD anchor descent. TabPFN's tighter decoder appears to bottleneck the signal whenever the shift does not register as a global feature-space displacement.

No baseline matches CALD's harm/benign trade-off. D-MMD, H-divergence, and KNN each score ≥ 0.95 on a majority of harmful columns but ≤ 0.10 on a majority of benign columns: their headline harmful-side TPR comes from *always alarming*, not from harm-calibration. RMD shows the opposite trade-off: benign-side silence comparable to **CALD**, but harmful-side TPR collapsing to

Table 1: **Detection performance (TPR@5% FPR) across 14 shifts ordered by harm.** Columns are sorted by ID→OOD accuracy drop Δ (second header row, %). Harmful/benign split via Hoeffding bound at $\delta = 0.05$, $N = \min(\text{rounds} \times \text{eval}, |\mathcal{D}_{\text{OOD}}|) = 1{,}280$ from the deterioration survey of Appendix C.3: $\Delta^* \approx 4.8\%$. For *harmful* shifts ($\Delta > \Delta^*$), TPR = rate of correct alarms. For *benign* shifts ($\Delta \leq \Delta^*$), TPR = rate of correct silence (1 − false-alarm rate). A good detector scores high in *both* regions.

	Benign ($\Delta \leq 4.8\%$)							*Harmful* ($\Delta > 4.8\%$)						
Method	Adult	Diab.	Amazon	Credit	Folktbl.	Bank	CelebA	CivilC.	Camel.	Waterb.	CIFAR	UCI	PACS	Covert.
Δ	−9.4	−2.3	+1.5	+2.8	+3.1	+4.4	+4.5	+7.4	+8.9	+11.6	+12.4	+15.6	+27.9	+83.5
MSP	.00	.64	.78	.55	.53	.00	.63	**1.00**	.97	**1.00**	.96	.74	**1.00**	**1.00**
Entropy	.00	.61	.80	.61	.50	.00	.39	**1.00**	.98	**1.00**	.97	.77	**1.00**	**1.00**
Energy	.00	.00	.77	.00	.86	**.42**	.04	**1.00**	**1.00**	**1.00**	.97	.66	**1.00**	.50
Mahal.	.00	.00	.71	.00	.12	.05	.00	.32	.43	.18	.55	**1.00**	**1.00**	**1.00**
RMD	.69	.59	.87	**.92**	.72	.05	.42	.51	.40	.52	.67	.47	**1.00**	**1.00**
KNN	.00	.00	.78	.00	.23	.00	.00	**1.00**	.95	.97	.10	.99	**1.00**	**1.00**
ViM	.00	.10	.87	.00	.63	.13	.81	.23	.54	.18	.15	.93	.70	**1.00**
BBSD-S	.00	.40	.75	.76	.71	.02	.64	.61	.94	.88	.89	.81	**1.00**	**1.00**
BBSD-E	.00	.00	.76	.67	.66	.00	.49	.58	**1.00**	.99	.72	**1.00**	**1.00**	**1.00**
D-MMD	.00	.00	.66	.00	.00	.00	.54	.91	**1.00**	**1.00**	.68	**1.00**	**1.00**	**1.00**
H-Div	.00	.00	.19	.00	.10	.32	.00	.70	**1.00**	**1.00**	.50	**1.00**	**1.00**	**1.00**
GradNorm	.00	.44	.85	.01	.78	.27	.60	**1.00**	.92	.98	.91	.54	.98	.70
GradOrth	.83	**.77**	.83	.69	.86	.02	.46	.66	.54	**1.00**	.95	.23	.97	**1.00**
GAIA-z	.00	.07	—	.01	.80	.29	—	—	—	—	—	.41	—	.59
WeiPer	.18	.17	.89	.00	.67	.17	.81	.34	.45	.26	.21	.89	.62	.73
CALD (TabPFN)	**1.00**	.08	**.92**	.00	**.98**	.02	.27	1.00	.19	.46	**1.00**	.71	**1.00**	**1.00**
CALD (TabICL)	**1.00**	.00	.87	.76	.42	.00	**.97**	.99	**1.00**	.99	.99	.71	**1.00**	**1.00**

0.40–0.67 on the moderate shifts (Camelyon, Waterbirds, CIFAR). **CALD** is the only detector in the table that simultaneously achieves high harmful-side TPR on a majority of datasets and benign-side silence on a majority of datasets. We emphasize that this is a relative claim, not an absolute one: **CALD** does false-alarm on a minority of benign cells (Bank and Diabetes for both PFNs; Credit, Folktables, and CelebA for one of the two), and we view this remaining benign sensitivity as the realistic ceiling of a feature-space monitor that is blind to test labels.

5.3 Calibration to harm

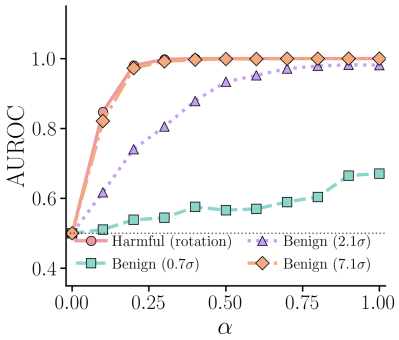
A deterioration monitor is only useful if its alarm scales with the *harm* a shift inflicts rather than with how novel the shifted inputs look. While Theorem 4.1 certifies a separation in the far-OOD limit, it does not determine how the statistic λ behaves as severity grows (§4.3). We therefore ask directly whether λ tracks severity, and whether it stays silent on shifts of comparable feature-space magnitude that do not hurt the model.

We sweep shift severity in two settings. The first is a controlled 2D Gaussian in which harm and novelty can be dialed independently: rotating the class boundary is harmful, while translating the inputs along the boundary is benign, and both are matched in feature-space magnitude. The second is real subpopulation shift on Waterbirds, where the fraction α of spuriously-correlated minority samples in the eval batch sets the accuracy drop.

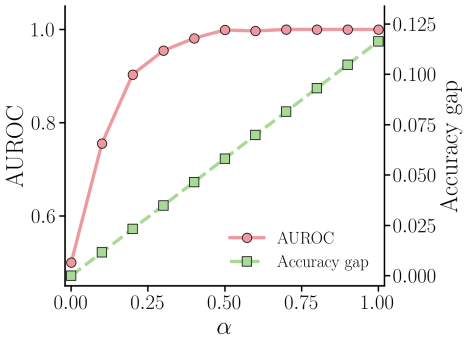
In Figure 2, on both the controlled and the real sweep, AUROC sits at chance when there is no shift, rises monotonically with the accuracy gap, and saturates at the most severe setting. On the benign axis, a small translation stays near chance throughout, while larger translations fire. This suggests that within these controlled regimes, λ tracks severity, and distinguishes harm from novelty only when the benign shift is small, which is the property that separates **CALD** from the always-alarming baselines in Table 1. The property is empirical and should only be treated as such, as on some real-world benign datasets the post-finetune anchor loss separates ID from OOD despite a small accuracy gap, and Lemma A.12 shows why no feature-space monitor can be perfectly harm-calibrated; we refer the reader to Appendix B.7 for a closer look.

5.4 The gradient cross-term

The proof of Theorem 4.1 tracks the anchor loss along the T finetuning steps and asks how much of its descent the eval-disagreement term interferes with, under ID versus OOD eval. That interference is carried by the inner product $\langle g_a, g_e \rangle$ between the anchor-agreement gradient $g_a = \nabla \mathcal{L}_{\text{anc}}$ and the

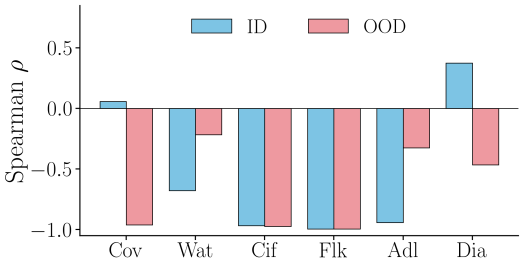


(a) Synthetic 2D Gaussian.

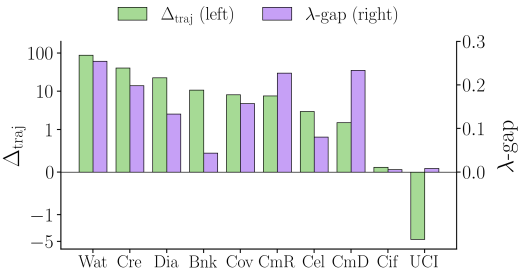


(b) Real (Waterbirds, α sweep).

Figure 2: **Detection AUROC scales with shift severity. CALD**'s AUROC rises monotonically with the induced accuracy gap on both controlled synthetic shift (left) and real subpopulation shift (right). At zero shift severity the AUROC is at chance; at maximum severity it saturates at 1.0. On the synthetic sweep, a small benign translation (0.7σ) stays near chance, while larger benign translations (2.1σ, 7.1σ) fire.



(a) Per-step Spearman ρ between $\cos(g_a, g_e)$ and $\Delta\mathcal{L}_{\text{anc}}$, ID vs. OOD; negative ρ is the prediction of Equation (9).



(b) The proof's premise Δ_{traj} of Equation (10) (left axis, symlog) and the λ-gap $\mathbb{E}[\lambda_{\text{ID}}] - \mathbb{E}[\lambda_{\text{OOD}}]$ it produces (right axis), per dataset; TabPFN, fixed $\eta = 10^{-3}$.

Figure 3: **The gradient cross-term governs anchor descent, and its ID/OOD gap is what the detector reads.** (a) The first-order law holds in both regimes; the two near-zero or positive ID bars (Covertype, Diabetes Readmission) are cells where the cosine is pinned at its fighting extreme with no spread, see text. (b) The premise $\Delta_{\text{traj}} > 0$ holds on nine of ten datasets, and the two cells where it is weakest or negative (CIFAR-10, UCI Heart) are exactly the two with near-zero λ-gap. *Datasets:* Cov=Covertype, Wat=Waterbirds, Cif=CIFAR-10, Flk=Folktables, Adl=Adult Income, Dia=Diabetes Readmission, Cre=Credit Default, Bnk=Bank Marketing, CmR/CmD=Camelyon17 (ResNet18/DINOv2), Cel=CelebA, UCI=UCI Heart.

eval-disagreement gradient $g_e = \nabla\mathcal{L}_{\text{eval}}$. Expanding one step of the joint update to first order,

$$\Delta\mathcal{L}_{\text{anc}}^{(t)} \;\approx\; -\eta\,\|g_a^{(t)}\|^2 \;-\; \eta\,\langle g_a^{(t)}, g_e^{(t)}\rangle, \tag{9}$$

so the anchor loss descends at the vanilla rate $-\eta\|g_a\|^2$ corrected by the cross-term. Theorem 4.1 makes two claims about it. First, that the cross-term is what *governs* anchor descent step by step, so that the first-order analysis describes the regime the detector actually runs in. Second, that its typical *value* differs by regime: strongly negative on ID eval, where the disagreement gradient fights anchor descent and holds λ up, and near zero on OOD eval, where the two gradients decouple and λ falls. We examine each in turn.

The cross-term governs anchor descent. We record $\cos(g_a^{(t)}, g_e^{(t)})$ and $\Delta\mathcal{L}_{\text{anc}}^{(t)}$ at every finetune step and compute the Spearman correlation between them per dataset, separately for ID and OOD eval. Equation (9) predicts a *negative* correlation: the more aligned the two gradients, the faster the anchor loss falls.

In Figure 3(a), the correlation is negative for every dataset under OOD eval, and for every dataset but two under ID eval. The two exceptions, Covertype and Diabetes Readmission under ID eval, are cells where the cosine is pinned near its fighting extreme across every step with the anchor loss rising throughout, which is the ID behavior the theorem describes; with no spread in the cosine there is nothing for a rank correlation to pick up. This suggests that the first-order law holds on all six datasets measured, and that it holds equally under ID and OOD eval. The law itself therefore does not separate the two regimes; what separates them must be the *value* of the cross-term, which we measure next.

The proof's premise. Theorem 4.1 lower-bounds the separation $\mathbb{E}[\lambda_{\mathrm{ID}}] - \mathbb{E}[\lambda_{\mathrm{OOD}}]$ by $T\eta(\gamma - \epsilon B)$, and the proof obtains this bound from the premise that by averaging the gradients over the T steps (Proposition A.8), the cross-term is more negative under ID eval than under OOD eval. Writing $\bar{g}_a$, $\bar{g}_e$ for the time-averaged gradients, the premise is

$$\Delta_{\mathrm{traj}} \;:=\; \mathbb{E}_{\mathrm{OOD}}\big[\langle \bar{g}_a, \bar{g}_e \rangle\big] \;-\; \mathbb{E}_{\mathrm{ID}}\big[\langle \bar{g}_a, \bar{g}_e \rangle\big] \;>\; 0, \tag{10}$$

and the constants γ and ϵ are the theorem's worst-case way of guaranteeing it. The certifying condition $\epsilon < \gamma/B$ turns out not to hold on any of the ten benchmarks on which we estimate it (Appendix D.8, Table 17): at finite severity, none of them is far-OOD in the theorem's sense. We therefore measure the premise itself, computing Δ_{traj} by exact full-decoder autograd along the finetuning trajectories, alongside the λ-gap $\mathbb{E}[\lambda_{\mathrm{ID}}] - \mathbb{E}[\lambda_{\mathrm{OOD}}]$ it is supposed to produce.

In Figure 3(b), Δ_{traj} is positive on nine of the ten datasets, including the moderate-overlap benchmarks on which the formal condition is vacuous. The two datasets on which it is weakest or negative, CIFAR-10 and UCI Heart, are also the two with a near-zero λ-gap under this configuration, so the measurement anticipates the detector's weakest cells rather than rationalizing its successes after the fact. This suggests that while Theorem 4.1 certifies detection in the far-OOD regime only, the gradient mechanism it is built on is present across the near-OOD benchmarks and is what drives detection there. We refer the reader to Appendix D.8 for the full analysis, including the estimated constants of the condition and the per-dataset λ-gap (Table 18).

6 Conclusion and Limitations

We introduced **CALD**, a deployment-time deterioration monitor for tabular PFNs that turns a brief decoder finetune of the frozen model into a detection signal. We ground our method in theory for harmful shift, and observe that across 14 datasets, **CALD** matches or outperforms 15 post-hoc baselines on harmful shifts, giving the most favorable harm/benign trade-off among the methods compared (Table 1).

The theory, however, is lacking on the benign side. Lemma A.12 essentially states that the accuracy gap depends only on mass moved on the model's mistake region M, while the post-finetune λ integrates the gradient cross-term over the whole input domain, so a feature-space monitor *cannot* achieve perfect harm-calibration. Empirically this surfaces on Bank, Diabetes, Credit, Folktables, and CelebA (each under at least one of the two PFNs in Table 1), benign shifts that nonetheless move substantial mass off M; Appendix B dissects the tabular cases. The conformal test also relies on ID exchangeability, so pure label shift, concept shift, open-set or adversarial inputs, and engineering errors are out of scope. Closing the harm gap likely requires a complementary signal tied to the density-ratio integral on M, sketched in Proposition A.21 and left to future work.

Concretely, a tabular PFN's impending failure leaves a mechanistic trace in the gradient geometry of a brief decoder fine-tune, reachable from a frozen pretrained model with no extra training data. The limitations above bound this particular feature-space readout. **CALD** thus probes this trace through the decoder, while PFNs expose gradients with respect to the support set itself. Probes that exploit this, such as measuring how disagreement-labeled deployment points placed in the support affect predictions on the anchor, are a natural direction for richer, genuinely in-context deterioration sensors.

Acknowledgements and Disclosure of Funding

RGK gratefully acknowledges support from the Canada Research Chairs Program (CRC-2022-00049) and the Canada CIFAR AI Chairs Program. This research was funded in part by an NFRF Special Call Award (NFRFR-2022-00526), an NSERC Discovery Grant (RGPIN-2022-04546) and a CIHR Team Grant (RN549848–534930). HB is supported by Swedish Research Council Grant 2022-04748, as well as the Sweden-America Foundation. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, John Schulman, and Dan Mané. Concrete problems in AI safety. 2016.

[3] Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Benson Li, Junwei Ma, Jesse C. Cresswell, and Rahul G. Krishnan. CausalPFN: Amortized causal effect estimation via in-context learning, 2025. URL `https://arxiv.org/abs/2506.07918`.

[4] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2): 550–560, 2018.

[5] Sima Behpour, Thang Doan, Xin Li, Wenbin He, Liang Gou, and Liu Ren. GradOrth: A simple yet efficient out-of-distribution detection with orthogonal projection of gradients. In *Advances in Neural Information Processing Systems*, 2023.

[6] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79: 151–175, 2010.

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

[8] Jinggang Chen, Junjie Li, Xiaoyang Qu, Jianzong Wang, Jiguang Wan, and Jing Xiao. GAIA: Delving into gradient-based attribution abnormality for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2023.

[9] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[10] Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

[11] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[12] Tom Ginsberg, Zhongyuan Liang, and Rahul G. Krishnan. A learning based hypothesis test for harmful covariate shift. In *International Conference on Learning Representations*, 2023.

[13] Maximilian Granz, Manuel Heurich, and Tim Landgraf. WeiPer: OOD detection using weight perturbations of class projections. In *Advances in Neural Information Processing Systems*, 2024.

[14] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

[15] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data? In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[16] Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, et al. Tabpfn-2.5: Advancing the state of the art in tabular foundation models. *arXiv preprint arXiv:2511.08667*, 2025.

[17] Kai Helli, David Schnurr, Noah Hollmann, Samuel Müller, and Frank Hutter. Drift-resilient TabPFN: In-context learning temporal distribution shifts on tabular data. In *Advances in Neural Information Processing Systems*, 2024.

[18] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

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

[20] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, 2022.

[21] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

[22] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

[23] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems*, 2021.

[24] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, 2021.

[25] Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J. Zico Kolter. Assessing generalization of SGD via disagreement. In *International Conference on Learning Representations*, 2022.

[26] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, 2017.

[27] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

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

[29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision*, 2017.

[30] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, 2020.

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

[32] Zachary C. Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, 2018.

[33] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning*, 2020.

[34] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2020.

[35] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

[36] Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Hamidreza Kamkari, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L. Caterini, and Maksims Volkovs. TabDPT: Scaling tabular foundation models on real data, 2024. URL `https://arxiv.org/abs/2410.18164`.

[37] John P. Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, 2021.

[38] Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. *arXiv preprint arXiv:2112.10510*, 2021.

[39] Thomas Nagler. Statistical foundations of prior-data fitted networks. In *International Conference on Machine Learning*, 2023.

[40] Preetum Nakkiran, Behnam Neyshabur, and Hanie Sedghi. Deep bootstrap framework: Good online learners are good offline generalizers. In *International Conference on Learning Representations*, 2021.

[41] Viet Nguyen, Changjian Shui, Vijay Giri, Siddharth Arya, Amol Verma, Fahad Razak, and Rahul G Krishnan. Reliably detecting model failures in deployment without labels. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[42] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, 2022.

[43] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[44] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019.

[45] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 2018.

[46] Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *International Conference on Machine Learning*, 2025.

[47] Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabiclv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

[48] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. Dataset shift in machine learning. *MIT Press*, 2009.

[49] Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*, 2019.

[50] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 classifiers generalize to CIFAR-10? 2018.

[51] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, 2019.

[52] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to Mahalanobis distance for improving near-OOD detection. *arXiv preprint arXiv:2106.09022*, 2021.

[53] Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-PFN: In-context learning for causal effect estimation, 2025. URL `https://arxiv.org/abs/2506.06039`.

[54] Elan Rosenfeld and Saurabh Garg. (almost) provable error bounds under distribution shift via disagreement discrepancy. In *Advances in Neural Information Processing Systems*, 2023.

[55] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

[56] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[57] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, 2022.

[58] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, 2020.

[59] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, 2020.

[60] Alexandru Ţifrea, Eric Stavarache, and Fanny Yang. Semi-supervised novelty detection using ensembles with regularized disagreement. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022.

[61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[62] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[63] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

[64] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-distribution with virtual-logit matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[65] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[66] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

[67] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 2024.

[68] Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei Koh, and Chelsea Finn. Wild-time: A benchmark of in-the-wild distribution shift over time. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[69] Shengjia Zhao, Abhishek Sinha, Yutong He, Aidan Perreault, Jiaming Song, and Stefano Ermon. Comparing distributions by measuring differences that affect decision making. In *International Conference on Learning Representations*, 2022.

A Theory

This appendix develops the theoretical foundations behind **CALD**. We first state the algorithm in pseudocode, then build up the formal argument that the post-finetuning anchor loss λ is, in expectation, lower on OOD eval batches than on ID eval batches.

A.1 Algorithm

We restate **CALD** in pseudocode for completeness; the verbal description in Section 3.4 suffices for the reader who does not need step-by-step detail.

Algorithm 1: Calibrated Anchor-Loss Disagreement (**CALD**)

Input: Support set $\mathcal{D}^{\text{sup}}$, anchor set $(X_{\text{anc}}, y_{\text{anc}})$, ID distribution p_0, steps T, rounds K, learning rate η, level α
Output: OOD verdict for test batch X^*_{eval}

`// -- Phase 1: Calibration on known-ID data --`
for $k = 1, \ldots, K$ **do**
  Sample $X^{(k)}_{\text{eval}} \sim p_0$
  $\hat{y}^{(k)} \leftarrow \arg\max\ q_{\theta_0}\Big(y \mid X^{(k)}_{\text{eval}},\, \mathcal{D}^{\text{sup}}\Big)$ `// pseudolabels (no grad)`
  $\theta \leftarrow \theta_0$
  for $t = 1, \ldots, T$ **do**
    $\ell_{\text{anc}} \leftarrow \mathcal{L}_{\text{CE}}(q_\theta(y \mid X_{\text{anc}},\, \mathcal{D}^{\text{sup}}),\ y_{\text{anc}})$
    $\ell_{\text{dis}} \leftarrow \mathcal{L}_{\text{DCE}}\Big(q_\theta(y \mid X^{(k)}_{\text{eval}},\, \mathcal{D}^{\text{sup}}),\ \hat{y}^{(k)}\Big)$
    $\theta \leftarrow \theta - \eta\, \nabla_\theta(\ell_{\text{anc}} + \ell_{\text{dis}})$
  end
  $\lambda^{(k)} \leftarrow \mathcal{L}_{\text{CE}}(q_\theta(y \mid X_{\text{anc}},\, \mathcal{D}^{\text{sup}}),\ y_{\text{anc}})$ `// post-finetune anchor loss`
end

`// -- Phase 2: Detection on test batch --`
$\hat{y}^* \leftarrow \arg\max\ q_{\theta_0}(y \mid X^*_{\text{eval}},\, \mathcal{D}^{\text{sup}})$
$\theta \leftarrow \theta_0$
for $t = 1, \ldots, T$ **do**
  $\ell_{\text{anc}} \leftarrow \mathcal{L}_{\text{CE}}(q_\theta(y \mid X_{\text{anc}},\, \mathcal{D}^{\text{sup}}),\ y_{\text{anc}})$
  $\ell_{\text{dis}} \leftarrow \mathcal{L}_{\text{DCE}}(q_\theta(y \mid X^*_{\text{eval}},\, \mathcal{D}^{\text{sup}}),\ \hat{y}^*)$
  $\theta \leftarrow \theta - \eta\, \nabla_\theta(\ell_{\text{anc}} + \ell_{\text{dis}})$
end
$\lambda^* \leftarrow \mathcal{L}_{\text{CE}}(q_\theta(y \mid X_{\text{anc}},\, \mathcal{D}^{\text{sup}}),\ y_{\text{anc}})$

`// -- Verdict --`
if $\lambda^* < \text{Quantile}_\alpha\big(\lambda^{(1)}, \ldots, \lambda^{(K)}\big)$ **then**
  return OOD
else
  return ID
end

A.2 Justification of the Detection Signal

We develop a formal argument that the post-finetuning anchor loss λ is, in expectation, *lower* on OOD eval batches than on ID eval batches. The argument is a first-order gradient-geometry analysis of the joint objective in (5), combined with the tilted-attention interpretation of transformer PFNs due to Nagler [39]. We state the main proposition as a one-step result, then extend to the full T-step trajectory used in Algorithm 1 using a standard small-step regime. The separation $\mathbb{E}[\lambda_{\text{OOD}}] < \mathbb{E}[\lambda_{\text{ID}}]$ is proven in the far-OOD regime (where the feature-orthogonality assumption holds with small parameter ϵ; see Theorem A.9). Conversely, in the near-OOD regime the signed direction is a heuristic prediction that is empirically supported by a severity sweep rather than a theorem.

The argument formalizes the informal intuition in Section 3.4 that *forcing disagreement on ID points tugs against anchor agreement, while forcing disagreement on OOD points does not*. We isolate the Frobenius inner product $\langle g_a, g_e \rangle_F$ of the anchor and disagreement gradients at initialization as the key quantity of interest: it controls whether anchor descent is impeded or free. We then show that $\langle g_a, g_e \rangle_F$ is controlled by a feature-space kernel $\langle \phi(x_a), \phi(x_e) \rangle$, and that this kernel collapses in the OOD limit under Nagler's exponential-tilting interpretation of attention.

The appendix builds up in six stages:

1. *Algebra* (Sections A.3–A.4): we write out the last-layer gradients of CE and DCE and observe that they factor as outer products of a label-space residual and a backbone-feature vector. This reduces the gradient cross-term to a sum of products of two scalar quantities (Lemma A.1).
2. *First-step geometry* (Sections A.5–A.7): a short Taylor expansion yields the anchor-loss decrement after one step as $-\eta \|g_a\|_F^2 - \eta \langle g_a, g_e \rangle_F$ (Proposition A.2). The sign and magnitude of the cross-term is then analyzed in two regimes: upper-bounded under a feature-orthogonality assumption (Corollary A.4), and strictly negative under an ID-confidence assumption (Proposition A.6).
3. *PFN theory* (Section A.8): we use Nagler [39, Theorems 5.4 & 6.3] to argue that the feature-orthogonality assumption is a structural consequence of transformer attention in the *far-OOD* limit. The key object is an exponentially tilted measure over the support set that collapses to the untilted marginal as $\|p_1 - p_0\|_{\mathrm{TV}} \to 1$.
4. *Main separation results* (Sections A.9–A.12): multi-step lift (Proposition A.8), the Main Separation Theorem that composes all preceding pieces into $\mathbb{E}[\lambda_{\mathrm{OOD}}] < \mathbb{E}[\lambda_{\mathrm{ID}}]$ (Theorem A.9), a conformal-prediction argument establishing Type-I error control (Proposition A.11), and a multiclass extension.
5. *Harm characterization* (Section A.13): we extend Nagler's tilting from features to predictions, derive a density-ratio decomposition of the accuracy gap Δ, and introduce a second signal $\hat{\mu}$ that estimates Δ directly (Proposition A.21). This bridges the gap between distributional-difference detection and predictive-harm detection.
6. *Summary of assumptions* (Section A.14): We summarize our main assumptions for both the main separation theorem as well as the harm detection extension.

Gradient interpretation. Two gradients live in a large parameter space $\mathbb{R}^{C \times d}$. The anchor gradient g_a points in the direction that, if walked along, most rapidly increases the probability assigned to the true labels on the anchor set. The disagreement gradient g_e points in the direction that most rapidly *reduces* the probability assigned to the (frozen) pseudo-labels on the eval set. Gradient descent moves in the direction $-(g_a + g_e)$. The question 'does this step also improve the anchor loss?' therefore reduces to: does $-(g_a + g_e)$ have positive inner product with the direction $-g_a$ that purely improves anchor loss? This inner product is $\|g_a\|_F^2 + \langle g_a, g_e \rangle_F$, and its sign determines whether the disagreement gradient aligns with anchor descent or interferes with it.

A.3 Setup and Notation

We document the formal setup in two pieces. We start with an important design choice that the subsequent analysis depends on. Then, we introduce the notation used throughout this appendix.

Why the analysis trains only the last layer. Our analysis models the finetune as acting on the last linear (or small MLP) layer with every other parameter frozen; the deployed detector updates the whole decoder, and Appendix D.2 shows the two are empirically indistinguishable. This restriction is what makes our analysis tractable. A frozen backbone provides a fixed feature map ϕ, and all gradient computations reduce to linear algebra in the last-layer parameter space. The per-sample gradient becomes a rank-one outer product of a label-space residual and a feature vector, which factorizes the gradient cross-term into a label $\times$ feature product.

Notation. Let $\phi : \mathcal{X} \times (\mathcal{X} \times \mathcal{Y})^n \to \mathbb{R}^d$ denote the PFN's backbone map, which takes a query x and a support set $\mathcal{D}^{\mathrm{sup}} = \{(X_i, Y_i)\}_{i=1}^n$ and returns a d-dimensional representation. We write $\phi(x) := \phi(x; \mathcal{D}^{\mathrm{sup}})$ when the support is held fixed; all arguments in this appendix treat $\mathcal{D}^{\mathrm{sup}}$ as a fixed conditioning set. Concretely, $\phi(x)$ is the transformer's penultimate activation at query x, with the support-set context incorporated through attention. The trainable parameter is the last-layer

weight matrix $W \in \mathbb{R}^{C\times d}$, which absorbs the corresponding bias by the usual one-augmentation of ϕ (i.e., appending a constant 1-coordinate). Predictions take the form

$$q_W(y \mid x) = \mathrm{softmax}(W\phi(x))_y. \tag{11}$$

All other PFN parameters are frozen at their pretrained values θ_0, so ϕ does not depend on W. Write W_0 for the initial decoder (the decoder weights of the pretrained PFN) and $p_x := q_{W_0}(\cdot \mid x) \in \Delta^{C-1}$ for the initial predicted distribution at query x. The initial prediction p_x is the PFN's usual in-context prediction before any finetuning.

Let $(X_{\mathrm{anc}}, y_{\mathrm{anc}}) = \{(x_a^{(i)}, y_a^{(i)})\}_{i=1}^{n_a}$ be the labeled ID anchor set and $X_{\mathrm{eval}} = \{x_e^{(j)}\}_{j=1}^{n_e}$ the unlabeled eval batch. Pseudo-labels $\hat{y}^{(j)} := \arg\max_c p_{x_e^{(j)}}(c)$ are frozen at initialization and do not change during finetuning (*cf.* the pseudolabel step of Algorithm 1). Define the per-sample losses

$$\ell_{\mathrm{CE}}(W; x, y) = -\log q_W(y \mid x), \tag{12}$$

$$\ell_{\mathrm{DCE}}(W; x, \hat{y}) = -\frac{1}{C-1}\sum_{c\neq\hat{y}} \log q_W(c \mid x), \tag{13}$$

and the batch losses

$$\mathcal{L}_{\mathrm{anc}}(W) = \frac{1}{n_a}\sum_{i=1}^{n_a} \ell_{\mathrm{CE}}(W; x_a^{(i)}, y_a^{(i)}), \qquad \mathcal{L}_{\mathrm{dis}}(W) = \frac{1}{n_e}\sum_{j=1}^{n_e} \ell_{\mathrm{DCE}}(W; x_e^{(j)}, \hat{y}^{(j)}). \tag{14}$$

The finetuning objective is $\mathcal{L}(W) = \mathcal{L}_{\mathrm{anc}}(W) + \mathcal{L}_{\mathrm{dis}}(W)$. The post-finetuning anchor loss, which is the detection statistic of interest, is

$$\lambda := \mathcal{L}_{\mathrm{anc}}(W_T), \qquad W_{t+1} = W_t - \eta\,\nabla_W \mathcal{L}(W_t). \tag{15}$$

A.4 Last-Layer Gradient Geometry

Gradient of CE. We derive (17) explicitly so that the reader can see why the outer-product structure emerges. Write the pre-softmax logits as $z := W\phi(x) \in \mathbb{R}^C$, so $q_W(c \mid x) = e^{z_c}/\sum_{c'} e^{z_{c'}}$. The per-sample CE loss is $\ell_{\mathrm{CE}} = -\log q_W(y \mid x) = -z_y + \log\sum_{c'} e^{z_{c'}}$. The gradient with respect to z is the standard softmax residual

$$\frac{\partial \ell_{\mathrm{CE}}}{\partial z_c} = q_W(c \mid x) - \mathbb{1}\{c = y\} = \big(q_W(\cdot \mid x) - e_y\big)_c. \tag{16}$$

To push this back to W, use the chain rule: $\partial z_c / \partial W_{c',d'} = \phi_{d'}(x)\mathbb{1}\{c = c'\}$, so

$$\frac{\partial \ell_{\mathrm{CE}}}{\partial W_{c,d}} = \big(q_W(c \mid x) - \mathbb{1}\{c = y\}\big)\,\phi_d(x).$$

Reading the right-hand side as the (c, d) entry of an outer product between the vector $q_W(\cdot \mid x) - e_y \in \mathbb{R}^C$ and the vector $\phi(x) \in \mathbb{R}^d$ yields

$$\nabla_W \ell_{\mathrm{CE}}(W; x, y) = \big(q_W(\cdot \mid x) - e_y\big)\,\phi(x)^\top. \tag{17}$$

This is a rank-one $C \times d$ matrix: the first factor lives in label space and records the per-class prediction error; the second factor lives in feature space and records 'which direction in W to push'.

Gradient of DCE. The disagreement cross-entropy $\ell_{\mathrm{DCE}}(W; x, \hat{y}) = -\frac{1}{C-1}\sum_{c\neq\hat{y}} \log q_W(c \mid x)$ is itself a cross-entropy, against the uniform-on-non-pseudolabel target $u_{\hat{y}} \in \Delta^{C-1}$ defined by $(u_{\hat{y}})_c = \mathbb{1}\{c \neq \hat{y}\}/(C-1)$. Indeed,

$$\mathrm{CE}\big(q_W(\cdot \mid x),\, u_{\hat{y}}\big) = -\sum_c (u_{\hat{y}})_c \log q_W(c \mid x) = -\tfrac{1}{C-1}\sum_{c\neq\hat{y}} \log q_W(c \mid x) = \ell_{\mathrm{DCE}}.$$

Applying the same chain-rule calculation as for CE, with target $u_{\hat{y}}$ replacing e_y,

$$\nabla_W \ell_{\mathrm{DCE}}(W; x, \hat{y}) = \big(q_W(\cdot \mid x) - u_{\hat{y}}\big)\,\phi(x)^\top. \tag{18}$$

This gradient is again rank-one and shares the feature factor $\phi(x)$ with the CE gradient in (17).

Batch gradients. Both gradients are rank-one matrices in $\mathbb{R}^{C\times d}$. Evaluating at initialization, write

$$r_a^{(i)} := p_{x_a^{(i)}} - e_{y_a^{(i)}}, \qquad r_e^{(j)} := p_{x_e^{(j)}} - u_{\hat{y}^{(j)}} \tag{19}$$

for the anchor- and eval-residual vectors in $\mathbb{R}^C$. $r_a^{(i)}$ measures the prediction error on anchor point i in label space, while $r_e^{(j)}$ is the corresponding direction in which the disagreement loss pushes the prediction away from the pseudolabel on eval point j. The initial batch gradients are the averaged rank-one matrices

$$g_a := \nabla_W \mathcal{L}_{\text{anc}}(W_0) = \frac{1}{n_a}\sum_{i=1}^{n_a} r_a^{(i)}\,\phi(x_a^{(i)})^\top, \tag{20}$$

$$g_e := \nabla_W \mathcal{L}_{\text{dis}}(W_0) = \frac{1}{n_e}\sum_{j=1}^{n_e} r_e^{(j)}\,\phi(x_e^{(j)})^\top. \tag{21}$$

Why we care about $\langle g_a, g_e\rangle_F$. The joint update moves W in the direction $-(g_a+g_e)$; whether this also descends the anchor loss depends on the alignment between $-(g_a+g_e)$ and the pure anchor-descent direction $-g_a$. Specifically, the first-order change in anchor loss is $\eta\langle g_a, g_a+g_e\rangle_F = \eta(\|g_a\|_F^2 + \langle g_a, g_e\rangle_F)$, a sum of a non-negative self-term and the gradient cross-term $\langle g_a, g_e\rangle_F$ whose sign determines whether the disagreement gradient aligns with (*sign* ≥ 0) or interferes with (*sign* < 0) anchor descent. Proposition A.2 makes this Taylor-level statement precise. Before that, it is useful to see that this inner product has a simple structure: it decomposes into a weighted sum over anchor-eval *pairs*, where each pair contributes a product of a label-space scalar and a feature-space scalar. This factorization is what makes the rest of the analysis tractable.

Lemma A.1 (Gradient cross-term factorization)**.** *The Frobenius inner product of g_a and g_e factorizes as*

$$\langle g_a,\, g_e\rangle_F = \frac{1}{n_a n_e}\sum_{i=1}^{n_a}\sum_{j=1}^{n_e} C_{ij}\,K_{ij}, \tag{22}$$

where $C_{ij} := \langle r_a^{(i)}, r_e^{(j)}\rangle \in [-2,2]$ is a label-space residual inner product and $K_{ij} := \langle \phi(x_a^{(i)}), \phi(x_e^{(j)})\rangle$ is a feature-space kernel.

Proof. We first establish the general outer-product identity. Let $A = u\,v^\top \in \mathbb{R}^{C\times d}$ and $B = u'\,v'^\top \in \mathbb{R}^{C\times d}$ be rank-one matrices with $u, u' \in \mathbb{R}^C$ and $v, v' \in \mathbb{R}^d$. The Frobenius inner product is

$$\begin{aligned}\langle A, B\rangle_F &= \sum_{c=1}^{C}\sum_{d'=1}^{d} A_{c,d'}\,B_{c,d'} = \sum_{c,d'}(u_c v_{d'})(u'_c v'_{d'}) \\ &= \Big(\sum_c u_c\,u'_c\Big)\Big(\sum_{d'} v_{d'}\,v'_{d'}\Big) = \langle u, u'\rangle\,\langle v, v'\rangle.\end{aligned}$$

In words: the Frobenius inner product of two rank-one matrices splits into the product of the two 'left' inner products and the two 'right' inner products. This is why decoder-only updates make the analysis tractable.
Now expand the cross-term using bilinearity:

$$\begin{aligned}\langle g_a, g_e\rangle_F &= \Big\langle \tfrac{1}{n_a}\sum_i r_a^{(i)}\phi(x_a^{(i)})^\top, \tfrac{1}{n_e}\sum_j r_e^{(j)}\phi(x_e^{(j)})^\top\Big\rangle_F \\ &= \tfrac{1}{n_a n_e}\sum_{i,j}\big\langle r_a^{(i)}\phi(x_a^{(i)})^\top, r_e^{(j)}\phi(x_e^{(j)})^\top\big\rangle_F \\ &= \tfrac{1}{n_a n_e}\sum_{i,j}\langle r_a^{(i)}, r_e^{(j)}\rangle\,\langle\phi(x_a^{(i)}), \phi(x_e^{(j)})\rangle = \tfrac{1}{n_a n_e}\sum_{i,j} C_{ij}\,K_{ij},\end{aligned}$$

where the third equality applies the rank-one identity to each pair (i,j).

For the bound $|C_{ij}| \le 2$: by Cauchy–Schwarz, $|C_{ij}| \le \|r_a^{(i)}\|_2 \, \|r_e^{(j)}\|_2$. Each residual is a difference of two distributions in Δ^{C-1}, so $\|r_a^{(i)}\|_2 = \|p_{x_a} - e_{y_a}\|_2 \le \sqrt{2}$ (the maximum squared ℓ_2 distance between two probability vectors on C classes is 2, attained when both are point masses on different classes). Hence $|C_{ij}| \le \sqrt{2} \cdot \sqrt{2} = 2$. ∎

Equation (22) is the basis of the analysis. The entire gradient interaction between the anchor and eval batches collapses onto the elementwise product of a label-space matrix $C \in \mathbb{R}^{n_a \times n_e}$ and a feature-space Gram matrix $K \in \mathbb{R}^{n_a \times n_e}$. The label part C_{ij} is bounded, easy to interpret, and typically negative under ID (we will see this in Proposition A.6). The feature part K_{ij} is the part that 'knows' whether the eval batch is OOD: as we will show in Section A.8, OOD queries produce feature vectors that collapse to a single input-independent direction, making K_{ij} small under OOD.

A.5 First-Step Anchor Descent

Why a first-step analysis is meaningful. Algorithm 1 takes $T = 50$ gradient steps per round at a learning rate selected from $\{10^{-2}, 10^{-3}, 10^{-4}\}$ (default 10^{-3}). One might worry that the first step tells us nothing about the cumulative effect of fifty. However, there are two insights that make the first-step analysis informative. First, in the small-step regime (Assumption A.7) the cumulative anchor-loss decrement after T steps is a linear combination of first-step quantities up to $O(T^2\eta^2)$ corrections (Proposition A.8). Second, the sign of the first-step cross-term is a discrete feature that typically persists along the trajectory when the backbone is frozen, because the feature-space kernel K_{ij} is the same at every step. This means that the sign analysis at step one carries to step T under mild regularity conditions.

Proposition A.2 (First-step anchor descent)**.** *Let $W_1 = W_0 - \eta(g_a + g_e)$ be the update after one gradient step on the joint loss $\mathcal{L}$. Then*

$$\mathcal{L}_{\text{anc}}(W_1) - \mathcal{L}_{\text{anc}}(W_0) = -\eta \, \|g_a\|_F^2 - \eta \, \langle g_a, \, g_e \rangle_F + R(\eta), \tag{23}$$

with remainder $|R(\eta)| \le \frac{1}{2} L_a \eta^2 \|g_a + g_e\|_F^2$, where L_a is the Lipschitz constant of $\nabla \mathcal{L}_{\text{anc}}$ along the segment $[W_0, W_1]$.

Proof. The loss $\mathcal{L}_{\text{anc}} : \mathbb{R}^{C \times d} \to \mathbb{R}$ is twice continuously differentiable, being a composition of the affine map $W \mapsto W\phi(x)$ with the softmax-cross-entropy function, both of which are smooth. On the bounded segment $[W_0, W_1] = \{W_0 + t(W_1 - W_0) : t \in [0, 1]\}$, $\nabla \mathcal{L}_{\text{anc}}$ is L_a-Lipschitz for some finite L_a; we bound L_a explicitly in the proof of Proposition A.8.
Step 1 (Taylor expansion with integral remainder). For any twice continuously differentiable $f : \mathbb{R}^{C \times d} \to \mathbb{R}$ and any W_0, W_1,

$$\begin{aligned} f(W_1) = & f(W_0) \, + \, \langle \nabla f(W_0), \, W_1 - W_0 \rangle_F \\ & + \int_0^1 (1 - t) \, \big\langle W_1 - W_0, \, \nabla^2 f(W_0 + t(W_1 - W_0))[W_1 - W_0] \big\rangle_F \, dt. \end{aligned}$$

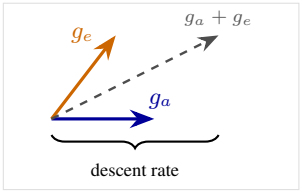
Applying this to $f = \mathcal{L}_{\text{anc}}$ and denoting the integral as $R(\eta)$ gives

$$\mathcal{L}_{\text{anc}}(W_1) - \mathcal{L}_{\text{anc}}(W_0) = \langle \nabla \mathcal{L}_{\text{anc}}(W_0), \, W_1 - W_0 \rangle_F + R(\eta).$$

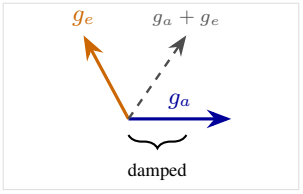
Step 2 (compute the linear term). By definition $\nabla \mathcal{L}_{\text{anc}}(W_0) = g_a$, and the update direction is $W_1 - W_0 = -\eta(g_a + g_e)$. Hence

$$\begin{aligned} \langle \nabla \mathcal{L}_{\text{anc}}(W_0), \, W_1 - W_0 \rangle_F &= \langle g_a, \, -\eta(g_a + g_e) \rangle_F \\ &= -\eta \langle g_a, g_a \rangle_F - \eta \langle g_a, g_e \rangle_F \\ &= -\eta \|g_a\|_F^2 - \eta \langle g_a, g_e \rangle_F, \end{aligned}$$

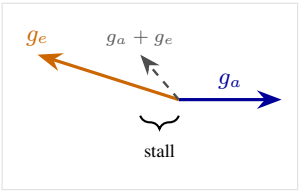
using bilinearity of the Frobenius inner product.



(a) **Cooperation.** $\langle g_a, g_e \rangle_F \geq 0$. Descent rate $\geq \eta \|g_a\|_F^2$ (OOD regime).



(b) **Damped.** $\langle g_a, g_e \rangle_F \in (-\|g_a\|_F^2, 0)$. Descent rate $< \eta \|g_a\|_F^2$ (typical ID regime).



(c) **Stall.** $\langle g_a, g_e \rangle_F \leq -\|g_a\|_F^2$. Descent rate ≤ 0 (extreme ID regime).

Figure 4: **The three regimes of Proposition A.2.** The first-order anchor-loss decrement after one joint-loss step equals η times the projection of $g_a + g_e$ (dashed) onto the direction of g_a (blue). When g_e (orange) makes an acute angle with g_a **(left)**, descent is unimpeded or better; this is the OOD regime (Corollary A.4). When g_e makes an obtuse but bounded angle **(middle)**, descent is damped but still positive (the typical ID regime, Proposition A.6). When g_e is strongly anti-aligned with g_a **(right)**, the projection turns negative and anchor loss does not decrease. The detector's signal comes from the ID regime sitting in the middle panel and the OOD regime sitting approximately on the left.

Step 3 (bound the remainder). By L_a-Lipschitzness of $\nabla \mathcal{L}_{\text{anc}}$ on $[W_0, W_1]$, the Hessian has operator norm $\leq L_a$ on the segment. Plugging this into the integral remainder,

$$|R(\eta)| \leq \int_0^1 (1-t)\, L_a \, \|W_1 - W_0\|_F^2 \, dt$$
$$= L_a \, \eta^2 \|g_a + g_e\|_F^2 \int_0^1 (1-t)\, dt = \tfrac{1}{2} L_a \, \eta^2 \|g_a + g_e\|_F^2.$$

Combining steps 1–3 gives (23). ∎

Interpretation: the cross-term controls the first-order descent rate. The significance of Proposition A.2 is that the first-order anchor descent rate is *entirely* controlled by the sign and magnitude of the cross-term $\langle g_a, g_e \rangle_F$. Fix η small enough such that the remainder $R(\eta) = O(\eta^2)$ is negligible compared to the first-order terms of size $O(\eta)$. Then

$$\mathcal{L}_{\text{anc}}(W_1) - \mathcal{L}_{\text{anc}}(W_0) \approx -\eta\big(\|g_a\|_F^2 + \langle g_a, g_e \rangle_F\big).$$

Since $\|g_a\|_F^2 \geq 0$, the three qualitatively distinct regimes are (see Figure 4):

1. **Alignment or orthogonality** $\langle g_a, g_e \rangle_F \geq 0$: anchor loss decreases at rate $\geq \eta \|g_a\|_F^2$. The disagree gradient either actively helps anchor descent (positive inner product) or is geometrically unrelated to it (zero inner product). *This is the regime we will argue for OOD eval batches.*
2. **Damped descent** $\langle g_a, g_e \rangle_F \in (-\|g_a\|_F^2, 0)$: anchor loss still decreases, but at a rate strictly smaller than $\eta \|g_a\|_F^2$. The disagree gradient partially cancels the anchor gradient. *This is the typical ID regime.*
3. **Stall or increase** $\langle g_a, g_e \rangle_F \leq -\|g_a\|_F^2$: anchor loss stalls or increases. The disagree gradient more than cancels the anchor gradient. *This is an extreme ID regime in which anchor and eval are particularly entangled.*

The informal claim that the disagreement gradient 'interferes with anchor agreement' in the ID regime but not in the OOD regime translates to this cross-term being negative under ID and near-zero under OOD. The next two subsections make each direction rigorous: Corollary A.4 upper-bounds $|\langle g_a, g_e \rangle_F|$ under an OOD feature-orthogonality assumption, while Proposition A.6 lower-bounds it by a negative quantity under an ID confidence assumption.

A.6 Feature-Orthogonality Corollary

Corollary A.4 below gives an upper bound on $|\langle g_a, g_e \rangle_F|$ when anchor and eval features are approximately orthogonal. This happens whenever $\cos(\phi(x_a), \phi(x_e))$ is small.

Assumption A.3 (Feature orthogonality)**.** *There exists $\epsilon \geq 0$ such that, for every anchor-eval pair (i, j),*

$$\left|\langle \phi(x_a^{(i)}),\, \phi(x_e^{(j)})\rangle\right| \leq \epsilon \, \|\phi(x_a^{(i)})\| \, \|\phi(x_e^{(j)})\|. \tag{24}$$

The parameter $\epsilon \in [0, 1]$ is the worst-case cosine between anchor and eval features. When $\epsilon = 0$ the two feature sets lie in orthogonal subspaces; when $\epsilon = 1$ they are identical up to scale.

Corollary A.4 (Orthogonal features imply unimpeded anchor descent)**.** *Assume $\|\phi(x)\| \leq B_\phi$ for all x in $X_{\text{anc}} \cup X_{\text{eval}}$ and set $B_C := \max_{i,j} |C_{ij}| \leq 2$ (from Lemma A.1). Under Assumption A.3,*

$$\left|\langle g_a,\, g_e\rangle_F\right| \leq \epsilon \, B_\phi^2 \, B_C. \tag{25}$$

Consequently, when $\epsilon = o\big(\|g_a\|_F^2/(B_\phi^2 B_C)\big)$, the first-step anchor descent in (23) *reduces to $-\eta\|g_a\|_F^2(1 + o(1))$, i.e., the disagreement gradient leaves anchor descent unimpeded at first order in the step size η.*

Proof. Start from the factorization (22). For each pair (i, j), bound the summand by the product of term-wise bounds:

$$\begin{aligned} |C_{ij}K_{ij}| &= |C_{ij}| \left|\langle \phi(x_a^{(i)}), \phi(x_e^{(j)})\rangle\right| \\ &\leq B_C \cdot \epsilon \, \|\phi(x_a^{(i)})\| \, \|\phi(x_e^{(j)})\| \qquad \text{(Assumption A.3; definition of } B_C\text{)} \\ &\leq B_C \cdot \epsilon \, B_\phi^2 \qquad \text{(boundedness of } \phi\text{)}. \end{aligned}$$

Averaging over the $n_a n_e$ pairs in (22) (the triangle inequality applied to the sum),

$$\left|\langle g_a, g_e\rangle_F\right| = \left|\tfrac{1}{n_a n_e} \sum_{i,j} C_{ij}K_{ij}\right| \leq \tfrac{1}{n_a n_e} \sum_{i,j} |C_{ij}K_{ij}| \leq \tfrac{1}{n_a n_e} \cdot n_a n_e \cdot \epsilon B_\phi^2 B_C = \epsilon B_\phi^2 B_C.$$

For the consequence, plug this bound into (23):

$$\mathcal{L}_{\text{anc}}(W_1) - \mathcal{L}_{\text{anc}}(W_0) = -\eta\|g_a\|_F^2 \left(1 + \tfrac{\langle g_a, g_e\rangle_F}{\|g_a\|_F^2}\right) + O(\eta^2),$$

and observe that the ratio $\langle g_a, g_e\rangle_F / \|g_a\|_F^2$ is $o(1)$ exactly when $\epsilon = o(\|g_a\|_F^2/(B_\phi^2 B_C))$. ∎

Intuitive reading. When anchor and eval features lie in approximately orthogonal subspaces of $\mathbb{R}^d$, moving the decoder in the direction $-g_e$ (which is a sum of rank-one outer products with eval features) has little effect on the decoder's behavior on anchor features. The disagreement gradient is, to first order, *invisible* to the anchor gradient. Therefore, the joint update $-(g_a + g_e)$ acts on the anchor loss as if only $-g_a$ had been applied, and anchor loss descends at the usual rate $\eta\|g_a\|_F^2$.

A.7 ID Interference

The OOD corollary of the previous subsection showed that $\langle g_a, g_e\rangle_F$ is *small* when features are orthogonal. But to get a detector we need separation: the cross-term must not just be zero but *negative* under ID, so anchor descent is visibly damped. This subsection establishes that separation.

Why this is not obvious. A casual reader might expect that any gradient that is not orthogonal to g_a contributes random noise. However, the claim here is stronger: under the mild ID conditions listed below, the sign of $\langle g_a, g_e\rangle_F$ is *systematically* negative. This asymmetry between the OOD regime (cross-term ≈ 0) and the ID regime (cross-term < 0) is the source of the detector's signal: $\mathbb{E}[\lambda_{\text{OOD}}] - \mathbb{E}[\lambda_{\text{ID}}]$ is bounded above by $-\eta\,\mathbb{E}_{p_0 \otimes p_0}[\langle g_a, g_e\rangle_F] < 0$ at first order, where $p_0 \otimes p_0$ denotes the product distribution under which the anchor pair (x_a, y_a) and the eval pair $(x_e, \hat{y})$ are drawn independently, each from p_0.

Assumption A.5 (ID confidence, correctness, and class symmetry)**.** *Recall from Section A.3 that $p_x := q_{W_0}(\cdot \mid x) \in \Delta^{C-1}$ is the initial PFN prediction at query x, and the pseudolabel function is $\hat{y}(x) := \arg\max_c p_x(c)$. Under sampling $(x_a, y_a), (x_e, \hat{y}) \sim p_0 \otimes p_0$, and for a constant $\tau < 1/(2(C-1))$:*

(i) The initial model is non-trivially correct on the anchor: $\mathbb{E}_{p_0}[p_{x_a}(y_a)] \geq 1/2$.

(ii) The initial model is non-trivially confident on eval: $\mathbb{E}_{p_0}[p_{x_e}(\hat{y}(x_e))] \geq 1 - \tau$.

(iii) Backbone features are p_0-centered and non-degenerate: $\mathbb{E}_{p_0}[\phi(x)] = 0$ and $\mathbb{E}_{p_0}[\|\phi(x)\|^2] > 0$.

(iv) (Class symmetry) The class-conditional feature means $\mu_c := \mathbb{E}_{p_0}[\phi(x) \mid y(x) = c]$ span C distinct directions, and the predictor satisfies $p_x(c) - 1/C \propto \langle \phi(x), \mu_c \rangle$ up to first order.

(v) (Accuracy) The initial classifier is accurate on ID: $\mathbb{P}_{p_0}(\hat{y}(x) = y(x)) \geq 1 - \tau$.

Condition (iii) can be arranged by subtracting the p_0-mean from ϕ; we revisit this when connecting to Nagler's limiting feature in Section A.8. Condition (iv) is the class-symmetry assumption that makes the label-disagreement case reverse the sign of the feature covariance; it is standard in linearized-network analyses. We state the proposition for general C but prove it first in the illustrative binary case, then give the multiclass extension.

Proposition A.6 (ID interference)**.** *Under Assumption A.5,*

$$\mathbb{E}_{p_0 \otimes p_0}\big[C_{ij}\, K_{ij}\big] < 0, \tag{26}$$

so in expectation $\langle g_a, g_e \rangle_F < 0$, and Proposition A.2 gives damped anchor descent.

Proof. We prove the binary ($C = 2$) case in full and then indicate the multiclass extension.
Step 1 (binary simplification of residuals). When $C = 2$ and the pseudolabel is $\hat{y}$, the uniform-on-non-pseudolabel distribution is $u_{\hat{y}} = e_{1-\hat{y}}$ (a one-hot on the flipped class). Hence $r_e^{(j)} = p_{x_e^{(j)}} - e_{1-\hat{y}^{(j)}}$, which is the same as the CE residual against the flipped label. For the anchor, $r_a^{(i)} = p_{x_a^{(i)}} - e_{y_a^{(i)}}$ as before.
Step 2 (binary residual inner product). Write $p_x = (1 - p_x^+, p_x^+)^\top$ with $p_x^+ := p_x(1)$, and identify $p_x^0 = 1 - p_x^+, p_x^1 = p_x^+$. A direct calculation gives

$$p_x - e_0 = (-p_x^+,\, p_x^+)^\top = -p_x^1\,(1, -1)^\top, \qquad p_x - e_1 = (1 - p_x^+,\, -(1 - p_x^+))^\top = p_x^0\,(1, -1)^\top.$$

Both residuals are scalar multiples of $e := (1, -1)^\top$. We can write this uniformly as

$$p_x - e_c = (-1)^{c+1}(1 - p_x^c)\, e.$$

Define $\alpha_x^c := (-1)^{c+1}(1 - p_x^c)$; note $|\alpha_x^c| = 1 - p_x^c$, and the sign depends on c. Then $r_a = \alpha_{x_a}^{y_a}\, e$ and $r_e = p_{x_e} - u_{\hat{y}} = p_{x_e} - e_{1-\hat{y}} = \alpha_{x_e}^{1-\hat{y}}\, e$. Hence

$$\langle r_a, r_e \rangle = \alpha_{x_a}^{y_a}\, \alpha_{x_e}^{1-\hat{y}}\, \langle e, e \rangle = 2\, \alpha_{x_a}^{y_a}\, \alpha_{x_e}^{1-\hat{y}}.$$

Step 3 (sign of each factor under Assumption A.5). Condition on y_a and $\hat{y}(x_e)$. By the definition of α:

- $\alpha_{x_a}^{y_a} = (-1)^{y_a+1}\,(1 - p_{x_a}^{y_a})$. Since $p_{x_a}^{y_a} \leq 1$, the magnitude is $1 - p_{x_a}^{y_a} \geq 0$. The *sign* is $(-1)^{y_a+1}$.
- $\alpha_{x_e}^{1-\hat{y}} = (-1)^{(1-\hat{y})+1}\,(1 - p_{x_e}^{1-\hat{y}}) = (-1)^{\hat{y}}\, p_{x_e}^{\hat{y}}$ (using $1 - p_{x_e}^{1-\hat{y}} = p_{x_e}^{\hat{y}}$). The magnitude is $p_{x_e}^{\hat{y}} = p_{x_e}(\hat{y}(x_e))$, which by Assumption A.5(ii) is $\geq 1 - \tau$. The *sign* is $(-1)^{\hat{y}}$.

Multiplying the signs:

$$\operatorname{sign}\big(\langle r_a, r_e \rangle\big) = (-1)^{y_a+1} \cdot (-1)^{\hat{y}} = -(-1)^{y_a+\hat{y}}.$$

When $y_a = \hat{y}$ (label agreement): sign $= -1$. When $y_a \neq \hat{y}$ (label disagreement): sign $= +1$.

Step 4 (sign of the feature kernel under class symmetry). Under Assumption A.5(iv) with $C = 2$, the class-conditional feature means are $\mu_0 = -\mu_1 =: -\mu$ (centering forces $\mu_0 + \mu_1 = 0$). Since the anchor draw x_a and the eval draw x_e are independent under $p_0 \otimes p_0$, conditioning on $(y_a, \hat{y}(x_e))$ makes $\phi(x_a)$ and $\phi(x_e)$ conditionally independent, so

$$\mathbb{E}\big[K_{ij} \mid y_a, \hat{y}\big] = \langle \mathbb{E}[\phi(x_a) \mid y_a],\, \mathbb{E}[\phi(x_e) \mid \hat{y}(x_e)] \rangle = \langle \mu_{y_a},\, m_{\hat{y}} \rangle,$$

where we write $m_c := \mathbb{E}[\phi(x_e) \mid \hat{y}(x_e) = c]$ for the *predicted*-class conditional mean. By Assumption A.5(v), $\mathbb{P}_{p_0}(\hat{y}(x) \neq y(x)) \leq \tau$, and a standard total-probability decomposition gives

$$m_c = \mu_c + O(\tau)$$

with the $O(\tau)$ constant controlled by the uniform backbone bound B_ϕ and the posterior $\mathbb{P}(y = c \mid \hat{y} = c) \geq 1 - O(\tau)$ implied by (v) and the law of total probability. Hence $\mathbb{E}[K_{ij} \mid y_a, \hat{y}] = \langle \mu_{y_a}, \mu_{\hat{y}} \rangle + O(\tau B_\phi^2)$, with $\langle \mu_{y_a}, \mu_{\hat{y}} \rangle = +\|\mu\|^2$ under label agreement and $-\|\mu\|^2$ under disagreement. For $\tau B_\phi^2 < \|\mu\|^2$ (denoting the regime of any useful ID classifier) the conditional feature kernel is strictly positive under label agreement and strictly negative under label disagreement.

Step 5 (combining Steps 3 and 4). The product $C_{ij}K_{ij}$ has sign equal to the product of the signs computed in Steps 3 and 4:

- Label agreement ($y_a = \hat{y}$): $C_{ij} < 0$ and $\mathbb{E}[K_{ij} \,|\, y_a = \hat{y}] > 0$, so the contribution to $\mathbb{E}[C_{ij}K_{ij}]$ is *negative*.
- Label disagreement ($y_a \neq \hat{y}$): $C_{ij} > 0$ and $\mathbb{E}[K_{ij} \,|\, y_a \neq \hat{y}] < 0$, so the contribution to $\mathbb{E}[C_{ij}K_{ij}]$ is also *negative*.

Both events contribute negatively, so $\mathbb{E}[C_{ij}K_{ij}] < 0$ (Figure 5). This proves the binary case.

Step 6 (multiclass extension). For general C, we again condition on the label pair $(y_a, \hat{y})$. Conditional independence of $\phi(x_a)$ and $\phi(x_e)$ gives

$$\mathbb{E}[K_{ij} \,|\, y_a = c, \hat{y} = c'] = \langle \mu_c, \mu_{c'} \rangle.$$

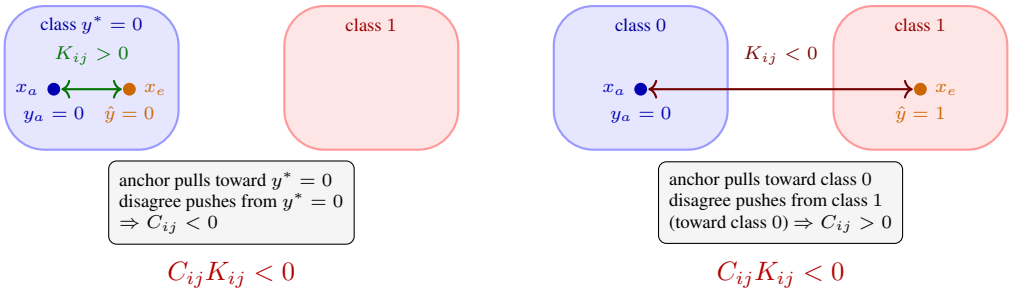
Centering ($\sum_c \mu_c = 0$) and the class-symmetry condition (iv) force $\langle \mu_c, \mu_{c'} \rangle$ to be positive when $c = c'$ and to have average value $-\|\mu\|^2/(C-1) < 0$ when $c \neq c'$. The residual inner product $\mathbb{E}[C_{ij} \,|\, y_a = c, \hat{y} = c']$ generalizes to a label-pair-indexed quantity that, under Assumption A.5(i)–(ii), is negative for $c = c'$ (the generalization of the binary '$-p_{x_e}(\hat{y})$' term) and positive for $c \neq c'$. The product $\mathbb{E}[C_{ij}K_{ij} \,|\, c, c']$ is therefore negative for both $c = c'$ and $c \neq c'$, and summing over (c, c') weighted by their joint probability completes the argument. ∎

Intuitive reading. Under label agreement, anchor and eval points tend to lie in the same class region, so their features are positively correlated ($K_{ij} > 0$), but the residual inner product is negative ($C_{ij} < 0$) because the anchor wants to push toward the correct class while the disagreement term wants to push away from it. Under label disagreement, features are negatively correlated ($K_{ij} < 0$), but the residual inner product is positive ($C_{ij} > 0$) for the mirror reason. The product $C_{ij}K_{ij}$ is negative in *both* cases, and these are the only two cases, so the expectation is negative regardless of how the labels mix. This is why the ID regime gives a strictly damped anchor descent rather than a random-sign perturbation: the mechanism aligns the two factors of the product so that the sign is consistent across both label configurations, $y_a = \hat{y}$ and $y_a \neq \hat{y}$ alike.

Contrast with OOD. Under Corollary A.4, the feature kernel K_{ij} shrinks to size $O(\epsilon)$ in the far-OOD regime, so the product $C_{ij}K_{ij}$ also shrinks to $O(\epsilon)$ regardless of the label configuration. The sign of C_{ij} no longer has a K_{ij} of matching sign to multiply with; the whole product goes to zero. Combining with Proposition A.6, we have *in the far-OOD regime (where $\epsilon B_\phi^2 B_C < \gamma$)*

$$\mathbb{E}\big[\lambda_{\mathrm{OOD}}\big] - \mathbb{E}\big[\lambda_{\mathrm{ID}}\big] \approx \eta\, \mathbb{E}\big[\langle g_a, g_e \rangle_F^{\mathrm{ID}}\big] - \eta\, \mathbb{E}\big[\langle g_a, g_e \rangle_F^{\mathrm{OOD}}\big] < 0,$$

where the first expectation is $-\gamma < 0$ by Proposition A.6 and the second is $O(\epsilon)$ by Corollary A.4. We note that for near-OOD, where ϵ is not small, the signed inequality is not rigorously established.



(a) **Label agreement** ($y_a = \hat{y}$). Anchor and eval lie in the same class region, giving $K_{ij} > 0$. The anchor gradient pulls toward that class while the disagreement gradient pushes away, giving $C_{ij} < 0$. The product $C_{ij}K_{ij}$ is therefore negative.

(b) **Label disagreement** ($y_a \neq \hat{y}$). Anchor and eval lie in different class regions, giving $K_{ij} < 0$. The anchor gradient and the disagreement gradient now both pull toward the same class, giving $C_{ij} > 0$. The product $C_{ij}K_{ij}$ is again negative.

Figure 5: **The "double negative" mechanism of Proposition A.6.** In both possible label configurations, the product $C_{ij}K_{ij}$ ends up negative: in the first case because $K > 0$ multiplies $C < 0$, in the second case because $K < 0$ multiplies $C > 0$. Since these are the only two cases, the expectation $\mathbb{E}[C_{ij}K_{ij}]$ is negative regardless of the label marginal. This is the mechanism that forces ID anchor descent into the "damped" regime of Figure 4 (middle panel).

A.8 Why Feature-Orthogonality Holds Under OOD: Nagler's Tilting

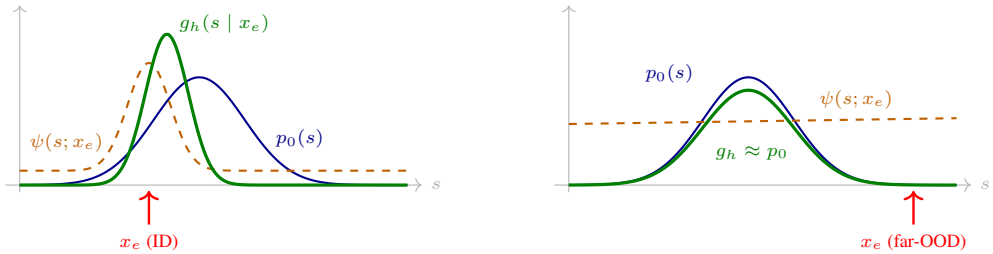
Assumption A.3 (ϵ small in the OOD regime) is the key empirical claim of this paper. We now argue that it is a structural consequence of the attention mechanism in trained transformer PFNs, using the framework of Nagler [39]. The goal of this subsection is to give a mechanistic reason why the feature vector $\phi(x_e)$ collapses to a prediction-uninformative direction when x_e is OOD.

Nagler's argument as stated concludes pointwise feature collapse: every far-OOD query maps to the same direction ϕ_∞. That conclusion depends on a few idealized assumptions: a one-layer attention mechanism (Nagler [39] assume a single layer; PFNs in practice are multi-layer) and the strict total-variation limit $\|p_1 - p_0\|_{\mathrm{TV}} \longrightarrow 1$. Hence, this behavior is best read as a limit point instead of a literal description of any specific experiment. We use the argument here to motivate Assumption A.3, which is much weaker than full collapse: it asks only that anchor and eval features be on average near-perpendicular. Note that the weaker condition can hold even when full collapse fails (OOD features fanning out within a subspace orthogonal to the span of the ID features give orthogonality without collapse), and the separation theorem (Theorem A.9) only needs the weaker condition.

What attention computes (informally). In a PFN's forward pass, the query x_e attends to the support set and produces a feature as a weighted average of transformed support tokens. The weights are softmax-normalized query-key similarities: $\alpha_j \propto \exp(v(x_e)^\top W_q\, s_j)$, where s_j is the j-th support token and $v(x_e)$ is the query's encoding. When the query resembles some support samples strongly, those attention weights dominate, and the feature is determined by a *local neighborhood* of x_e in the support set. When no support sample resembles the query, the weights are approximately uniform, and the feature becomes a *global average* over the support set. The rest of this subsection formalizes this explanation of OOD feature collapse.

Locality of trained PFNs. Nagler [39, Theorem 5.4] establishes that any PFN with vanishing bias on a sufficiently rich class $\mathcal{P}$ of ID distributions must be *local*: the prediction at a query x depends only on support samples (X_i, Y_i) with $\|X_i - x\| \leq \epsilon_n$, where $\epsilon_n \longrightarrow 0$. Labels of samples far from x can be flipped arbitrarily without changing the prediction. A well-trained TabPFN approximates this property empirically.

Attention as exponential tilting. Nagler [39, Theorem 6.3] promotes the informal picture above to an exact statement about the *limiting* feature (in the regime where the bias converges to a deterministic



(a) **ID query.** The tilt has a sharp peak near x_e, so the tilted measure $g_h(\cdot \mid x_e)$ concentrates on a small neighborhood of x_e within the support. The mean feature depends strongly on where x_e is.

(b) **OOD query.** The tilt is nearly flat over the support of p_0, so $g_h(\cdot \mid x_e) \approx p_0$. The mean feature degenerates to $\mathbb{E}_{V\sim p_0}[V]$, the same input-independent direction for every OOD query.

Figure 6: **Exponential tilting of the support measure (Nagler Theorem 6.3).** Each attention head reweights the support distribution p_0 (solid blue) by a tilt factor $\psi(s; x_e) = \exp(v(x_e)^\top W_q s)$ (dashed orange) to produce the head's effective sampling distribution $g_h(s \mid x_e)$ (thick green). When the query x_e is close to the support, the tilt concentrates the measure on a small neighborhood of x_e and the resulting feature is query-specific. When x_e is far from the support, the tilt is approximately constant and the measure collapses back to p_0, giving the same feature ϕ_∞ for every far-OOD query.

function $\bar{q}_\theta$). The limiting feature takes the form

$$\bar{u}(x) = \sum_{h=1}^{H} W_v^{(h)}\, \mathbb{E}_{V\sim g_h(\cdot|x)}[V], \tag{27}$$

where the attention head h induces the exponentially tilted measure

$$g_h(s \mid x) = \frac{\exp\big(v(x)^\top W_q^{(h)}\, s\big)}{\mathbb{E}_{V\sim p_0}\big[\exp(v(x)^\top W_q^{(h)} V)\big]}\, p_0(s) \tag{28}$$

on the support marginal p_0. The vector $v(x) \in \mathbb{R}^{d+1}$ is the query's encoding and $W_q^{(h)}, W_v^{(h)}$ are the learned query and value projection matrices for head h.

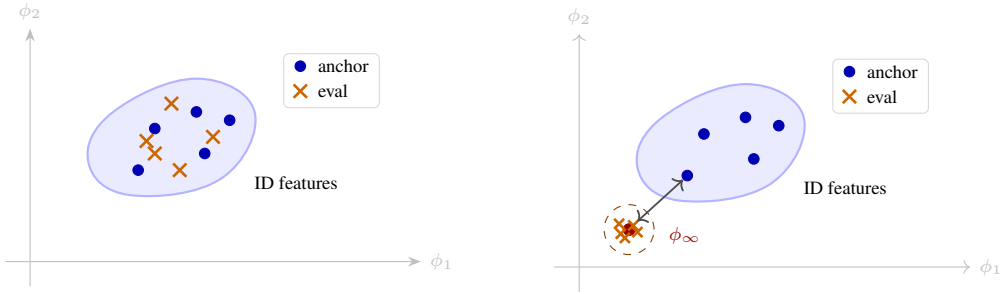
Interpreting Equation (28). The tilted measure $g_h(\cdot \mid x)$ is p_0 reweighted by the exponential factor $\exp(v(x)^\top W_q^{(h)} s)$, normalized to sum to one. Two extreme cases help build intuition:

- If the tilt $\exp(v(x)^\top W_q^{(h)} s)$ has a sharp peak at some $s^* \in \mathrm{supp}(p_0)$, then $g_h(\cdot \mid x)$ concentrates on a small neighborhood of s^*. The resulting feature $\mathbb{E}_{V\sim g_h}[V]$ is close to s^*, and depends strongly on x (which determines where the peak is).
- If the tilt is approximately constant over $s \sim p_0$, the normalized measure $g_h(\cdot \mid x)$ is indistinguishable from p_0. The feature is $\mathbb{E}_{V\sim p_0}[V]$, the p_0-mean, and depends on x only through higher-order perturbations.

Implication for OOD queries. Consider a query x_e drawn from a shifted distribution $p_1 \neq p_0$. The tilt (28) assigns high mass to support samples s that are similar to x_e in the metric induced by $W_q^{(h)}$. Two regimes emerge, corresponding to the two extreme cases above (Figure 6):

- *ID query* ($x_e \sim p_0$, close to the support): $v(x_e)^\top W_q^{(h)} s$ has a sharp maximum over $s \sim p_0$, so $g_h(\cdot \mid x_e)$ concentrates on a localized neighborhood of x_e within the support. The feature $\bar{u}(x_e)$ is then a kernel-smoothed average over that neighborhood and depends strongly on x_e; distinct ID queries produce distinct features.
- *Far-OOD query* (x_e outside the support of p_0): no support sample s is especially similar to x_e, so $v(x_e)^\top W_q^{(h)} s$ is approximately constant over $s \sim p_0$. The tilt degenerates, $g_h(\cdot \mid x_e) \longrightarrow p_0$, and the feature collapses to

$$\bar{u}(x_e) \longrightarrow \phi_\infty := \sum_{h=1}^{H} W_v^{(h)}\, \mathbb{E}_{V\sim p_0}[V], \tag{29}$$



(a) **ID eval.** Anchor features (blue dots) and eval features (orange crosses) both lie in the region occupied by ID features and intermix. Their inner products $K_{ij} = \langle \phi(x_a^{(i)}), \phi(x_e^{(j)}) \rangle$ are on average positive (Assumption A.5(iv) $\to$ Proposition A.6).

(b) **OOD eval.** By Nagler's tilting (Figure 6), eval features (orange crosses) collapse onto the single input-independent direction ϕ_∞ (red dot), far from the ID features. After centering by ϕ_∞, anchor–eval inner products shrink to $O(\epsilon)$.

Figure 7: **Feature-space picture of Corollary A.4.** Under ID **(left)**, anchor and eval features coexist in the ID feature region and the kernel K_{ij} is typically positive. Under OOD **(right)**, eval features are forced by the attention mechanism onto a single input-independent direction ϕ_∞, so the centered kernel $\langle \tilde{\phi}(x_a), \tilde{\phi}(x_e) \rangle$ becomes small (Assumption A.3). This is what switches the gradient geometry from Figure 4 (middle) to Figure 4 (left).

a single input-*independent* vector. Figure 7 visualizes the resulting ID-vs-OOD feature geometry: under ID, features occupy a rich, input-dependent region and anchor-eval inner products are $\Theta(1)$; under far-OOD they collapse to ϕ_∞ and centered inner products shrink to $O(\epsilon)$.

From feature collapse to feature orthogonality. Define the centered feature $\tilde{\phi}(x) := \phi(x) - \phi_\infty$. For far-OOD x_e, $\tilde{\phi}(x_e) \to 0$. For ID x_a, $\tilde{\phi}(x_a)$ retains input-specific structure inherited from the local tilt. Consequently,

$$\left|\langle \tilde{\phi}(x_a),\, \tilde{\phi}(x_e) \rangle\right| = \|\tilde{\phi}(x_a)\|\, \|\tilde{\phi}(x_e)\|\, \left|\cos(\tilde{\phi}(x_a), \tilde{\phi}(x_e))\right| \longrightarrow 0 \tag{30}$$

as x_e recedes from $\mathrm{supp}(p_0)$: the factor $\|\tilde{\phi}(x_e)\| \to 0$ forces the product to zero regardless of the cosine. This verifies Assumption A.3 in the far-OOD limit, with ϵ controlled by $\|\tilde{\phi}(x_e)\|$.

Note on centering. Assumption A.5(iii) asked for feature centering $\mathbb{E}_{p_0}[\phi(x)] = 0$. This is either automatic (under the tilting picture with a centering layer in the architecture) or can be enforced by working with the empirically centered feature $\tilde{\phi}$; either way, all preceding statements apply verbatim.

The feature-collapse picture above gives $\epsilon \to 0$ in Assumption A.3 only in the far-OOD limit $\|p_1 - p_0\|_{\mathrm{TV}} \to 1$, via Proposition A.15. For near-OOD, OOD features stay close to ID features and the anchor-eval cosine remains $\Theta(1)$, so Theorem A.9's signed lower bound is vacuous.

A.9 Multi-Step Extension

Proposition A.2 analyzes the first gradient step. Algorithm 1 uses $T = 50$ steps, so we need a multi-step statement. The main idea is that in a small-step regime, the T-step decrement is the sum of T first-step decrements plus a telescoping error of size $O(T^2\eta^2 L)$. The sign of the cross-term at the first step therefore determines the sign of the cumulative descent, provided that the cross-term does not flip sign along the trajectory.

Assumption A.7 (Small-step regime)**.** *The joint loss $\mathcal{L}$ has L-Lipschitz gradient on the trajectory $\{W_t\}_{t=0}^{T}$, and $T\eta L \le 1/2$.*

The smoothness constant. For the last-layer parametrization $q_W(y \mid x) = \mathrm{softmax}(W\phi(x))_y$, the Hessian of the per-sample CE loss with respect to W is the Kronecker product

$$\nabla_W^2\, \ell_{\mathrm{CE}}(W; x, y) = J(q_W) \otimes \phi(x)\phi(x)^\top,$$

where $J(q) := \mathrm{diag}(q) - qq^\top$ is the Jacobian of the softmax at $q = q_W(\cdot \mid x)$. For any unit vector $v \in \mathbb{R}^C$, $v^\top J(q) v = \sum_c q_c v_c^2 - (\sum_c q_c v_c)^2 = \mathrm{Var}_{c\sim q}(v_c) \le 1/2$, with the maximum achieved by a two-class concentrated q and v pointing along the corresponding axis. Hence $\|J(q)\|_{\mathrm{op}} \le 1/2$. The second Kronecker factor has operator norm $\|\phi(x)\|^2 \le B_\phi^2$, and the operator norm of a Kronecker product is the product of operator norms, so

$$L \le B_\phi^2/2. \tag{31}$$

The DCE Hessian admits the same bound: DCE is itself a cross-entropy against $u_{\hat{y}}$, so its Hessian has the same $J(q) \otimes \phi\phi^\top$ form. Averaging over the batch preserves the bound.

Proposition A.8 (T-step anchor descent)**.** *Under Assumption A.7, let* $\bar{g}_a := \frac{1}{T}\sum_{t=0}^{T-1} \nabla\mathcal{L}_{\mathrm{anc}}(W_t)$ *and* $\bar{g}_e := \frac{1}{T}\sum_{t=0}^{T-1} \nabla\mathcal{L}_{\mathrm{dis}}(W_t)$ *denote the time-averaged gradients along the trajectory. Then*

$$\mathcal{L}_{\mathrm{anc}}(W_T) - \mathcal{L}_{\mathrm{anc}}(W_0) = -T\eta\big(\|\bar{g}_a\|_F^2 + \langle \bar{g}_a,\, \bar{g}_e\rangle_F\big) + O(T^2\eta^2 L\, B_g^2), \tag{32}$$

where $B_g := \max_t \|\nabla\mathcal{L}(W_t)\|_F$ *is a trajectory-uniform gradient bound. Furthermore, if* $\langle g_a^{(t)},\, g_a^{(0)}\rangle_F \ge (1-\kappa)\,\|g_a^{(t)}\|_F\,\|g_a^{(0)}\|_F$ *for all* $t \le T$ *(gradients do not rotate substantially), then* $\|\bar{g}_a\|_F^2 = \|g_a^{(0)}\|_F^2\,(1 + O(\kappa))$.

Proof. We apply Proposition A.2 at each step of the trajectory and sum, controlling the accumulated error from the Hessian terms.
Step 1 (first-step expansion at each t). For each $t = 0, 1, \ldots, T-1$, apply Taylor's theorem to $\mathcal{L}_{\mathrm{anc}}$ on the segment $[W_t, W_{t+1}]$ with $W_{t+1} = W_t - \eta(g_a^{(t)} + g_e^{(t)})$ and $g_\star^{(t)} := \nabla\mathcal{L}_\star(W_t)$ for $\star \in \{\mathrm{anc}, \mathrm{dis}\}$:

$$\mathcal{L}_{\mathrm{anc}}(W_{t+1}) - \mathcal{L}_{\mathrm{anc}}(W_t) = -\eta\,\|g_a^{(t)}\|_F^2 - \eta\,\langle g_a^{(t)}, g_e^{(t)}\rangle_F + R_t,$$

with $|R_t| \le \frac{1}{2}L\eta^2\|g_a^{(t)} + g_e^{(t)}\|_F^2 \le \frac{1}{2}L\eta^2\, B_g^2$.
Step 2 (telescoping). Summing over $t = 0, \ldots, T-1$,

$$\mathcal{L}_{\mathrm{anc}}(W_T) - \mathcal{L}_{\mathrm{anc}}(W_0) = -\eta\sum_{t=0}^{T-1}\|g_a^{(t)}\|_F^2 - \eta\sum_{t=0}^{T-1}\langle g_a^{(t)}, g_e^{(t)}\rangle_F + \sum_{t=0}^{T-1} R_t,$$

with the Hessian remainder bounded by $\sum_t |R_t| \le \frac{1}{2}TL\eta^2 B_g^2$.
Step 3 (replacing sums by time-averaged gradients). The trajectory displacement is $\|W_t - W_0\| \le t\eta B_g \le T\eta B_g$, and L-Lipschitzness of $\nabla\mathcal{L}$ gives $\|g_\star^{(t)} - g_\star^{(0)}\|_F \le LT\eta B_g$ for $\star \in \{\mathrm{anc}, \mathrm{dis}\}$ at every t. Hence each $g_\star^{(t)}$ lies within an $LT\eta B_g$-ball of $\bar{g}_\star$, and

$$\big|\tfrac{1}{T}\sum_t \|g_a^{(t)}\|_F^2 - \|\bar{g}_a\|_F^2\big| \le (LT\eta B_g)^2, \qquad \big|\tfrac{1}{T}\sum_t \langle g_a^{(t)}, g_e^{(t)}\rangle_F - \langle \bar{g}_a, \bar{g}_e\rangle_F\big| \le (LT\eta B_g)^2,$$

each by direct expansion and Cauchy–Schwarz. Multiplying by $T\eta$ converts this to an error of size $T^3\eta^3L^2B_g^2$ in the anchor-loss decrement. Under Assumption A.7 ($T\eta L \le 1/2$), this is dominated by the Hessian remainder $T^2\eta^2 LB_g^2$, so the total remainder in (32) is $O(T^2\eta^2 LB_g^2)$.
Step 4 (small-rotation corollary). If $g_a^{(t)}$ stays within angle $\arccos(1-\kappa)$ of $g_a^{(0)}$ for all t, then $\|\bar{g}_a\|_F^2$ is within a $(1 - O(\kappa))$ factor of $\|g_a^{(0)}\|_F^2$ by direct calculation: decompose each $g_a^{(t)} = \alpha_t g_a^{(0)} + \beta_t h_t$ with $\|h_t\| = 1$, $h_t \perp g_a^{(0)}$, and $\alpha_t\|g_a^{(0)}\|_F \ge (1-\kappa)\|g_a^{(t)}\|_F$, then average. ∎

Proposition A.8 inherits the sign structure of the one-step result. If the OOD feature-orthogonality of Corollary A.4 holds *uniformly* along the trajectory (which requires only that the backbone is frozen, not that W stays close to W_0: the feature-space kernel K_{ij} does not depend on W at all), then OOD anchor descent at the time-averaged rate is $\approx -T\eta\|\bar{g}_a\|_F^2$. By Proposition A.6, ID anchor descent is strictly damped to $-T\eta(\|\bar{g}_a\|_F^2 + \langle \bar{g}_a, \bar{g}_e\rangle_F)$ with $\langle \bar{g}_a, \bar{g}_e\rangle_F < 0$.

A.10 Main Separation Theorem

The preceding subsections have established: (i) anchor descent is controlled by the gradient cross-term (Proposition A.2); (ii) under OOD feature orthogonality the cross-term is $O(\epsilon)$ (Corollary A.4); (iii) under ID confidence the cross-term has a strictly negative expectation $-\gamma$ for some $\gamma > 0$ (Proposition A.6); (iv) the first-step analysis lifts to T steps in the small-step regime (Prop A.8). We now compose these into a separation theorem that underpins the detector.

Theorem A.9 (Main separation)**.** *Let p_1 be an eval distribution and write $\mathbb{E}[\lambda_{\mathrm{ID}}]$, $\mathbb{E}[\lambda_{\mathrm{OOD}}]$ for the expected post-finetuning anchor losses when X_{eval} is drawn i.i.d. from p_0 and from p_1 respectively. Suppose the core gradient-geometry assumptions hold: A1–A3 of Table 2 (design and boundedness), together with Assumption A.7 (small-step regime), Assumption A.3 (with parameter ϵ under p_1), and Assumption A.5 (under p_0). Let $\gamma := -\mathbb{E}_{p_0 \otimes p_0}[C_{ij} K_{ij}] > 0$ denote the ID interference strength from Proposition A.6. Then*

$$\mathbb{E}[\lambda_{\mathrm{ID}}] - \mathbb{E}[\lambda_{\mathrm{OOD}}] \geq T\eta\big(\gamma - \epsilon B_\phi^2 B_C\big) - O(T^2\eta^2 L\, B_g^2), \tag{33}$$

where B_ϕ, B_C, L, B_g are the constants defined earlier. In particular, whenever $\epsilon < \gamma/(B_\phi^2 B_C)$ and $T\eta$ is in the small-step regime of Assumption A.7, the right-hand side is strictly positive and $\mathbb{E}[\lambda_{\mathrm{OOD}}] < \mathbb{E}[\lambda_{\mathrm{ID}}]$.

Proof. Apply Proposition A.8 separately under p_0 and p_1. Let $\bar{g}_a^{\mathrm{ID}}, \bar{g}_e^{\mathrm{ID}}$ and $\bar{g}_a^{\mathrm{OOD}}, \bar{g}_e^{\mathrm{OOD}}$ denote the respective time-averaged gradients, and let $\lambda(W_0) = \mathcal{L}_{\mathrm{anc}}(W_0)$ be the pre-finetuning anchor loss (the same in both regimes since it depends only on the anchor, not on the eval batch). Then

$$\begin{aligned}
\lambda_{\mathrm{ID}} - \lambda(W_0) &= -T\eta\big(\|\bar{g}_a^{\mathrm{ID}}\|_F^2 + \langle \bar{g}_a^{\mathrm{ID}}, \bar{g}_e^{\mathrm{ID}} \rangle_F\big) + O(T^2\eta^2 L B_g^2), \\
\lambda_{\mathrm{OOD}} - \lambda(W_0) &= -T\eta\big(\|\bar{g}_a^{\mathrm{OOD}}\|_F^2 + \langle \bar{g}_a^{\mathrm{OOD}}, \bar{g}_e^{\mathrm{OOD}} \rangle_F\big) + O(T^2\eta^2 L B_g^2).
\end{aligned}$$

Subtract:

$$\begin{aligned}
\mathbb{E}[\lambda_{\mathrm{ID}} - \lambda_{\mathrm{OOD}}] = T\eta \Big[&\mathbb{E}\|\bar{g}_a^{\mathrm{OOD}}\|_F^2 - \mathbb{E}\|\bar{g}_a^{\mathrm{ID}}\|_F^2 \\
&+ \mathbb{E}\langle \bar{g}_a^{\mathrm{OOD}}, \bar{g}_e^{\mathrm{OOD}} \rangle_F - \mathbb{E}\langle \bar{g}_a^{\mathrm{ID}}, \bar{g}_e^{\mathrm{ID}} \rangle_F \Big] + O(T^2\eta^2 L B_g^2).
\end{aligned} \tag{34}$$

The anchor gradient $\bar{g}_a$ depends only on the anchor set and the classifier at W_t; it is the same in expectation at $t = 0$ whether the eval batch is from p_0 or p_1, because the finetuning trajectory is anchored at W_0 and the leading term is the same first-step gradient. Along the trajectory, W_t^{ID} and W_t^{OOD} diverge by at most $O(T\eta B_g)$ (both trajectories have per-step displacement $\leq \eta B_g$), so by L-Lipschitz continuity of $\nabla\mathcal{L}_{\mathrm{anc}}$ and Cauchy–Schwarz, $\mathbb{E}\|\bar{g}_a^{\mathrm{OOD}}\|_F^2 - \mathbb{E}\|\bar{g}_a^{\mathrm{ID}}\|_F^2 = O(T\eta L B_g^2)$. Multiplied by the outer $T\eta$ factor, this is $O(T^2\eta^2 L B_g^2)$ and is subsumed in the second-order remainder. Hence

$$\mathbb{E}[\lambda_{\mathrm{ID}} - \lambda_{\mathrm{OOD}}] = T\eta\left(\mathbb{E}\langle \bar{g}_a, \bar{g}_e^{\mathrm{OOD}} \rangle_F - \mathbb{E}\langle \bar{g}_a, \bar{g}_e^{\mathrm{ID}} \rangle_F\right) + O(T^2\eta^2 L B_g^2),$$

where we have identified $\bar{g}_a^{\mathrm{ID}} \approx \bar{g}_a^{\mathrm{OOD}} =: \bar{g}_a$ to leading order.
Apply Corollary A.4 to the OOD term:

$$\left|\mathbb{E}\langle \bar{g}_a, \bar{g}_e^{\mathrm{OOD}} \rangle_F\right| \leq \epsilon\, B_\phi^2\, B_C.$$

Apply Proposition A.6 to the ID term:

$$\mathbb{E}\langle \bar{g}_a, \bar{g}_e^{\mathrm{ID}} \rangle_F \leq -\gamma.$$

Substituting and using $|a| \leq \epsilon B_\phi^2 B_C \Rightarrow -\epsilon B_\phi^2 B_C \leq a$,

$$\begin{aligned}
\mathbb{E}[\lambda_{\mathrm{ID}} - \lambda_{\mathrm{OOD}}] &\geq T\eta\left(-\epsilon B_\phi^2 B_C - (-\gamma)\right) + O(T^2\eta^2 L B_g^2) \\
&= T\eta\left(\gamma - \epsilon B_\phi^2 B_C\right) + O(T^2\eta^2 L B_g^2).
\end{aligned}$$

∎

What the theorem says. The separation between expected ID and OOD anchor losses is at least $T\eta$ times an architecture-independent gap $\gamma - \epsilon B_\phi^2 B_C$, which is positive whenever the OOD feature orthogonality is strong enough to overwhelm the ID interference constant. The factor $T\eta$ means longer finetuning amplifies the separation linearly (within the small-step regime), while the constant γ is a structural property of the ID data and the pretrained classifier and is independent of the shift's geometry. Severity dependence enters through ϵ: Proposition A.15's feature collapse drives $\tilde{\phi}(x_e) \longrightarrow 0$ (and hence $\epsilon \longrightarrow 0$) in the far-OOD regime $\|p_1 - p_0\|_{\mathrm{TV}} \longrightarrow 1$, but ϵ can be $\Theta(1)$ for near-OOD. The theorem's signed lower bound is therefore only applicable in the far-OOD regime.

When the theorem fails. Theorem A.9's guarantee becomes vacuous when $\epsilon B_\phi^2 B_C \geq \gamma$, i.e., when OOD features retain too much ID-like structure relative to the ID interference strength. In this regime the theorem imposes no signed constraint on the separation, and the observed direction is an empirical matter: the ! flags in Table 10 mark the dataset–model pairs where it reverses ($\lambda_{\mathrm{OOD}} > \lambda_{\mathrm{ID}}$ for the majority of seeds). All four such cells are benign shifts, and all four coincide with the cells where the learning-rate stability rule falls back to $\eta = 10^{-4}$.

A.11 Statistical Validity of the Hypothesis Test

Theorem A.9 establishes a separation between expected anchor losses under ID and OOD distributions. The practical detector of Algorithm 1 does not observe these expectations directly; it observes K calibration samples $\lambda^{(1)}, \ldots, \lambda^{(K)}$ and a single test statistic λ^*, and rejects H_0 when λ^* falls below the empirical α-quantile of the calibration set. We now verify that this decision rule controls Type-I error at level α (up to a finite-K correction), via a standard conformal prediction argument.

We first make precise what "in-distribution" means for this test. In a continuous feature space the data need not occupy a sharply bounded set, and density typically decays gradually rather than defining a natural in/out boundary.

Definition A.10 (In-distribution evaluation batch)**.** *Let p_0 denote the data-generating distribution on inputs and labels, and p_0^X its covariate marginal. The support set, the anchor set $(X_{\mathrm{anc}}, y_{\mathrm{anc}})$, and the K calibration batches are sampled from p_0. Conditional on the fixed support and anchor sets, a deployment evaluation batch is* in-distribution *when*

$$H_0 : \; X^*_{\mathrm{eval}} \sim (p_0^X)^{\otimes n_e},$$

that is, when it is drawn from the same covariate distribution as — and is therefore exchangeable with — the calibration batches.

This is a distribution-level definition. One could instead posit a pointwise ID region, for example a density level set $\{x : \mathrm{d}p_0^X / \mathrm{d}\mu\,(x) > c\}$ or a support neighborhood $\{x : d(x, \operatorname{supp} p_0^X) \leq r\}$, but both require an arbitrary choice of c or r. Note that **CALD** uses neither: the detector thresholds the calibrated statistic λ, and the validity of Proposition A.11 below follows from the exchangeability in Definition A.10 alone, independent of dimension, support geometry, or the existence of a density. The feature-separation parameter ϵ of Theorem A.9 plays a different role: it governs the *power* analysis, not the definition of ID versus OOD — in the far-OOD limit $\|p_1 - p_0\|_{\mathrm{TV}} \longrightarrow 1$ the tilting argument predicts $\epsilon \longrightarrow 0$, while for near-OOD shifts ϵ may remain $\Theta(1)$, making the power bound vacuous even where the detector has empirical power.

Proposition A.11 (Finite-sample Type-I error control)**.** *Consider the calibrated-disagreement-detection test defined in Algorithm 1. Assume that the anchor set $(X_{\mathrm{anc}}, y_{\mathrm{anc}})$, the support set $\mathcal{D}^{\mathrm{sup}}$, and all finetuning hyperparameters (T, η) are fixed; that the eval batches $X^{(1)}_{\mathrm{eval}}, \ldots, X^{(K)}_{\mathrm{eval}}, X^*_{\mathrm{eval}}$ across rounds are i.i.d. draws; and that the finetuning inner loop is deterministic given its inputs. Under the null hypothesis $H_0 : X^*_{\mathrm{eval}} \sim p_0$ (same distribution as the calibration eval batches), the one-sided test*

$$\text{Reject } H_0 \iff \lambda^* < \lambda^{(\lceil \alpha(K+1) \rceil)} \tag{35}$$

(where $\lambda^{(k)}$ denote the calibration statistics in ascending order) satisfies

$$\mathbb{P}_{H_0}(\text{Reject}) \leq \frac{\lceil \alpha(K+1) \rceil}{K+1} \leq \alpha + \frac{1}{K+1}. \tag{36}$$

Proof. Under H_0 and the stated i.i.d. conditions, the $K+1$ random variables

$$\lambda^{(1)}, \ldots, \lambda^{(K)}, \lambda^*$$

are i.i.d. because each $\lambda^{(k)}$ and λ^* is a deterministic function of a fresh eval batch $X_{\text{eval}}^{(k)}$ (resp. X_{eval}^*) applied through the same fixed anchor, support, and finetuning procedure. In particular they are *exchangeable*: the joint law is invariant under any permutation of the indices.
Let R denote the rank of λ^* among the $K+1$ statistics when they are sorted in ascending order (breaking ties uniformly at random). By exchangeability, R is uniformly distributed on $\{1, 2, \ldots, K+1\}$. The test rejects exactly when $\lambda^* < \lambda^{(\lceil \alpha(K+1) \rceil)}$, i.e., when λ^* is among the $\lceil \alpha(K+1) \rceil - 1$ smallest of the $K+1$ statistics, i.e., when $R \leq \lceil \alpha(K+1) \rceil - 1$. Hence

$$\mathbb{P}_{H_0}(\text{Reject}) = \mathbb{P}\big(R \leq \lceil \alpha(K+1) \rceil - 1\big) = \frac{\lceil \alpha(K+1) \rceil - 1}{K+1} \leq \frac{\alpha(K+1)}{K+1} = \alpha,$$

using $\lceil \alpha(K+1) \rceil \leq \alpha(K+1)+1$ in the inequality. This gives the tight bound $\mathbb{P}_{H_0}(\text{Reject}) \leq \alpha$. The weaker (but standard) form quoted in (36), $\mathbb{P}_{H_0}(\text{Reject}) \leq \lceil \alpha(K+1) \rceil / (K+1) \leq \alpha + 1/(K+1)$, is the discretization version obtained by dropping the -1: its slack $1/(K+1)$ quantifies the cost of using a finite calibration set. ∎

Power analysis. Proposition A.11 controls Type-I error only; power depends on the separation established in Theorem A.9. A power statement requires two ingredients. First, a mean separation: $\mathbb{E}[\lambda_{\text{OOD}}] \leq \mathbb{E}[\lambda_{\text{ID}}] - \Delta_\lambda$ for some $\Delta_\lambda > 0$. Theorem A.9 supplies one, with $\Delta_\lambda \geq T\eta(\gamma - \epsilon B_\phi^2 B_C)$. Second, concentration: λ_{ID} and λ_{OOD} must be sufficiently tight around their means. The batch-gradient structure gives an $O(1/\sqrt{n_e})$ standard deviation under the small-step regime, but a full Hoeffding-type bound would need to be proved separately. Given both ingredients, the detector's power approaches 1 as $K \to \infty$ for any Δ_λ exceeding the concentration noise. We leave a detailed power analysis as a direction for future work.

Assumption satisfied in the experimental setup. The assumptions of Proposition A.11 match Algorithm 1's protocol: anchor and support are fixed across calibration rounds (only eval batches are resampled), the finetuning inner loop is deterministic given its inputs (full-batch gradient descent with no randomness beyond the initial draw), and the eval batches are i.i.d. from p_0 during calibration.

A.12 Multiclass Case

All statements in Lemma A.1, Proposition A.2, Corollary A.4, and Proposition A.8 hold for arbitrary $C \geq 2$: the outer-product structure $\nabla_W \ell = (q - \text{target})\, \phi^\top$ produces a C-dimensional first factor but carries through unchanged. The ID-interference argument in Proposition A.6 uses binary structure for concreteness; the extension to $C > 2$ replaces the two-component sign analysis with a joint distribution over label pairs $(y_a, \hat{y}(x_e)) \in [C]^2$. Under a class-symmetry condition (p_0 marginally balanced over classes, or ϕ covariance block-diagonal across class regions), the label-agreement contribution dominates and gives the same negative expected sign for the cross-term.

A.13 Distributional Difference vs. Predictive Harm

The theory developed so far controls the gradient cross-term $\langle g_a, g_e \rangle_F$ via the TV distance $\|p_1 - p_0\|_{\text{TV}}$: an *unsigned* measure of distributional difference. But a detector of *performance deterioration* should fire only on shifts that actually hurt predictive accuracy. As the benign-shift experiments (Adult Income, Diabetes Readmission) show, these notions are not equivalent.

This subsection bridges the gap. We:

1. Formally decompose the accuracy gap into a density-ratio integral that makes the sign structure of harm explicit.
2. Extend Nagler's tilting from features to *predictions*: at far-OOD inputs, not only does the feature collapse, but so does the classifier's output, converging to a constant distribution q_∞.
3. Propose a second, accuracy-sensitive signal $\hat{\mu}$ that estimates the accuracy gap Δ directly (Proposition A.21), with $\hat{\lambda}$ serving as a shift-confirmation gate.

The additional assumptions required by Proposition A.21 are catalogued in Section A.14.

A.13.1 Accuracy gap as a signed density-ratio integral

Let $\ell_{01}(q, x, y) := \mathbb{1}\{y \neq \arg\max_c q(c \mid x)\}$ be the 0-1 loss and $\mathrm{acc}_q(p) := 1 - \mathbb{E}_{(x,y)\sim p}[\ell_{01}(q, x, y)]$. The accuracy gap between p_0 and p_1 under the fixed (pretrained) classifier $q = q_{W_0}$ is

$$\Delta(q, p_0, p_1) := \mathrm{acc}_q(p_0) - \mathrm{acc}_q(p_1). \tag{37}$$

Lemma A.12 (Density-ratio decomposition of the accuracy gap)**.** *Suppose p_0 and p_1 have densities with respect to a common reference measure on $\mathcal{X} \times \mathcal{Y}$, with $p_0(x, y) > 0$ wherever $p_1(x, y) > 0$, and let $r(x, y) := p_1(x, y)/p_0(x, y)$ denote the density ratio. Then*

$$\Delta(q, p_0, p_1) = \mathbb{E}_{p_0}\big[(r - 1)\,\ell_{01}\big]. \tag{38}$$

Proof. Step 1 (rewrite Δ in loss form). Since $\mathrm{acc}_q(p) = 1 - \mathbb{E}_p[\ell_{01}(q, x, y)]$,

$$\Delta = \mathrm{acc}_q(p_0) - \mathrm{acc}_q(p_1) = (1 - \mathbb{E}_{p_0}[\ell_{01}]) - (1 - \mathbb{E}_{p_1}[\ell_{01}]) = \mathbb{E}_{p_1}[\ell_{01}] - \mathbb{E}_{p_0}[\ell_{01}].$$

Step 2 (change of measure). The support condition $\{p_1 > 0\} \subseteq \{p_0 > 0\}$ ensures the density ratio $r = p_1/p_0$ is well-defined and finite p_0-almost everywhere. For any bounded $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$,

$$\mathbb{E}_{p_1}[f] = \int f\, p_1 = \int f \cdot \frac{p_1}{p_0} \cdot p_0 = \int f\, r\, p_0 = \mathbb{E}_{p_0}[r\, f].$$

Applied to $f = \ell_{01}$: $\mathbb{E}_{p_1}[\ell_{01}] = \mathbb{E}_{p_0}[r\,\ell_{01}]$.
Step 3 (subtract and combine).

$$\Delta = \mathbb{E}_{p_0}[r\,\ell_{01}] - \mathbb{E}_{p_0}[\ell_{01}] = \mathbb{E}_{p_0}[(r - 1)\,\ell_{01}].$$

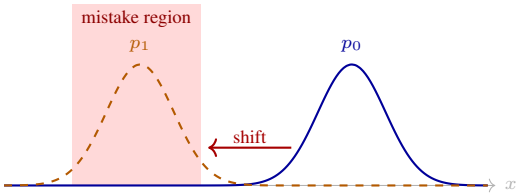
∎

Reading the decomposition. The gap is the p_0-expectation of the 0-1 loss weighted by the *signed* excess density $r - 1$:

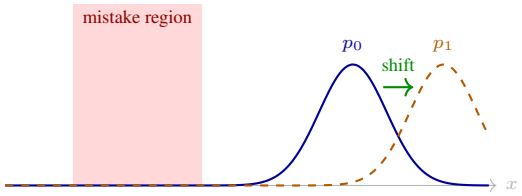
- On the mistake region $M := \{(x, y) : \ell_{01}(q, x, y) = 1\}$: if $r > 1$ (shift puts more mass there), the integrand $(r-1)\ell_{01}$ is positive, so the contribution to Δ is positive (i.e., $\mathrm{acc}_{p_1} < \mathrm{acc}_{p_0}$; harmful). Conversely, if $r < 1$ on M, the shift pulls mass *out* of the mistake region, the contribution is negative, and $\Delta < 0$ (OOD accuracy exceeds ID accuracy; benign).
- On the correct region M^c: the integrand vanishes because $\ell_{01} = 0$, so changes in mass on correctly-classified regions do not affect Δ directly.

Thus Δ is controlled *entirely* by how the shift re-weights the mistake region. In contrast, $\|p_1 - p_0\|_{\mathrm{TV}}$ is controlled by re-weighting of the *entire* domain. Two shifts with the same TV distance can produce very different Δ values depending on whether they transfer mass into or out of the mistake region.

Corollary A.13 (TV upper bound is loose)**.** $|\Delta(q, p_0, p_1)| \leq 2\,\|p_1 - p_0\|_{\mathrm{TV}}$*, with equality if and only if $\{|r - 1| > 0\} \subseteq M$ up to p_0-null sets and $(r - 1)\ell_{01}$ has constant sign on M (i.e., the shift is supported on the mistake region and the density ratio is monotone in one direction).*



(a) **Harmful shift.** p_1 moves mass into the classifier's mistake region (shaded), so the accuracy gap $\Delta > 0$ and accuracy drops on the shifted distribution.



(b) **Benign shift.** p_1 moves mass away from the mistake region (shaded) into the correct region, so $\Delta \leq 0$ and accuracy is preserved or improved.

Figure 8: **Same TV distance, opposite signs of Δ.** Both panels show a shift $p_0 \to p_1$ of the same magnitude (same $\|p_1 - p_0\|_{\mathrm{TV}}$) under the same classifier, which has a mistake region on the left of the axis (shaded red). In **(a)**, the shift transfers mass into the mistake region, so accuracy drops. In **(b)**, the shift transfers mass away from the mistake region, so accuracy is preserved or improved. The anchor-loss detector $\hat{\lambda}$ (Section A.8) has the same value in both cases; it is insensitive to the sign of Δ. An additional signal is needed.

Proof. Step 1 (absolute-value bound inside expectation). By Lemma A.12 and the triangle inequality,

$$|\Delta| = \left|\mathbb{E}_{p_0}[(r-1)\,\ell_{01}]\right| \leq \mathbb{E}_{p_0}\left[|r-1|\,\ell_{01}\right].$$

Since $\ell_{01}(q, x, y) \in \{0, 1\}$, $|r-1|\,\ell_{01} \leq |r-1|$ pointwise, hence

$$\mathbb{E}_{p_0}\left[|r-1|\,\ell_{01}\right] \leq \mathbb{E}_{p_0}[|r-1|].$$

Step 2 (identify $\mathbb{E}_{p_0}[|r-1|]$ with TV distance). The standard density-based formula for total variation distance reads

$$\|p_1 - p_0\|_{\mathrm{TV}} = \tfrac{1}{2}\int |p_1 - p_0| = \tfrac{1}{2}\int |r-1|\,p_0 = \tfrac{1}{2}\,\mathbb{E}_{p_0}[|r-1|].$$

Therefore $\mathbb{E}_{p_0}[|r-1|] = 2\,\|p_1 - p_0\|_{\mathrm{TV}}$. Combining with Step 1 gives $|\Delta| \leq 2\,\|p_1 - p_0\|_{\mathrm{TV}}$.
Step 3 (equality conditions). The two inequalities used above are tight under the following conditions:

- $|\mathbb{E}[g]| = \mathbb{E}[|g|]$ requires $g = (r-1)\ell_{01}$ to have constant sign p_0-almost surely.
- $|r-1|\,\ell_{01} = |r-1|$ p_0-almost surely requires $\ell_{01}(q, x, y) = 1$ on $\{|r-1| > 0\}$, i.e., the shift is supported on the mistake region.

Both must hold simultaneously for equality in the TV bound. ∎

Corollary A.13 formalizes the central tension: TV distance upper-bounds $|\Delta|$ but the bound is rarely tight on benchmark data, because shifts on real-world data rarely align perfectly with the classifier's mistake region. Figure 8 shows two shifts with similar TV distances but opposite signs of Δ.

A.13.2 Far-OOD prediction collapse

Nagler's tilting argument (Section A.8) established that far-OOD features collapse to a single input-independent direction ϕ_∞. We now observe that this extends to the classifier's output: all far-OOD queries receive the *same* prediction q_∞.

Assumption A.14 (Far-OOD regime)**.** *The eval distribution p_1 satisfies $\|p_1 - p_0\|_{\mathrm{TV}} \longrightarrow 1$ in the sense that, for p_1-almost every x_e, the tilt $\exp(v(x_e)^\top W_q^{(h)} s)$ is approximately constant over $s \sim p_0$ for every head h.*

We acknowledge that this is the most aggressive assumption in this derivation and it is therefore best read as an idealization; it is defensible only for one-layer attention (Nagler [39]'s tilting theorem

assumes a single layer, while TabPFN and TabICL are multi-layer), and only in the strict limit $\|p_1 - p_0\|_{\mathrm{TV}} \to 1$, while typical benchmark shifts have TV distance bounded well away from 1. The conclusion of Proposition A.15 below, that *all* far-OOD queries receive the same prediction q_∞, should accordingly be read as the limit point of a trend rather than a description of any specific experiment. We note that the subsequent separation theorem (Theorem A.9) does not depend on this assumption; only the harm theorem (Proposition A.21) does.

Proposition A.15 (Far-OOD prediction collapse)**.** *Under Assumption A.14, the PFN's prediction collapses pointwise: for p_1-almost every x_e,*

$$q_{W_0}(\cdot \mid x_e) \longrightarrow q_\infty := \mathrm{softmax}(W_0 \phi_\infty) \tag{39}$$

in total variation, where $\phi_\infty := \sum_h W_v^{(h)} \mathbb{E}_{V \sim p_0}[V]$ is the input-independent direction defined in (29)*; equivalently, q_∞ is the same constant prediction for every far-OOD query.*

Proof. We show successively that (i) the tilted measures at each head approach p_0 in total variation, (ii) the tilted mean features approach $\mathbb{E}_{p_0}[V]$ in norm, (iii) the backbone feature approaches ϕ_∞, and (iv) the softmax output approaches q_∞ in total variation.
Step 1 (tilted measure converges to p_0). Assumption A.14 posits a sequence $\epsilon_n \to 0$ and a set of x_e's of p_1-mass $\to 1$ such that, for every head h,

$$\sup_{s \in \mathrm{supp}(p_0)} \big|\psi_h(s; x_e) - Z_h(x_e)\big| \le \epsilon_n,$$

where $\psi_h(s; x) := \exp(v(x)^\top W_q^{(h)} s)$ and $Z_h(x) := \mathbb{E}_{V \sim p_0}[\psi_h(V; x)]$ is the partition function. Using this uniform bound,

$$\big\|g_h(\cdot \mid x_e) - p_0\big\|_{\mathrm{TV}} = \tfrac{1}{2}\, \mathbb{E}_{V \sim p_0}\left[\left|\frac{\psi_h(V; x_e)}{Z_h(x_e)} - 1\right|\right] \le \frac{\epsilon_n}{2 Z_h(x_e)}.$$

Step 2 (tilted mean feature converges). For any bounded V (say $\|V\| \le B_s$),

$$\big\|\mathbb{E}_{V \sim g_h(\cdot \mid x_e)}[V] \;-\; \mathbb{E}_{V \sim p_0}[V]\big\| \le 2B_s \cdot \big\|g_h(\cdot \mid x_e) - p_0\big\|_{\mathrm{TV}} \le \frac{B_s\, \epsilon_n}{Z_h(x_e)}.$$

Step 3 (feature converges to ϕ_∞). Substituting Step 2 into the limiting-feature expression (27),

$$\big\|\bar{u}(x_e) - \phi_\infty\big\| = \Big\|\sum_{h=1}^{H} W_v^{(h)} \big(\mathbb{E}_{g_h}[V] - \mathbb{E}_{p_0}[V]\big)\Big\| \le \sum_{h=1}^{H} \|W_v^{(h)}\| \, \frac{B_s\, \epsilon_n}{Z_h(x_e)} =: C_\phi(x_e)\, \epsilon_n,$$

where $C_\phi(x_e) := B_s \sum_h \|W_v^{(h)}\| / Z_h(x_e) < \infty$ for p_1-a.e. x_e (the partition functions are bounded away from zero on the assumed far-OOD set).
Step 4 (softmax is Lipschitz). The softmax map $z \mapsto \mathrm{softmax}(z)$ is 1-Lipschitz from $(\mathbb{R}^C, \|\cdot\|_2)$ to the probability simplex under $\|\cdot\|_2$ (a standard consequence of the Jacobian bound $\|J(q)\|_{\mathrm{op}} \le 1/2$ used in Section A.9, which gives a tighter constant $1/2$ for differentiable paths; the 1-Lipschitz statement suffices here):

$$\big\|\mathrm{softmax}(z) - \mathrm{softmax}(z')\big\|_2 \le \|z - z'\|_2.$$

Converting to the ℓ_1 norm used by TV distance via $\|\cdot\|_1 \le \sqrt{C}\, \|\cdot\|_2$,

$$\big\|\mathrm{softmax}(z) - \mathrm{softmax}(z')\big\|_1 \le \sqrt{C}\, \|z - z'\|_2.$$

Applied to $z = W_0 \bar{u}(x_e)$ and $z' = W_0 \phi_\infty$, and using $\|z - z'\|_2 \le \|W_0\|_{\mathrm{op}} \|\bar{u}(x_e) - \phi_\infty\|_2$,

$$\begin{aligned} \big\|q_{W_0}(\cdot \mid x_e) - q_\infty\big\|_{\mathrm{TV}} \;&=\; \tfrac{1}{2}\, \big\|\mathrm{softmax}(W_0 \bar{u}(x_e)) - \mathrm{softmax}(W_0 \phi_\infty)\big\|_1 \\ &\le\; \tfrac{\sqrt{C}}{2}\, \|W_0\|_{\mathrm{op}}\, C_\phi(x_e)\, \epsilon_n. \end{aligned}$$

As $\epsilon_n \to 0$ (Assumption A.14), the right-hand side vanishes p_1-almost surely, proving $q_{W_0}(\cdot \mid x_e) \to q_\infty$ in total variation. ∎

Corollary A.16 (Far-OOD accuracy is constant-predictor accuracy)**.** *Under Assumption A.14,*

$$\mathrm{acc}_{q_{W_0}}(p_1) \longrightarrow \mathrm{acc}_{q_\infty}(p_1) = p_{Y|p_1}(\hat{y}_\infty), \tag{40}$$

where $\hat{y}_\infty := \arg\max_c q_\infty(c)$ *is the constant prediction at* ϕ_∞ *and* $p_{Y|p_1}$ *is the label marginal under* p_1*.*

Proof. Write the accuracy as an expectation:

$$\mathrm{acc}_q(p_1) = \mathbb{P}_{(x,y)\sim p_1}\left(\arg\max_c q(c \mid x) = y\right).$$

Step 1 (arg-max converges). Proposition A.15 gives $q_{W_0}(\cdot \mid x_e) \longrightarrow q_\infty$ in TV for p_1-almost every x_e. We assume that q_∞ has a unique maximizer, so the arg-max function is continuous at q_∞; ties, if they occur, can be broken by an arbitrary fixed rule without changing the statement. Hence

$$\arg\max_c q_{W_0}(c \mid x_e) \longrightarrow \arg\max_c q_\infty(c) = \hat{y}_\infty$$

for p_1-almost every x_e.
Step 2 (accuracy difference is bounded by mismatch probability). By the triangle inequality applied to $\mathbb{1}\{\arg\max q_{W_0}(\cdot \mid x) = y\}$ versus $\mathbb{1}\{\hat{y}_\infty = y\}$:

$$\begin{aligned}\left|\mathrm{acc}_{q_{W_0}}(p_1) - \mathrm{acc}_{q_\infty}(p_1)\right| &= \left|\mathbb{P}_{p_1}(\arg\max q_{W_0} = y) - \mathbb{P}_{p_1}(\hat{y}_\infty = y)\right| \\ &\le \mathbb{P}_{p_1}\left(\arg\max_c q_{W_0}(c \mid x) \neq \hat{y}_\infty\right).\end{aligned}$$

By Step 1, the right-hand side tends to zero.
Step 3 (evaluate the limit). Since $\hat{y}_\infty$ is a *constant* independent of x,

$$\mathrm{acc}_{q_\infty}(p_1) = \mathbb{P}_{(x,y)\sim p_1}(\hat{y}_\infty = y) = \mathbb{P}_{y\sim p_{Y|p_1}}(y = \hat{y}_\infty) = p_{Y|p_1}(\hat{y}_\infty).$$

∎

Far-OOD accuracy equals the probability that the constant prediction $\hat{y}_\infty$ matches the true OOD label. Whether this is harmful depends on the alignment between $\hat{y}_\infty$ and the OOD label marginal:

- *Harmful shift*: $\hat{y}_\infty$ poorly aligned with $p_{Y|p_1}$, so $\mathrm{acc}_{q_\infty}(p_1)$ is low (e.g., Covertype: the classifier predicts a near-uniform distribution over forest types, but OOD elevation concentrates on subalpine species; mismatch).
- *Benign shift*: $\hat{y}_\infty$ happens to align with $p_{Y|p_1}$, so $\mathrm{acc}_{q_\infty}(p_1)$ equals or exceeds ID accuracy (e.g., Adult Income: female subgroup has lower base rate of high income, so predicting 'low income' as a constant happens to be more accurate than the ID-calibrated conditional prediction).

This gives us a handle on harm that is invisible to the TV-distance-based bound of Corollary A.13.

A.13.3 The two-signal decomposition

We combine the feature-collapse signal (our original λ) with a prediction-collapse signal that exploits Proposition A.15.

Definition A.17 (Two-signal quantities)**.** *Define the population quantities*

$\lambda := \mathbb{E}[\lambda_{\mathrm{ID}}] - \mathbb{E}[\lambda_{\mathrm{OOD}}]$, *(anchor-loss drop under eval batches; Proposition A.2)*
$\mu := \mathbb{E}_{(x,y)\sim p_0}[p_x(y)] - \mathbb{E}_{x\sim p_1}[p_x(\hat{y}(x))]$, *(mean confidence drop from anchor to eval).*

The sample estimators $\hat{\lambda}$*,* $\hat{\mu}$ *replace each population expectation with its empirical average over the relevant batch. Recall* $\hat{y}(x) = \arg\max_c p_x(c)$*, so* $p_x(\hat{y}(x)) = \max_c p_x(c)$ *is the maximum softmax probability (MSP) at* x*.*

Roles of the two signals. λ tracks distributional difference (Proposition A.2, Corollary A.4): it is large when features shift, regardless of whether the shift hurts. μ tracks predictive confusion: when the shift pushes eval points onto ϕ_∞, each eval MSP collapses to the constant $q_\infty(\hat{y}_\infty)$, and μ grows proportionally to the gap between ID accuracy and this constant.

Assumption A.18 (ID calibration)**.** *The initial classifier is approximately calibrated on p_0:*

$$\left|\mathbb{E}_{(x,y)\sim p_0}[p_x(y)] - \mathrm{acc}_{q_{W_0}}(p_0)\right| \le \mathrm{cal}_0, \qquad \left|\mathbb{E}_{x\sim p_0}[\max_c p_x(c)] - \mathrm{acc}_{q_{W_0}}(p_0)\right| \le \mathrm{cal}_0',$$

for small constants $\mathrm{cal}_0, \mathrm{cal}_0' \ll 1$.

The first inequality is standard expected-calibration-error on the true-label posterior; the second says the classifier's mean MSP on ID is close to its ID accuracy. Both hold for well-calibrated classifiers.

Assumption A.19 (OOD calibration at the collapse point)**.** *The limiting prediction q_∞ and the OOD label marginal $p_{Y|p_1}$ agree at the collapse class $\hat{y}_\infty$:*

$$\left|q_\infty(\hat{y}_\infty) - p_{Y|p_1}(\hat{y}_\infty)\right| \le \mathrm{cal}_\infty.$$

This is weaker than full calibration on p_1 (which is unavailable without OOD labels): it only constrains the classifier's confidence in its constant prediction to match the frequency of that class under p_1. We treat cal_∞ as a structural miscalibration that is dataset-dependent but bounded.

Assumption A.20 (Label marginal on OOD)**.** *The OOD label marginal is bounded away from a point mass:* $\max_c p_{Y|p_1}(c) < 1$.

Mixture model for partial shifts. To state the theorem, we model partial shifts as two-component mixtures. Let $p_1 = (1-\alpha)\,p_0 + \alpha\,p_\star$, where $\alpha \in [0,1]$ is the OOD fraction and $p_\star$ is an auxiliary far-OOD distribution satisfying Assumption A.14. Pure ID is $\alpha = 0$, pure OOD is $\alpha = 1$, and partial shifts are $\alpha \in (0,1)$. This parametrization captures a large class of distribution shifts and is convenient for tracking linear dependence on shift magnitude.

Proposition A.21 (Harm characterization)**.** *Under Assumptions A.14–A.20 and the mixture model $p_1 = (1-\alpha)p_0 + \alpha p_\star$, the mean confidence drop tracks the accuracy gap:*

$$\mu = \Delta(q_{W_0}, p_0, p_1) + O\big(\alpha\,\mathrm{cal}_\infty \;+\; \mathrm{cal}_0 + \mathrm{cal}_0' \;+\; o(1)\big). \tag{41}$$

In particular, if $\mathrm{cal}_0, \mathrm{cal}_0', \mathrm{cal}_\infty \to 0$, then $\mu \to \Delta$.

Proof. We compute μ and Δ separately for the mixture and compare.
Step 1 (anchor term of μ). By Assumption A.18 (first inequality),

$$\mathbb{E}_{(x,y)\sim p_0}[p_x(y)] = \mathrm{acc}_{q_{W_0}}(p_0) + O(\mathrm{cal}_0). \tag{42}$$

Step 2 (eval term of μ, ID component). Since $p_x(\hat{y}(x)) = \max_c p_x(c)$, the mixture gives

$$\mathbb{E}_{x\sim p_1}[\max_c p_x(c)] = (1-\alpha)\,\mathbb{E}_{x\sim p_0}[\max_c p_x(c)] + \alpha\,\mathbb{E}_{x\sim p_\star}[\max_c p_x(c)].$$

By Assumption A.18 (second inequality),

$$\mathbb{E}_{x\sim p_0}[\max_c p_x(c)] = \mathrm{acc}_{q_{W_0}}(p_0) + O(\mathrm{cal}_0'). \tag{43}$$

Step 3 (eval term of μ, OOD component). Under Assumption A.14, Proposition A.15 gives $q_{W_0}(\cdot \mid x) \to q_\infty$ in total variation for $p_\star$-a.e. x, so $\max_c q_{W_0}(c \mid x) \to q_\infty(\hat{y}_\infty)$. Since $\max_c q_{W_0}(c \mid x) \in [0,1]$ is uniformly bounded, dominated convergence yields

$$\mathbb{E}_{x\sim p_\star}[\max_c p_x(c)] = q_\infty(\hat{y}_\infty) + o(1). \tag{44}$$

Step 4 (combine eval terms). Substituting (43) and (44),

$$\mathbb{E}_{x\sim p_1}[\max_c p_x(c)] = (1-\alpha)\big[\mathrm{acc}(p_0) + O(\mathrm{cal}_0')\big] + \alpha\big[q_\infty(\hat{y}_\infty) + o(1)\big].$$

Step 5 (compute μ). Subtracting from (42),

$$\begin{aligned}\mu &= \mathbb{E}_{p_0}[p_x(y)] - \mathbb{E}_{p_1}[\max_c p_x(c)] \\ &= \big[\mathrm{acc}(p_0) + O(\mathrm{cal}_0)\big] - \big[(1-\alpha)(\mathrm{acc}(p_0) + O(\mathrm{cal}_0')) + \alpha(q_\infty(\hat{y}_\infty) + o(1))\big] \\ &= \alpha\left(\mathrm{acc}(p_0) - q_\infty(\hat{y}_\infty)\right) + O(\mathrm{cal}_0 + \mathrm{cal}_0') + \alpha\, o(1). \end{aligned} \tag{45}$$

Step 6 (compute Δ). Corollary A.16 applied to $p_\star$ gives $\mathrm{acc}_{q_{W_0}}(p_\star) = p_{Y|p_\star}(\hat{y}_\infty) + o(1)$. For the mixture,

$$\mathrm{acc}_{q_{W_0}}(p_1) = (1-\alpha)\,\mathrm{acc}_{q_{W_0}}(p_0) + \alpha\,\mathrm{acc}_{q_{W_0}}(p_\star) = (1-\alpha)\,\mathrm{acc}(p_0) + \alpha\, p_{Y|p_\star}(\hat{y}_\infty) + o(1),$$

hence

$$\Delta = \mathrm{acc}(p_0) - \mathrm{acc}_{q_{W_0}}(p_1) = \alpha\big(\mathrm{acc}(p_0) - p_{Y|p_\star}(\hat{y}_\infty)\big) + o(1). \tag{46}$$

Step 7 (compare). Subtract (46) from (45):

$$\mu - \Delta = \alpha\left(p_{Y|p_\star}(\hat{y}_\infty) - q_\infty(\hat{y}_\infty)\right) + O(\mathrm{cal}_0 + \mathrm{cal}_0') + o(1).$$

By Assumption A.19 applied to $p_\star$, $|p_{Y|p_\star}(\hat{y}_\infty) - q_\infty(\hat{y}_\infty)| \le \mathrm{cal}_\infty$, so

$$|\mu - \Delta| \le \alpha\,\mathrm{cal}_\infty + O(\mathrm{cal}_0 + \mathrm{cal}_0') + o(1),$$

which is (41). ∎

Sample-version corollary. The sample estimator $\hat{\mu}$ differs from the population μ by a Hoeffding-style concentration term (both summands take values in $[0,1]$):

$$|\hat{\mu} - \mu| = O\big(1/\sqrt{n_a} + 1/\sqrt{n_e}\big)$$

with high probability. Combining with Proposition A.21,

$$\hat{\mu} = \Delta + O\big(\alpha\,\mathrm{cal}_\infty \;+\; \mathrm{cal}_0 + \mathrm{cal}_0' \;+\; 1/\sqrt{n}\big), \tag{47}$$

where $n = \min(n_a, n_e)$.

The role of $\hat{\lambda}$. Proposition A.21 establishes that $\hat{\mu}$ *alone* estimates Δ up to calibration and sampling errors. The anchor-loss signal $\hat{\lambda}$ does not appear in the statement. Its role in the paper is auxiliary but practically important:

- *Robustness gate.* When $\hat{\mu}$ is small, it could mean either no harm ($\Delta \approx 0$) or severe miscalibration (the confidence-based estimator is unreliable). Checking $\hat{\lambda} > \tau_\lambda$ confirms that a distributional shift is present and therefore that $\hat{\mu}$ is being estimated in a non-trivial regime.
- *Calibration-free fallback.* $\hat{\lambda}$ requires no calibration assumptions; it fires under Assumption A.3 alone. If Assumptions A.18–A.19 are violated, $\hat{\lambda}$ can still detect distributional shift, though not necessarily harm.
- *Two-signal operating point.* A practical harm detector declares "harmful shift" when both $\hat{\lambda} > \tau_\lambda$ and $\hat{\mu} > \tau_\mu$ (Figure 9). The upper-right quadrant of the $(\hat{\lambda}, \hat{\mu})$ plane is the high-precision region; the lower-right contains benign shifts; the upper-left corresponds to miscalibrated estimators or rare silent-harm cases.

Definition A.22 (Composed two-signal detector)**.** *The composed detector alerts if and only if the conformal λ-test of Proposition A.11 rejects* and $\hat{\mu} > \tau_\mu$.

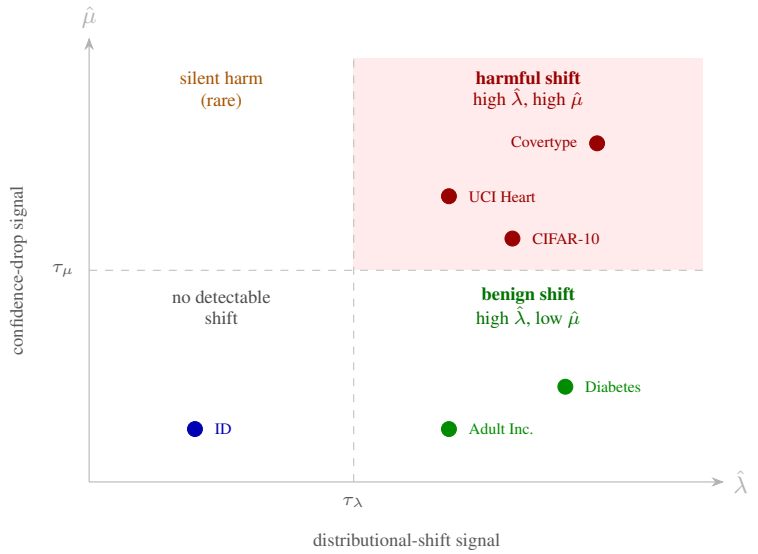


Figure 9: **Two-signal detector space.** Each dataset is a point in the $(\hat{\lambda}, \hat{\mu})$ plane. The current $\hat{\lambda}$-only detector uses only the horizontal axis and cannot distinguish benign shifts (right-lower) from harmful shifts (upper-right). By Proposition A.21, $\hat{\mu}$ alone estimates Δ up to calibration errors, so the vertical axis is the harm-magnitude estimator. Gating on $\hat{\lambda} > \tau_\lambda$ ensures that a distributional shift is actually present before trusting $\hat{\mu}$ (robustness gate); the conjunction $\hat{\lambda} > \tau_\lambda \wedge \hat{\mu} > \tau_\mu$ selects the upper-right region (red box). The upper-left "silent harm" quadrant is theoretically possible but rare in practice, because Nagler's tilting links feature shift to prediction confusion.

Remark A.23 (False-alarm containment under the null)**.** *Under H_0 the composed detector alerts only when both tests alert, so its false-alarm rate is at most the smaller of the two individual rates — for arbitrary dependence between $\hat{\lambda}$ and $\hat{\mu}$, and with no multiple-testing correction. In particular, the conformal bound $\alpha + 1/(K+1)$ of Proposition A.11 carries over to the composed detector unchanged.*

Remark A.24 (The gate never adds alarms)**.** *Under a shifted distribution the deployment and calibration rounds are no longer exchangeable, so no absolute false-alarm bound is possible. However, the $\hat{\mu}$ gate can only remove alerts produced by the λ-test, never add them; the composed detector's benign false-alarm rate is therefore bounded by that of the λ-test under* every *distribution.*

We evaluate the composed detector in Appendix D.9; λ remains the primary, conformally certified statistic.

Visualizing the two-signal plane. Proposition A.21 predicts that $\hat{\mu} \approx \Delta$ up to calibration and sampling errors, and that harmful shifts populate a specific region of the $(\hat{\lambda}, \hat{\mu})$ plane (Figure 9).

A.14 Summary of Assumptions

Table 2 collects all assumptions used in the appendix. Of the ten, only Assumption A.3 and Assumption A.19 are genuinely empirical (i.e., not provable from the setup alone). The remainder are design choices, standard regularity, or consequences derived from Nagler's tilting.

Key assumptions. Two assumptions are inherently empirical:

Table 2: Summary of assumptions used in the theoretical argument.

Label	Statement	Source
Core gradient-geometry assumptions (Theorem A.9)		
A1 (design)	Only the last-layer weight W is updated; backbone ϕ is frozen.	Analysis restriction; the deployed detector updates the full decoder (Appendix D.2).
A2 (design)	Pseudolabels $\hat{y}$ are fixed at W_0 and do not change during finetuning.	Pseudolabel step of Algorithm 1.
A3 (standard)	$\|\phi(x)\| \le B_\phi$ and softmax residuals bounded.	Standard for trained models.
Assn. A.3 (key)	$\|\langle \phi(x_a), \phi(x_e)\rangle\| \le \epsilon \|\phi(x_a)\| \|\phi(x_e)\|$ for $x_e \sim p_1$.	Consequence of Nagler [39, Thm. 6.3] in the far-OOD limit; genuine empirical claim for near-OOD.
Assn. A.5	Anchor correctness + eval confidence + centered, class-symmetric features + ID accuracy $\ge 1 - \tau$.	Holds for any useful PFN on ID data.
Assn. A.7	$T\eta L \le 1/2$ for L the smoothness constant of $\mathcal{L}$.	Verified numerically in Section A.9.
Harm-characterization assumptions (Proposition A.21)		
Assn. A.14	Far-OOD regime: the tilt $\psi_h(s; x_e)$ is approximately constant over $s \sim p_0$ for p_1-a.e. x_e.	Consequence of Nagler's tilting in the TV-limit $\|p_1 - p_0\|_{\mathrm{TV}} \to 1$.
Assn. A.18	ID calibration: mean true-label confidence $\mathbb{E}_{p_0}[p_x(y)]$ and mean MSP $\mathbb{E}_{p_0}[\max_c p_x(c)]$ both track ID accuracy up to small errors $\mathrm{cal}_0, \mathrm{cal}_0'$.	Holds for well-calibrated classifiers; measurable via held-out split.
Assn. A.19 (key)	OOD calibration at the collapse point: $\|q_\infty(\hat{y}_\infty) - p_{Y\|p_1}(\hat{y}_\infty)\| \le \mathrm{cal}_\infty$.	Genuine empirical claim. Not directly verifiable without OOD labels.
Assn. A.20	OOD label marginal is bounded away from a point mass.	Holds on all real benchmarks.

- **Assumption A.3 (feature orthogonality)** is the scope boundary of the *separation* argument (Theorem A.9). Where it fails, when OOD features retain substantial ID-like structure along the directions relevant for classification, the theorem does not guarantee $\mathbb{E}[\lambda_{\mathrm{OOD}}] < \mathbb{E}[\lambda_{\mathrm{ID}}]$.

- **Assumption A.19 (OOD calibration at the collapse point)** is the scope boundary of the *harm* argument (Proposition A.21). Where it fails, $\hat{\mu}$ no longer tracks Δ and can systematically over- or under-estimate harm. It is also the most delicate assumption: it is the only one that constrains behavior on the shifted distribution, and without OOD labels we cannot directly verify it.

Plausibility of the remaining harm-specific assumptions. Beyond the key pair, four further conditions are relevant to the harm theorem (Proposition A.21):

- Assumption A.18 (ID calibration) is routinely violated by modern deep networks (since they are often overconfident). A rigorous version of the theorem can work with (temperature-)scaled predictions using a held-out calibration set.

- Assumption A.14 (far-OOD regime) is an idealized limit. For near-OOD, Proposition A.15 holds only approximately, so Proposition A.21 degrades correspondingly; a formal near-OOD rate is left to future work.

- The mixture model $p_1 = (1 - \alpha)p_0 + \alpha p_\star$ used to state Proposition A.21 is a modeling choice. Real shifts may not decompose this way, but for general p_1 the same argument applies with α interpreted as the effective OOD mass $\int \mathbb{1}\{x \text{ is far-OOD}\}\, dp_1$.

- Nagler's tilting theorem is proved only for one-layer transformers. Extending the argument to multilayer transformers (which TabPFN and TabICL use) is a nontrivial open question. Until this is resolved, this part of the argument is a *mechanism hypothesis* rather than a proven property.

What holds unconditionally. The density-ratio decomposition of Lemma A.12 and the TV upper bound of Corollary A.13 are both exact and require no assumptions. They already suffice to rebut the framing "$\hat{\lambda}$ measures harm" and to motivate the need for a second signal beyond $\hat{\lambda}$.

B Scope and failure modes

B.1 Catalog of failure modes

A PFN at deployment can exhibit many distinct kinds of failure, and **CALD** is targeted at a specific subset of them. We describe the failure modes a PFN may encounter, and indicate for each one whether **CALD** is empirically evaluated against it and whether the theory of Appendix A covers it. The space decomposes along three axes (see Table 3):

- *Distribution shift* can be **covariate** ($p_0(x) \neq p_1(x)$ with $p(y \mid x)$ unchanged), **label** ($p_0(y) \neq p_1(y)$ with $p(x \mid y)$ unchanged), **concept** ($p(y \mid x)$ itself changes), or structured variants such as **subpopulation** (minority group reweighted) and **open-set** (novel classes appear at test time).
- Within distribution shift, the shift may be **harmful** (predictive accuracy drops) or **benign** (accuracy preserved or improved). **CALD** is calibrated to *harm* rather than to *distance*.
- *Non-distributional failure modes* sit outside the i.i.d. framing of conformal calibration entirely: schema mismatch (different feature columns or dtypes), adversarial inputs, insufficient support set, numerical issues (NaNs, out-of-range values), and miscalibration without distributional change.

Table 3: **Scope of CALD.** Failure modes a PFN may encounter at deployment, with whether **CALD** is empirically evaluated against it and whether the theory of Appendix A covers it. ✓ indicates full empirical coverage; *partial* indicates a heuristic check (e.g., a severity sweep without a formal guarantee); × indicates absence.

Failure mode	Tested	Theory	Notes
*Distribution shift – harmful (**CALD** should fire)*			
Covariate shift	✓	Thm. A.9 (far-OOD)	main use case
Subpopulation shift	✓	severity sweep (heuristic)	spurious attributes
Domain shift	✓	Thm. A.9 (far-OOD)	subset of covariate shift
Near-OOD (overlapping support)	partial	vacuous in Thm. A.9	empirical only
Distribution shift – benign by accuracy criterion			
Covariate shift, Δ small	✓	Lem. A.12	**CALD** *fires* via gradient signature; see §B.2
Out of scope			
Pure label shift ($p(y)$ only)	×	none	test statistic is feature-driven
Concept shift ($p(y \mid x)$)	×	none	undetectable without test labels
Open-set / novel class	×	partial (handled as far-OOD)	special case of label shift
Adversarial inputs	×	none	breaks i.i.d. calibration
Schema / engineering errors	×	none	pre-test sanity check, not a monitor
Predictive miscalibration	×	none	orthogonal to distributional shift

B.2 Empirical investigation: covariate shift without an accuracy drop

The benign-by-accuracy row of Table 3 deserves a closer look. Among the seven benign datasets of Table 1 (below the Hoeffding harm threshold $\Delta^\star \approx 4.84$ pp at $\delta = 0.05$, $N_{\text{exp2}} = 1280$, defined in Appendix C.3), four deserve a closer look:

$$\mathcal{D}_{\text{benign}} = \{\text{diabetes_readmission},\ \text{bank_marketing},\ \text{credit_default},\ \text{folktables}\}.$$

On these four datasets, the post-finetune anchor-loss test of **CALD** nevertheless rejects $H_0 : X^*_{\text{eval}} \sim p_0$ at high power for at least one of the two methods (cald-tabpfn, cald-tabicl) on every dataset. The full numbers are reproduced in Table 4.

Reading the table. The AUROC values for KNN, MMD, Mahalanobis, Energy, and MSP show that the four datasets are not edge cases: their ID and OOD distributions are easily distinguishable in feature space by every standalone shift detector we evaluate. The benign label is purely a property of the predictive accuracy gap $\Delta(q, p_0, p_1)$ as decomposed in Lemma A.12: it places strong constraints on how mass moves on the mistake region $M = \{(x, y) : \mathbb{1}\{y \neq \arg\max_c q_{W_0}(c \mid x)\} = 1\}$, but no constraints whatsoever on how mass moves on M^c. In all four datasets we observe heavy reweighting of M^c (which is most of the input domain when the pretrained classifier q_{W_0} is accurate on p_0): **CALD**, like every shift detector that uses information about M^c, picks this up.

Table 4: **CALD and standalone shift detectors on the four datasets where $\Delta < \Delta^\star$.** AUROC computed on the same calibration protocol as Table 10 (10 seeds, 50 calibration rounds, support and anchor sets fixed within each seed). Every distribution-shift detector that does not condition on harm (KNN, MMD, Mahalanobis, Energy, MSP) also fires on these datasets at AUROC ≥ 0.9, indicating real distributional shift in feature space.

	Δ_{acc} (pp)	cald-tabpfn	cald-tabicl	KNN	MMD	Mahal.	Energy	MSP
diabetes_readmission	-2.34	0.976	1.000	1.000	1.000	1.000	1.000	0.869
bank_marketing	$+4.37$	0.997	1.000	1.000	1.000	0.989	0.934	1.000
credit_default	$+2.81$	1.000	0.698	1.000	1.000	1.000	1.000	0.764
folktables	$+3.12$	$0.624^\dagger$	0.899	0.926	1.000	0.968	0.722	0.833

$\Delta_{\text{acc}} := \text{acc}(p_0) - \text{acc}(p_1)$ (positive = OOD accuracy drop). † The LR-stability heuristic (Algorithm 1) flags this cell as *unstable* and falls back to $\eta = 10^{-4}$, where the signal collapses; see §B.6.

B.3 Two sub-mechanisms within Theorem A.9

The above empirical observation is consistent with, and indeed predicted by, the multi-step main separation theorem. Recall the central inequality (Theorem A.9):

$$\mathbb{E}[\lambda_{\text{ID}}] - \mathbb{E}[\lambda_{\text{OOD}}] \;\geq\; T\eta\,\big(\gamma - \epsilon\, B_\phi^2\, B_C\big) - O(T^2\eta^2 L\, B_g^2), \tag{48}$$

where $\gamma := -\mathbb{E}_{p_0 \otimes p_0}[C_{ij} K_{ij}] > 0$ is the ID interference strength (Proposition A.6) and ϵ is the OOD feature-orthogonality parameter (Assumption A.3 and Corollary A.4). The detector fires when the right-hand side is positive, i.e. when

$$\gamma - \epsilon\, B_\phi^2\, B_C > 0. \tag{49}$$

The proof of Theorem A.9 (Appendix A.10) makes the firing condition slightly richer than (49) suggests. Specializing equation (34) to the T-step regime and decomposing the cross-term in polar form $\langle \bar{g}_a, \bar{g}_e \rangle_F = \|\bar{g}_a\|_F \|\bar{g}_e\|_F \cos\theta$, where θ is the angle between the time-averaged anchor and disagree gradients, yields

$$\begin{aligned}\mathbb{E}[\lambda_{\text{ID}} - \lambda_{\text{OOD}}] = T\eta \Big[&\mathbb{E}\big(\|\bar{g}_a\|_F \|\bar{g}_e^{\text{ID}}\|_F \cos\theta_{\text{ID}}\big) \\ &- \mathbb{E}\big(\|\bar{g}_a\|_F \|\bar{g}_e^{\text{OOD}}\|_F \cos\theta_{\text{OOD}}\big)\Big] \\ &+ O(T^2\eta^2 L\, B_g^2),\end{aligned} \tag{50}$$

under the leading-order identification $\bar{g}_a^{\text{ID}} \approx \bar{g}_a^{\text{OOD}} =: \bar{g}_a$ established in the theorem's proof. The integrated separation is therefore positive whenever

$$\underbrace{\mathbb{E}\big(\|\bar{g}_a\|_F \|\bar{g}_e^{\text{ID}}\|_F \cos\theta_{\text{ID}}\big)}_{\text{ID conflict, large negative}} \;<\; \underbrace{\mathbb{E}\big(\|\bar{g}_a\|_F \|\bar{g}_e^{\text{OOD}}\|_F \cos\theta_{\text{OOD}}\big)}_{\text{OOD signed cross-term}}. \tag{51}$$

The right-hand side can be made larger than the left-hand side by either of the following two mutually compatible mechanisms:

(M1) Cosine shift toward less conflict.
$\cos\theta_{\text{OOD}} > \cos\theta_{\text{ID}}$ with $\|\bar{g}_e^{\text{OOD}}\|_F \approx \|\bar{g}_e^{\text{ID}}\|_F$. Even a small positive shift $\Delta\cos := \cos\theta_{\text{OOD}} - \cos\theta_{\text{ID}}$ is enough to make (51) positive when θ_{ID} is in a strongly-conflicting regime ($\cos\theta_{\text{ID}} \to -1$).

(M2) Eval-gradient magnitude shrinkage.
$\|\bar{g}_e^{\text{OOD}}\|_F < \|\bar{g}_e^{\text{ID}}\|_F$ with $\cos\theta_{\text{OOD}} \approx \cos\theta_{\text{ID}}$ (or even $\cos\theta_{\text{OOD}} < \cos\theta_{\text{ID}}$, provided the magnitude attenuation dominates). This is the canonical Nagler-tilting picture of Appendix A.8: under far-OOD inputs the prediction collapses, the disagree loss saturates, and $\|\bar{g}_e\|_F \to 0$.

The original informal argument in the body of the paper (Section 4.1) emphasized M2; it presented OOD as a regime where "the PFN is uncertain on its own pseudolabels, $\nabla\mathcal{L}_{\text{eval}}$ is small in magnitude and approximately orthogonal to $\nabla\mathcal{L}_{\text{anc}}$". Theorem A.9's actual statement is more permissive: the firing condition (51) holds whenever *either* mechanism produces enough integrated effect. M1 is sufficient on its own. M2 is sufficient on its own. The two can also reinforce or partially cancel.

Why benign datasets satisfy (51)**.** On the benign-by-accuracy datasets, q_{W_0} classifies the OOD batch with accuracy comparable to or better than ID, so the prediction-collapse picture of Proposition A.15 does not apply: the model is confident on OOD, and $\|\bar{g}_e^{\mathrm{OOD}}\|_F$ is comparable to $\|\bar{g}_e^{\mathrm{ID}}\|_F$. M2 is muted. But the geometry of the OOD batch in feature space is still different: a non-trivial fraction of OOD points fall in regions where the kernel $K_{ij} = \langle \tilde{\phi}(x_a^i), \tilde{\phi}(x_e^j) \rangle$ has a distribution that is no longer centered exactly where it sits under p_0. This perturbs $\cos\theta$ by a small amount, which integrated over T steps suffices for (51) to hold.

The benign-firing cells therefore correspond to the γ-driven branch of (49): γ remains close to its ID value (p_0 is unchanged on the anchor side) and $\epsilon\, B_\phi^2\, B_C$ does not need to vanish for the bound to be active: a small $\Delta\cos$ produced by partial feature reorganization is enough.

B.4 Gradient-probe diagnostics

To verify the M1/M2 picture empirically, we instrument the inner finetune loop of **CALD** to log the per-step gradient quantities of equation (50). For each of the four benign datasets and two harmful controls (camelyon17, covertype), and for each of the two methods (cald-tabpfn, cald-tabicl), we run $n_{\mathrm{rounds}} = 10$ probe rounds at three seeds with the LR selected by the same heuristic as in Table 10. At every step $t = 1, \ldots, T = 50$ within every round we record $\|g_a^{(t)}\|_F^2$, $\|g_e^{(t)}\|_F^2$, $\langle g_a^{(t)}, g_e^{(t)} \rangle_F$, and $\mathcal{L}_{\mathrm{anc}}(W_t)$. We additionally log the pre-finetune cosine similarity $\cos\theta_0 = \langle g_a^{(0)}, g_e^{(0)} \rangle_F / (\|g_a^{(0)}\|_F \, \|g_e^{(0)}\|_F)$ on each round.

The probe code (`detection/viz/cald_gradient_probe.py`) calls into the reference implementation `detection/core/finetune.py:run_one_iteration` for the cald-tabpfn arm and uses a self-contained instrumented loop for cald-tabicl. Aggregating across seeds and rounds for each (dataset, method) cell, we report:

- $\lambda_{\mathrm{gap}} := \mathbb{E}[\lambda_{\mathrm{ID}}] - \mathbb{E}[\lambda_{\mathrm{OOD}}]$, the directly observed integrand of the conformal test (Appendix A.11);
- $\rho_e^{(50)} := \mathbb{E}\|\bar{g}_e^{\mathrm{OOD}}\|_F^2 / \mathbb{E}\|\bar{g}_e^{\mathrm{ID}}\|_F^2$, averaged over all 50 finetune steps and across rounds and seeds;
- $\Delta\cos := \mathbb{E}[\cos\theta_0^{\mathrm{OOD}}] - \mathbb{E}[\cos\theta_0^{\mathrm{ID}}]$, the pre-finetune cosine shift;
- $H_{\mathrm{ID}}, H_{\mathrm{OOD}}$, the mean entropy of $q_{W_0}(\cdot \mid x_e)$ on the eval batch (a check on whether the prediction-collapse picture of Proposition A.15 is active).

Table 5 reports these four quantities together with the conformal AUROC for the (dataset, method) cell.

Table 5: **Gradient-probe summary across the eight benign cells and four harmful-control cells.** $\rho_e^{(50)} < 1$ indicates eval-gradient magnitude shrinkage on OOD (mechanism M2). $\Delta\cos > 0$ indicates a cosine shift toward less conflict on OOD (mechanism M1). Cells with AUROC ≥ 0.99 exhibit at least one of M1, M2 with the correct sign (and often both); cells with AUROC < 0.7 exhibit neither.

dataset	method	kind	AUROC	Δ_{acc} (pp)	λ_{gap}	$\rho_e^{(50)}$	$\cos\theta_0^{\mathrm{ID}}$	$\cos\theta_0^{\mathrm{OOD}}$	$\Delta\cos$
Diab.	cald-tabpfn	benign	0.976	−2.34	+0.352	0.421	−1.000	−0.973	+0.027
Bank	cald-tabpfn	benign	0.997	+4.37	+0.058	0.909	−0.336	−0.335	+0.001
Credit	cald-tabpfn	benign	1.000	+2.81	+0.269	0.155	−0.336	−0.300	+0.036
Folktbl.	cald-tabpfn	benign	0.624	+3.12	−0.004	0.878	+0.992	+0.990	−0.002
Diab.	cald-tabicl	benign	1.000	−2.34	+0.085	0.982	−0.253	−0.240	+0.013
Bank	cald-tabicl	benign	1.000	+4.37	+0.155	0.749	+0.969	+0.985	+0.016
Credit	cald-tabicl	benign	0.698	+2.81	+0.012	0.948	+0.442	+0.440	−0.002
Folktbl.	cald-tabicl	benign	0.899	+3.12	+0.025	0.909	+0.483	+0.376	−0.108
Camel.	cald-tabpfn	harmful	0.675	+8.91	+0.069	0.718	+0.059	−0.231	−0.289
Camel.	cald-tabicl	harmful	1.000	+8.91	+0.180	0.719	−0.189	−0.611	−0.422
Covert.	cald-tabpfn	harmful	1.000	+83.52	+0.163	0.269	−0.697	+0.240	+0.937
Covert.	cald-tabicl	harmful	1.000	+83.52	+0.468	0.060	+0.075	+0.087	+0.011

Reading Table 5. The eight benign cells fall into a consistent pattern with respect to mechanism M1: in five of eight cells, $\Delta\cos > 0$ at magnitudes between $+0.001$ and $+0.036$. The exceptions are folktables (both methods, $\Delta\cos$ negative or zero) and credit_default/cald-tabicl ($\Delta\cos = -0.002$). On those exception cells, M2 carries the signal: $\rho_e^{(50)}$ is well below 1, indicating the magnitude attenuation. The only cell where neither mechanism is active is folktables/cald-tabpfn at the unstable LR, with $\rho_e^{(50)} = 0.878$ and $\Delta\cos = -0.002$ both close to neutral; the λ_{gap} is correspondingly -0.004, and the conformal test does not reject (§B.6).

The four harmful controls are consistent with M2 being the dominant mechanism: $\rho_e^{(50)}$ ranges from 0.060 to 0.719, all well below 1, while $\Delta\cos$ takes both signs depending on dataset. The covertype/cald-tabpfn cell is the cleanest example of M1 and M2 firing together, with the cosine flipping sign ($\cos\theta_0^{\text{ID}} = -0.697 \rightarrow \cos\theta_0^{\text{OOD}} = +0.240$, $\Delta\cos = +0.937$) and the magnitude collapsing by a factor of $\rho_e^{(50)} = 0.27$.

Per-seed consistency. The aggregate $\Delta\cos$ values reported above are seed-means; the underlying distribution of $\cos\theta_0$ is sensitive to the random init of the decoder, which can flip its sign across seeds. To guard against the pathology of two seeds with opposite signs averaging to zero, we also computed $\Delta\cos$ per-seed for the eight benign cells (24 seed–cell pairs). On diabetes_readmission and bank_marketing, $\Delta\cos$ is positive in all six seeds for both methods; on credit_default/cald-tabpfn, two of three seeds have $\Delta\cos > 0$ and the third is small negative (-0.020). On folktables/cald-tabpfn, $\Delta\cos$ is essentially zero in all three seeds. The aggregate signs in Table 5 are not seed-mean artifacts.

B.5 Case studies

We highlight three (dataset, method) cells that span the qualitative regimes.

Case 1: diabetes_readmission/cald-tabpfn (M1, benign). Despite $\Delta_{\text{acc}} = -2.34$ pp (model is more accurate on OOD than ID), the gradient probe records a λ_{gap} of $+0.352$, larger than every harmful control except covertype/cald-tabicl. The mechanism is purely M1: $\rho_e^{(50)} = 0.42$ shows mild magnitude attenuation but $\cos\theta_0^{\text{ID}} = -1.000$ vs $\cos\theta_0^{\text{OOD}} = -0.973$ shows the cosine moving very slightly out of the strict-conflict regime. Integrated over $T = 50$ steps with $\eta = 10^{-2}$, this $\Delta\cos = +0.027$ is enough to satisfy (51). The pre-finetune entropy $H_{\text{OOD}} = 0.297$ is in fact slightly *below* $H_{\text{ID}} = 0.318$, ruling out the prediction-collapse picture for this dataset.

Figure 10 plots $\|g_e^{(t)}\|_F^2$, $\langle g_a^{(t)}, g_e^{(t)}\rangle_F$, and $\mathcal{L}_{\text{anc}}(W_t)$ versus t for ID and OOD on this cell. The middle panel is the diagnostic visual: the cross-term on ID descends to ≈ -80 over the trajectory while the cross-term on OOD plateaus near -10. Anchor descent on OOD therefore proceeds with five-fold less integrated interference than on ID, and the right panel shows the consequence: $\mathcal{L}_{\text{anc}}(W_T)$ on OOD reaches ≈ 0.4 while on ID it stays around 0.75.

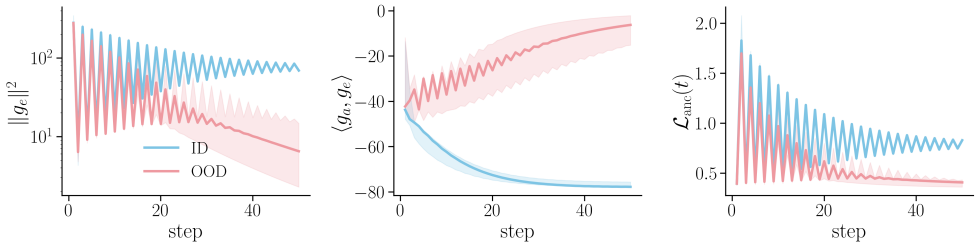


Figure 10: **Mechanism M1 on a benign dataset (diabetes_readmission/cald-tabpfn).** Per-step trajectories of $\|g_e\|_F^2$ (left), $\langle g_a, g_e\rangle_F$ (middle), and $\mathcal{L}_{\text{anc}}(W_t)$ (right), median across rounds and seeds with interquartile shading. The cross-term diverges between ID (more negative) and OOD (less negative), driving the integrated $\lambda_{\text{gap}} = +0.352$ even though $\Delta_{\text{acc}} = -2.34$ pp.

Case 2: covertype/cald-tabpfn (M1+M2, harmful). The two mechanisms both fire and reinforce each other. $\rho_e^{(50)} = 0.269$ shows strong magnitude collapse (M2): the OOD prediction is near-uniform on 7 classes (consistent with $H_{\text{OOD}} = 1.479$ approaching the maximal entropy $\log 7 \approx 1.95$), and the disagree loss saturates. Simultaneously, $\Delta\cos = +0.937$ (M1): the cosine flips from anti-aligned

($\cos\theta_0^{\mathrm{ID}} = -0.697$) to mildly aligned ($\cos\theta_0^{\mathrm{OOD}} = +0.240$). The per-step trajectories (Figure 11) are textbook for both mechanisms.

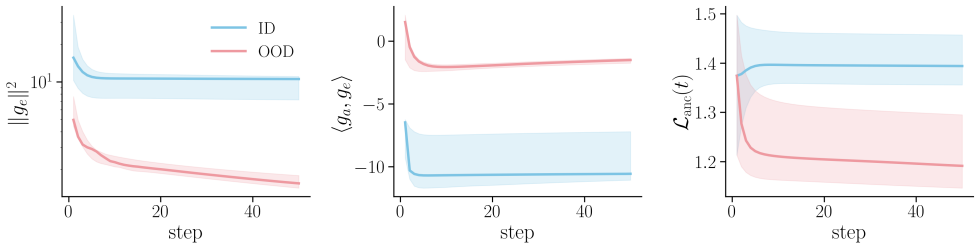


Figure 11: **Mechanisms M1+M2 firing together on a harmful dataset (covertype/cald-tabpfn).** The OOD eval-gradient magnitude is uniformly an order of magnitude smaller than ID throughout the trajectory (left), the cross-term sign on OOD is much closer to zero than on ID (middle), and the anchor loss on OOD descends to a clearly distinct level (right).

Case 3: camelyon17/cald-tabicl (M2 only, harmful). The cosine on this cell moves *further* into the conflict regime ($\Delta\cos = -0.422$) but the AUROC is 1.000. The mechanism is pure M2: $\rho_e^{(50)} = 0.719$ (eval gradient $\sim 30\%$ smaller on OOD). Mechanism M1's contribution to (51) is negative on this cell, but the integrated M2 contribution dominates: the magnitude attenuation reduces the absolute size of the OOD cross-term in (50) below the ID cross-term, even though the OOD angle is more anti-aligned.

Case 4: diabetes_readmission/cald-tabpfn vs. covertype/cald-tabpfn (head-to-head). Figure 12 places the cleanest M1 example side-by-side with the cleanest M1+M2 example. Both produce $\lambda_{\mathrm{OOD}} < \lambda_{\mathrm{ID}}$ with high confidence; the gradient signatures are visibly distinct.

Figure 13 plots λ_{gap} against Δ_{acc} for all twelve cells in Table 5. The visualization makes the central claim concrete: the largest λ_{gap} in the benign region is $+0.352$ (diabetes_readmission/cald-tabpfn), which exceeds the λ_{gap} of the harmful camelyon17 (both methods). The two axes are not monotonically related on this 12-cell sample.

B.6 Fallback behavior of the LR-stability rule

The single benign cell in Table 5 where the conformal test does not reject is folktables/cald-tabpfn, with AUROC 0.624, $\lambda_{\mathrm{gap}} = -0.004$, $\rho_e^{(50)} = 0.878$, and $\Delta\cos = -0.002$. Both mechanisms M1 and M2 are dormant here: the gradient signature of equation (51) is essentially neutral, and the detector's reported AUROC is the only one below 0.7 across the eight benign cells.

The low AUROC and the fallback have a common cause. Algorithm 1 selects a finetuning learning rate per (dataset, method) cell with a tail-stability rule: it checks the per-step direction indicator $\mathbb{1}\{\mathbb{E}\lambda_{\mathrm{OOD}}^{(t)} < \mathbb{E}\lambda_{\mathrm{ID}}^{(t)}\}$ for $t \in [30, 50]$ across calibration rounds, and rejects a candidate η if the indicator flips more than twice or has tail-mean below 0.5. On folktables/cald-tabpfn the rule rejects $\eta = 10^{-2}$ and falls back to $\eta = 10^{-4}$, flagging the cell `unstable`; by (49), when neither M1 nor M2 produces a coherent integrated effect, $\bar{g}_a$ and $\bar{g}_e$ are too noisy to define a reliable post-finetune λ, and the rule registers this as tail oscillation. We report the modest AUROC on this cell as a measured limitation of the detector in this regime: the stability rule localizes the unreliable cell, but it does not repair it.

A census over all 30 method $\times$ dataset cells shows the fallback is rare and confined to benign shifts: four cells fall back to $\eta = 10^{-4}$ (adult_income on both models, at AUROC 1.000 despite the fallback; celebA/cald-tabicl at 0.845; and folktables/cald-tabpfn at 0.624), while every severely shifted dataset selects its learning rate stably. The selection is also insensitive to the choice of tail window: re-running the rule with windows 25–45, 30–50, and 35–50 changes the selected rate in 1 of 60 cell–window comparisons (folktables/cald-tabpfn under 25–45, ΔAUROC -0.052). Appendix D.5 reports the full census together with the frozen-configuration comparison.

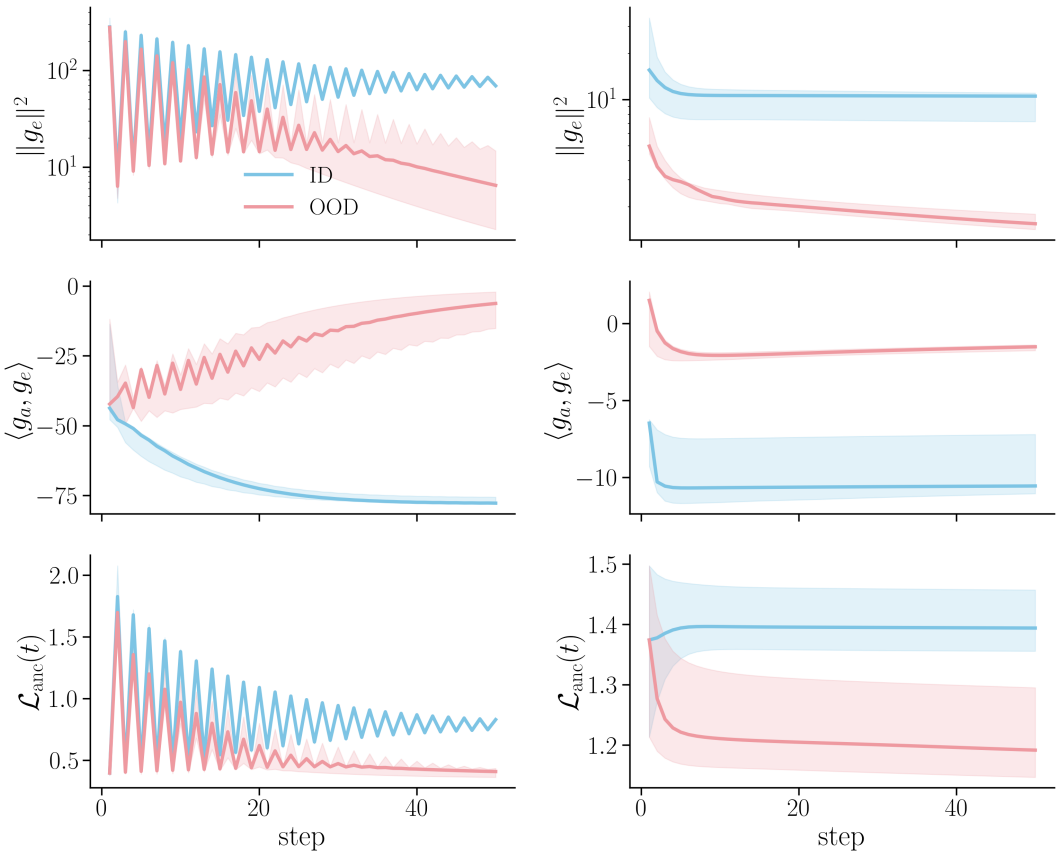


Figure 12: **Same detector outcome ($\lambda_{\text{OOD}} < \lambda_{\text{ID}}$), two different mechanisms.** Left column: diabetes_readmission (benign, $\Delta_{\text{acc}} = -2.34$ pp, AUROC 0.976) fires via cosine shift (M1). Right column: covertype (harmful, $\Delta_{\text{acc}} = +83.5$ pp, AUROC 1.000) fires via the combined M1+M2 mechanism, with a much cleaner log-scale separation in $\|g_e\|_F^2$.

B.7 Implications for the calibration-to-harm claim

Section 4.3 and §5.3 of the main text present the empirical observation that AUROC scales monotonically with shift severity (calibration to harm). The probe data above show that this monotonicity does not hold on the four-dataset benign group; in particular, the diabetes_readmission/cald-tabpfn cell has $\lambda_{\text{gap}} = +0.352$ at $\Delta_{\text{acc}} = -2.34$ pp, while camelyon17/cald-tabpfn has $\lambda_{\text{gap}} = +0.069$ at $\Delta_{\text{acc}} = +8.91$ pp.

The right interpretation, supported by Lemma A.12, is that λ_{gap} and Δ_{acc} measure formally different objects:

- Δ_{acc} depends only on how the shift moves mass on the mistake region M (Lemma A.12). In particular, Δ_{acc} is invariant under any reweighting of M^c that preserves the marginal on M.
- λ_{gap} depends on the integrated cross-term (50), which is a feature-space quantity. Reweighting that occurs entirely on M^c still perturbs the geometry of $\bar{g}_a, \bar{g}_e$ and therefore still affects λ_{gap}.

Two shifts with the same $|\Delta_{\text{acc}}|$ can therefore produce very different λ_{gap} values. The four benign-by-accuracy datasets are the regime where $|\Delta_{\text{acc}}|$ is small but feature-space reweighting on M^c is large; the harm decomposition of Lemma A.12 predicts exactly this kind of dissociation.

The accuracy-sensitive auxiliary signal $\hat{\mu}$ proposed in Proposition A.21 is the formal vehicle for restoring calibration to harm: $\hat{\mu}$ is constructed to track the density-ratio integral of Lemma A.12 directly, rather than the unsigned TV distance that drives λ. A natural reading of the present appendix's empirical findings is that $\hat{\lambda}$ and $\hat{\mu}$ should be reported jointly: $\hat{\lambda}$ confirms shift, $\hat{\mu}$ adjudicates whether the shift is harmful. The dissociation observed on the four benign datasets is then not a defect of

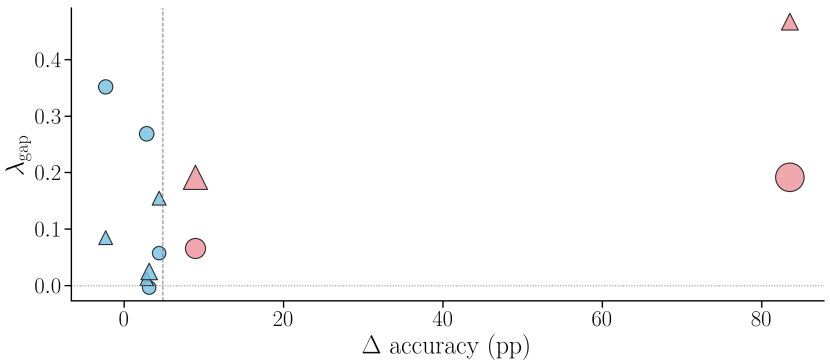


Figure 13: λ_{gap} **versus** Δ_{acc} **across the twelve probe cells.** Marker size scales with $|\Delta\cos|$. The dashed vertical line is the Hoeffding harm threshold $\Delta^\star \approx 4.84$ pp. Points to the left of the threshold are benign; points to the right are harmful. Both regions contain cells with $\lambda_{\mathrm{gap}} \geq 0.05$, confirming that the conformal test fires on both kinds of distributional shift.

CALD but a quantitative reading of the same $\hat{\lambda}/\hat{\mu}$ separation that the harm theorem (Proposition A.21) makes formal.

C Experimental Details

This appendix collects dataset-, baseline-, and protocol-level details deferred from Section 5, together with a survey of TabPFN/TabICL accuracy under the shifts considered. §C.1 describes the 14 datasets and their ID/OOD splits; §C.2 catalogs the 15 post-hoc baselines we compare against; §C.3 surveys the per-dataset accuracy drops and applies a Hoeffding bound to classify each shift as benign or harmful; and §C.4 reports our main TPR@5% FPR results with confidence intervals and explains the LR-stability protocol used for **CALD**.

C.1 Dataset Description

We evaluate on 14 datasets covering tabular, vision-foundation-model, and text-foundation-model representations. For each dataset we describe the modality and the criterion that splits in-distribution (ID) from out-of-distribution (OOD); in many cases the OOD pool corresponds to the dataset's own test split, with the criterion baked in by the dataset designers. Sample-size and protocol details are deferred to §C.4.

1. **UCI Heart Disease.** Tabular medical features (13 covariates: age, sex, chest-pain type, cholesterol, ECG measurements, exercise-induced angina, etc.). *ID/OOD:* Cleveland Clinic patients are ID, Hungarian Institute of Cardiology patients are OOD. Split is *site* shift: different hospital, different population, different measurement conventions, with non-trivial covariate redistribution across the 13 fields.
2. **Diabetes Readmission** (UCI #296). Tabular hospital records (49 covariates: demographics, admission codes, lab counts, medication lists, length of stay). *ID/OOD:* `admission_type` = *Emergency* is ID, `Elective` is OOD. Split is *patient-acuity* shift: emergency admissions skew toward acute conditions and shorter stays, elective toward planned procedures and longer stays, redistributing nearly every covariate.
3. **Adult Income** (UCI #2). Tabular 1994 US Census records (14 covariates: age, work class, education, occupation, hours-per-week, etc.). *ID/OOD:* `sex` = *Male* is ID, *Female* is OOD. Split is *demographic* shift: in 1994 the labor-force composition for men and women differs sharply on hours, occupation, and capital gains, so the covariate distributions are far from identical.
4. **Bank Marketing** (UCI #222). Tabular Portuguese-bank telemarketing call records (20 covariates including age, job, balance, contact channel, prior-campaign outcome, and macroeconomic indicators). *ID/OOD:* calls in *May–August 2008* are ID, calls in *October–November 2008* are OOD. Split is *temporal/macroeconomic* shift bracketing the September 2008 Lehman collapse, after which euro-zone interest rates and consumer-confidence indicators in the dataset shift abruptly.
5. **Credit Default** (UCI #350). Tabular Taiwanese credit-card client records (23 covariates: limit balance, sex, age, education, six months of bill amounts and payment statuses). *ID/OOD:* `EDUCATION` = *University* is ID, *Graduate school* is OOD. Split is *education-stratum* shift: graduate-degree holders systematically have higher credit limits and different payment patterns.
6. **Covertype** (sklearn `fetch_covtype`). Tabular forest-cover records from Roosevelt National Forest (54 covariates: 10 quantitative including elevation/slope/aspect/distances, 4 binary wilderness areas, 40 binary soil types). *ID/OOD: 2,000–2,500 m* elevation is ID, *>3,000 m* is OOD. Split is *ecological-zone* shift: above 3,000 m the forest transitions from mixed broadleaf–conifer to pure subalpine conifer, redistributing soil type and slope.
7. **Folktables ACS Income**. Tabular American Community Survey records (10 covariates: age, work class, education, marital status, occupation, etc.). *ID/OOD:* we use a state-pair split following the Folktables release. Split is *geographic* shift: different states have different industrial mixes and educational distributions.
8. **Camelyon17** (Bandi et al. 4, Koh et al. 27). Whole-slide histopathology image patches encoded by a frozen DINOv2 ViT-B/14 to 768-dim representations. *ID/OOD:* patches from three of the five WILDS hospitals are ID, the held-out OOD hospital is OOD. Split is *hospital/staining* shift: different scanners and H&E staining protocols produce systematic color and texture differences.
9. **CIFAR-10 $\rightarrow$ CIFAR-10.1** (Recht et al. 50, 51). Natural-image patches encoded by an ImageNet-pretrained ResNet-18 to 512-dim representations. *ID/OOD:* CIFAR-10 test set is ID, CIFAR-10.1

v6 is OOD. Split is *benchmark-replication* shift: Recht et al. [51] re-collected CIFAR-10 following the original protocol; the resulting class marginals match but the covariate distributions diverge enough to drop most classifiers' accuracy by 6–10 pp.

10. **PACS** (Li et al. 29). Object images from four artistic domains, encoded by frozen DINOv2 ViT-B/14 to 768-dim representations. *ID/OOD: Photo* is ID, *Art-painting / Cartoon / Sketch* are OOD. Split is *style/domain* shift, classically used as a domain-generalization benchmark.
11. **Waterbirds** (WILDS, Koh et al. 27). Bird photographs encoded by frozen DINOv2 ViT-B/14 to 768-dim representations. *ID/OOD:* the majority groups (waterbird-on-water, landbird-on-land) are ID; the minority groups (waterbird-on-land, landbird-on-water) are OOD. Split is *spurious-correlation* shift: the background–label correlation in the training distribution is broken at test time, and the OOD test set is a known failure mode for ERM-trained classifiers.
12. **CelebA** (WILDS subpopulation, Koh et al. 27, Sagawa et al. 55). Face photographs encoded by frozen DINOv2 ViT-B/14 to 768-dim representations. *ID/OOD:* with target y = blond hair and attribute a = male, the groups with $y = a$ (*non-blond female / blond male*) are ID and the groups with $y \neq a$ (*non-blond male / blond female*) are OOD. Split is a *group-composition* shift that reverses the hair-color–gender pairing between ID and OOD; it is not the WILDS majority/minority split.
13. **CivilComments** (WILDS, Koh et al. 27). Online-comment text encoded by Jina-v3 to 1024-dim representations. *ID/OOD:* comments with *no* demographic-identity mention (LGBTQ, Muslim, Black, etc.) are ID; comments with at least one identity mention are OOD. Split is *topic-stratum* shift: identity-mentioning comments are systematically more toxic in the training distribution, making the topic–toxicity correlation a known shortcut.
14. **Amazon Reviews** (WILDS, Koh et al. 27). Customer-review text encoded by Jina-v3 to 1024-dim representations, 5-class star rating. *ID/OOD:* a held-out 20% of training-set reviewers are ID; the WILDS test set, drawn from a disjoint reviewer pool, is OOD. Split is *reviewer/domain* shift: different reviewers favor different vocabulary, sentence length, and star-distribution profiles.

C.2 Baselines

The goal of every baseline below is the same as **CALD**'s: given the unlabeled OOD pool described in §C.1, decide whether the eval batch is drawn from the support distribution. Each baseline computes a scalar score per round; we calibrate a 5%-FPR threshold on the ID-round scores and report TPR on the OOD-round scores. Notation: X_{anc} and y_{anc} denote the labeled anchor batch with PFN softmax outputs $p_a \in \Delta^{C-1}$, X_{eval} denotes the eval batch with softmax outputs p, and $h(x) \in \mathbb{R}^d$ denotes the PFN's pre-decoder representation. We adapt the conventional "low score = OOD" direction so that AUROC is reported in $[0.5, 1]$; baseline implementations live at `detection/baselines/posthoc.py`.

1. **MSP** [19]. Maximum softmax probability, $s_{\text{MSP}}(x) = \max_c p(c \mid x)$. Confident ID predictions score high; uncertain OOD predictions score low.
2. **Entropy** [44]. Predictive entropy, $s_{\text{Ent}}(x) = -\sum_c p(c \mid x) \log p(c \mid x)$. Higher entropy is taken as more OOD.
3. **Energy** [34]. Negative log-sum-exp of the logits, $s_{\text{En}}(x) = -\log \sum_c e^{z_c(x)}$, the softmax-free analogue of MSP. ID samples produce a peakier logit distribution and therefore a more negative energy.
4. **Mahalanobis** [28]. With a class-conditional Gaussian fit on the support representations, $s_{\text{MH}}(x) = \min_c (h(x) - \mu_c)^\top \Sigma^{-1} (h(x) - \mu_c)$. OOD samples lie far from every class cluster.
5. **Relative Mahalanobis** (RMD, Ren et al. 52). RMD subtracts a class-agnostic background term, $s_{\text{RMD}}(x) = \min_c \text{MD}_c(x) - \text{MD}_{\text{bg}}(x)$ with $\text{MD}_{\text{bg}}(x) = (h(x) - \mu)^\top \Sigma^{-1} (h(x) - \mu)$. Cancelling population-level density isolates the discriminative signal and improves near-OOD detection where standard MD saturates.
6. **KNN** [57]. Average ℓ_2 distance from the eval representation to its $k = 5$ nearest neighbors among the support representations, $s_{\text{KNN}}(x) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} \|h(x) - h(x_i)\|_2$. A non-parametric distance-to-feature-cloud measure.

7. **ViM** [64]. Virtual-Logit Matching decomposes pre-head representations into a classifier-subspace component and a null-space residual; the OOD score combines residual norm with energy, $s_{\mathrm{ViM}}(x) = \alpha\|\Pi^{\perp}(h(x) - u)\|_2 - \log\sum_c e^{z_c(x)}$, where $\Pi^{\perp}$ projects onto the bottom eigendirections of the ID covariance and $u = -W^{+}b$ is the virtual origin. ID samples lie within the classifier's principal subspace; OOD samples carry significant null-space mass.
8. **BBSD-Softmax** [32, 49]. Black-Box Shift Detection on softmax probabilities: per-class two-sample Kolmogorov–Smirnov tests between anchor and eval, take the minimum p-value, apply Bonferroni correction, return $1 - p_{\mathrm{corr}}$. Sensitive to label-marginal shifts.
9. **BBSD-Embed** (BBSD-E). Same procedure as BBSD-Softmax but applied per-dimension on the pre-decoder embeddings $h(x)$ instead of the softmax. Sensitive to feature-distribution shifts that may not propagate to the labels.
10. **D-MMD** [14, 33]. Maximum Mean Discrepancy between support and eval feature distributions, $\widehat{\mathrm{MMD}}^2(P, Q) = \frac{1}{n^2}\sum_{i,j} k(x_i, x_j) + \frac{1}{m^2}\sum_{i,j} k(y_i, y_j) - \frac{2}{nm}\sum_{i,j} k(x_i, y_j)$, with an RBF kernel at the median pairwise-distance bandwidth. The classical batch-level two-sample test.
11. **H-Divergence** (H-Div, Ben-David et al. 6, Lopez-Paz and Oquab 35). Trains a logistic-regression domain classifier to distinguish support representations from eval representations; the held-out classification accuracy serves as the OOD score. An accuracy near 0.5 indicates indistinguishable distributions.
12. **GradNorm** [23]. Gradient norm of the KL divergence to uniform, $s_{\mathrm{GN}}(x) = \|\nabla_W \mathrm{KL}(\mathbf{u} \,\|\, p_\theta(\cdot \mid x))\|_F$. Confident ID samples produce large gradients toward uniform; near-uniform OOD predictions produce small gradients.
13. **GradOrth** [5]. At the last linear layer $\nabla_W L = \mathrm{error} \otimes h$, so the gradient lives in the row span of the representations; GradOrth projects $h(x)$ onto the top-k singular subspace of the ID-representation matrix and uses the projection norm as an ID-likelihood score (low projection $\Rightarrow$ OOD).
14. **GAIA-Z** [8]. Activity of the gradient of the maximum logit at an internal hidden activation, $s_{\mathrm{GAIA}}(x) = \mathbb{E}\big[\big|\partial \max_c z_c(x)/\partial h(x)\big|\big]$. Originally counts exact-zero gradient channels under ReLU; we adapt to the GELU decoder by reporting mean absolute gradient magnitude. ID samples activate more channels and yield larger mean activity.
15. **WeiPer** [13]. Mean MSP under random weight perturbations: for r samples of unit-direction noise Δ_r scaled to $\delta\|w_c\|$ added to each row of the classifier head, $s_{\mathrm{WeiPer}}(x) = \frac{1}{r}\sum_r \max_c \mathrm{softmax}(W + \Delta_r)h(x)$. ID predictions are stable under perturbation and retain a high max-softmax; OOD predictions collapse.

C.3 Performance Deterioration

A deterioration detector is only meaningful when shift is genuinely harmful. Table 6 reports TabPFN accuracy on ID and OOD splits for all 14 datasets (10 rounds, eval size 128, $N = 1{,}280$ pooled samples per dataset). To classify each shift as *benign* or *harmful* we use a principled threshold derived from Hoeffding's inequality: for a fixed classifier evaluated on N i.i.d. samples,

$$\Pr\Big[\hat{a}_{\mathrm{ID}} - \hat{a}_{\mathrm{OOD}} > \Delta^* \;\Big|\; \text{no shift}\Big] \le e^{-N(\Delta^*)^2}, \tag{52}$$

which is distribution-free and applies to any non-adaptive classifier. Setting the right-hand side to $\delta = 0.05$ gives

$$\Delta^*(N) = \sqrt{\ln(1/\delta)/N}. \tag{53}$$

For $N = 1{,}280$ this yields $\Delta^* \approx 4.8\%$; any observed accuracy drop exceeding this threshold is statistically distinguishable from sampling noise at 95% confidence under the null of no shift, and we therefore label the dataset as **harmful**. Datasets whose accuracy gap falls below Δ^* in absolute value (including those where OOD accuracy is *higher* than ID) are labeled **benign**: the model's predictive integrity is preserved up to sampling noise, and a label-blind detector that fires on these datasets is incurring an unnecessary false alarm. The Hoeffding threshold is intentionally conservative: it makes no parametric assumption about the accuracy distribution, only bounded support, and is a per-dataset quantity (Δ^* scales as $1/\sqrt{N}$). Our benchmark spans the full severity spectrum, from catastrophic (**Covertype**, $\Delta = 83.5\%$) through moderate (**Camelyon17**, $\Delta = 8.9\%$) to negligible-or-benign (**Amazon**, $\Delta = +1.5\%$; **Adult Income**, $\Delta = -9.5\%$ where OOD accuracy is higher than ID), providing a demanding testbed for both correct alarm and correct silence.

Table 6: **TabPFN accuracy under distribution shift, ordered by severity.** $\Delta = \mathrm{acc}_{\mathrm{ID}} - \mathrm{acc}_{\mathrm{OOD}}$ (positive = harmful drop). $\Delta^*(N) = \sqrt{\ln(1/\delta)/N}$ is the Hoeffding 95%-confidence noise floor (Eq. 52, $\delta = 0.05$, $N = 1{,}280$); shifts with $|\Delta| \leq \Delta^*$ are labeled *benign* (†). Accuracy estimated over 10 rounds of 128 eval samples each.

Dataset	ID Acc	OOD Acc	Δ (%)	Δ^* (%)
Benign ($\lvert\Delta\rvert \leq \Delta^*$)				
Adult Income†	0.820	0.914	−9.5	4.8
Diabetes Readm.†	0.886	0.909	−2.3	4.8
Amazon†	0.645	0.630	+1.5	4.8
Credit Default†	0.847	0.819	+2.8	4.8
Folktables†	0.799	0.768	+3.1	4.8
Bank Marketing†	0.902	0.859	+4.4	4.8
CelebA†	0.952	0.908	+4.5	4.8
Harmful ($\Delta > \Delta^*$)				
CivilComments	0.918	0.844	+7.4	4.8
Camelyon17	0.953	0.864	+8.9	4.8
Waterbirds	0.988	0.871	+11.6	4.8
CIFAR-10→10.1	0.795	0.670	+12.4	4.8
UCI Heart	0.824	0.668	+15.5	5.0
PACS	0.998	0.720	+27.9	4.8
Covertype	0.904	0.069	+83.5	4.8

Table 7: **Full TPR@5% FPR results across 7 benign shifts ($\Delta \leq 4.8\%$, Hoeffding bound at $N = 1{,}280$, Appendix C.3; TPR = correct silence rate).** Each cell is mean $_{\text{std}}$ across 10 seeds (seeds 42–51); std is the spread of per-seed TPR (each seed's threshold is its own 5th-percentile ID-score). Means match Table 1 exactly. Bold = best mean in column.

Method Δ (%)	Adult −9.4	Diab. −2.3	Amazon +1.5	Credit +2.8	Folktbl. +3.1	Bank +4.4	CelebA +4.5
MSP	.00 $_{.00}$	.64 $_{.42}$	.78 $_{.18}$	.55 $_{.31}$	.53 $_{.28}$	.00 $_{.00}$	.63 $_{.26}$
Entropy	.00 $_{.00}$	.61 $_{.42}$	.80 $_{.15}$	.61 $_{.30}$	.50 $_{.24}$	.00 $_{.00}$	.39 $_{.22}$
Energy	.00 $_{.00}$	.00 $_{.00}$	.77 $_{.19}$	.00 $_{.00}$	.86 $_{.17}$	**.42** $_{.45}$	.04 $_{.07}$
Mahal.	.00 $_{.00}$	.00 $_{.00}$	.71 $_{.19}$	.00 $_{.00}$	.12 $_{.11}$	.05 $_{.08}$	.00 $_{.00}$
RMD	.69 $_{.32}$	.59 $_{.39}$	.87 $_{.13}$	**.92** $_{.11}$	.72 $_{.24}$	.05 $_{.08}$	.42 $_{.20}$
KNN	.00 $_{.00}$	.00 $_{.00}$	.78 $_{.15}$	.00 $_{.00}$	.23 $_{.13}$	.00 $_{.00}$	.00 $_{.00}$
ViM	.00 $_{.00}$	.10 $_{.30}$	.87 $_{.13}$	.00 $_{.00}$	.63 $_{.28}$	.13 $_{.13}$	.81 $_{.23}$
BBSD-S	.00 $_{.00}$	.40 $_{.36}$	.75 $_{.29}$	.76 $_{.29}$	.71 $_{.21}$	.02 $_{.06}$	.64 $_{.40}$
BBSD-E	.00 $_{.00}$	.00 $_{.00}$	.76 $_{.27}$	.67 $_{.28}$	.66 $_{.34}$	.00 $_{.00}$	.49 $_{.43}$
D-MMD	.00 $_{.00}$	.00 $_{.00}$	.66 $_{.14}$	.00 $_{.00}$	.00 $_{.00}$	.00 $_{.00}$	.54 $_{.32}$
H-Div	.00 $_{.00}$	.00 $_{.00}$	.19 $_{.38}$	.00 $_{.00}$	.10 $_{.18}$	.32 $_{.17}$	.00 $_{.00}$
GradNorm	.00 $_{.00}$	.44 $_{.47}$	.85 $_{.19}$	.01 $_{.03}$	.78 $_{.18}$	.27 $_{.42}$	.60 $_{.19}$
GradOrth	.83 $_{.21}$	**.77** $_{.18}$	.83 $_{.17}$	.69 $_{.18}$	.86 $_{.18}$	.02 $_{.06}$	.46 $_{.28}$
GAIA-z	.00 $_{.00}$	.07 $_{.21}$	—	.01 $_{.03}$	.80 $_{.20}$	.29 $_{.36}$	—
WeiPer	.18 $_{.36}$	.17 $_{.35}$	.89 $_{.10}$	.00 $_{.00}$	.67 $_{.30}$	.17 $_{.35}$	.81 $_{.19}$
CALD (TabPFN)	**1.00** $_{.00}$	.08 $_{.25}$	**.92** $_{.04}$	.00 $_{.00}$	**.98** $_{.03}$	.02 $_{.04}$	.27 $_{.28}$
CALD (TabICL)	**1.00** $_{.00}$	.00 $_{.00}$	.87 $_{.14}$	.76 $_{.12}$	.42 $_{.15}$	.00 $_{.00}$	**.97** $_{.05}$

Adult Income deserves a remark. It is the only dataset in our benchmark where OOD accuracy is substantially *higher* than ID (+9.5 pp); this is consistent with the underlying covariate shift (ID = Male, OOD = Female) being structurally easier under the 1994 Census labels (the female labor-force composition of that era leaves a sharper income decision boundary). Adult Income is therefore a particularly stringent benign test: a label-blind detector *will* register the substantial covariate redistribution, but the predictive harm is not just absent; it is reversed. A well-calibrated harm detector should remain silent here.

C.4 Main Results with Confidence Intervals

For completeness we report each cell of the main-paper Table 1 as mean $_{\text{std}}$ over 10 seeds, split into the benign (Table 7) and harmful (Table 8) halves. Bolding marks the best mean in each column.

Protocol. For each (dataset, method) cell we run **10 independent seeds** (seeds 42–51); within each seed we run **50 calibration rounds** and **50 deployment rounds** of Algorithm 1. Each round draws a fresh ID anchor / eval pair (calibration) or a fresh OOD eval batch (deployment) of size $|X_{\text{eval}}| = 128$, with the support set $\mathcal{D}^{\text{sup}}$ (size $n = 1024$) and the labeled anchor $(X_{\text{anc}}, y_{\text{anc}})$ ($|X_{\text{anc}}| = 128$) held fixed within the seed. The detection threshold is set per seed at the empirical 5th percentile of that seed's 50 calibration anchor-loss values, and the per-seed TPR is the fraction of the 50 deployment anchor-losses that fall below it. The reported cell is the mean and standard deviation of the 10 per-seed TPRs and, crucially, *not* a pooled estimate across seeds, because different seeds draw different supports and thus have anchor losses on different scales. Baselines (§C.2) follow the same per-seed thresholding protocol on their own scalar score, again calibrated at the seed's own 5th-percentile ID-side cutoff. AUROC, when reported, is computed on the union of the seed's 50 ID and 50 OOD scores and then averaged across the 10 seeds.

LR-stability protocol for CALD. For each (dataset, method) cell we sweep **CALD**'s finetune learning rate over $\eta \in \{10^{-2}, 10^{-3}, 10^{-4}\}$ and *select the largest η that produces a stable signal*, where stability is defined on the per-step direction indicator

$$d_t := \mathbb{1}\left\{\overline{\lambda_t^{\text{OOD}}} < \overline{\lambda_t^{\text{ID}}}\right\}, \qquad t = 1, \ldots, T, \tag{54}$$

where $\overline{\lambda_t^{\text{ID}}}$ and $\overline{\lambda_t^{\text{OOD}}}$ are the round-averaged anchor losses at step t. We accept the candidate η if and only if d_t flips at most twice over $t \in [30, 50]$ (i.e. the trajectory has settled into a single direction in the second half of finetuning), *regardless of which direction it has settled into*. If no LR in the sweep is stable we mark the cell unstable (flagged '!' in the direction column of Table 10); otherwise we report the AUROC and TPR at the largest stable η.

The stability criterion is *not* a post-hoc filter for direction. **CALD** only fires when $\lambda_{\text{OOD}} < \lambda_{\text{ID}}$ (Eq. 6); a stable settled-into-$\lambda_{\text{OOD}} > \lambda_{\text{ID}}$ trajectory simply produces a low TPR, which is what we report. The reason we filter on stability rather than direction is that an unstable trajectory, one whose d_t oscillates between 0 and 1 across the second-half steps, contains no reliable signal. Reading off λ at any single step in such a trajectory is essentially a coin flip on transient gradient noise: stop one step earlier or later and the test verdict reverses. In particular, at high learning rates the ID/OOD trajectories cross each other multiple times before $t = T$, so the value of $\lambda_T = \mathcal{L}_{\text{anc}}(\theta_T)$ depends on whether T happens to land in a momentary alignment of the gradient cross-term. Reporting that single sample without checking for oscillation would be reading a randomly-timed slice of a chaotic curve, effectively "getting lucky" that the finetune horizon ends inside a brief favorable window. The stability criterion ensures that whatever direction we report, the trajectory has stopped flipping by step 30 and the value of λ_T reflects a converged regime rather than an arbitrary point in an oscillating one.

Table 8: **Full TPR@5% FPR results across 7 harmful shifts ($\Delta > 4.8\%$, Hoeffding bound at $N = 1{,}280$, Appendix C.3; TPR = correct alarm rate).** Each cell is mean$_{\text{std}}$ across 10 seeds (seeds 42–51); std is the spread of per-seed TPR (each seed's threshold is its own 5th-percentile ID-score). Means match Table 1 exactly. Bold = best mean in column.

Method Δ (%)	CivilC. +7.4	Camel. +8.9	Waterb. +11.6	CIFAR +12.4	UCI +15.6	PACS +27.9	Covert. +83.5
MSP	**1.00** $_{.00}$	.97 $_{.06}$	**1.00** $_{.00}$	.96 $_{.09}$	.74 $_{.32}$	**1.00** $_{.00}$	**1.00** $_{.00}$
Entropy	**1.00** $_{.00}$	.98 $_{.04}$	**1.00** $_{.00}$	.97 $_{.09}$	.77 $_{.31}$	**1.00** $_{.00}$	**1.00** $_{.00}$
Energy	**1.00** $_{.00}$	**1.00** $_{.00}$	**1.00** $_{.00}$	.97 $_{.09}$	.66 $_{.42}$	**1.00** $_{.00}$	.50 $_{.50}$
Mahal.	.32 $_{.19}$	.43 $_{.16}$	.18 $_{.13}$	.55 $_{.20}$	**1.00** $_{.00}$	**1.00** $_{.00}$	**1.00** $_{.00}$
RMD	.51 $_{.30}$	.40 $_{.38}$	.52 $_{.27}$	.67 $_{.25}$	.47 $_{.36}$	**1.00** $_{.00}$	**1.00** $_{.00}$
KNN	**1.00** $_{.00}$	.95 $_{.08}$	.97 $_{.05}$	.10 $_{.10}$	.99 $_{.03}$	**1.00** $_{.00}$	**1.00** $_{.00}$
ViM	.23 $_{.26}$	.54 $_{.36}$	.18 $_{.22}$	.15 $_{.11}$	.93 $_{.21}$	.70 $_{.36}$	**1.00** $_{.00}$
BBSD-S	.61 $_{.40}$	.94 $_{.15}$	.88 $_{.26}$	.89 $_{.10}$	.81 $_{.34}$	**1.00** $_{.00}$	**1.00** $_{.00}$
BBSD-E	.58 $_{.25}$	**1.00** $_{.00}$	.99 $_{.03}$	.72 $_{.24}$	**1.00** $_{.00}$	**1.00** $_{.00}$	**1.00** $_{.00}$
D-MMD	.91 $_{.21}$	**1.00** $_{.00}$	**1.00** $_{.00}$	.68 $_{.30}$	**1.00** $_{.00}$	**1.00** $_{.00}$	**1.00** $_{.00}$
H-Div	.70 $_{.27}$	**1.00** $_{.00}$	**1.00** $_{.00}$	.50 $_{.17}$	**1.00** $_{.00}$	**1.00** $_{.00}$	**1.00** $_{.00}$
GradNorm	**1.00** $_{.00}$	.92 $_{.14}$	.98 $_{.06}$	.91 $_{.15}$	.54 $_{.42}$	.98 $_{.06}$	.70 $_{.46}$
GradOrth	.66 $_{.24}$	.54 $_{.31}$	**1.00** $_{.00}$	.95 $_{.08}$	.23 $_{.28}$	.97 $_{.06}$	**1.00** $_{.00}$
GAIA-z	—	—	—	—	.41 $_{.42}$	—	.59 $_{.48}$
WeiPer	.34 $_{.36}$	.45 $_{.42}$	.26 $_{.27}$	.21 $_{.18}$	.89 $_{.30}$	.62 $_{.42}$	.73 $_{.42}$
CALD (TabPFN)	1.00 $_{.01}$	.19 $_{.13}$	.46 $_{.46}$	**1.00** $_{.01}$	.71 $_{.35}$	**1.00** $_{.00}$	**1.00** $_{.00}$
CALD (TabICL)	.99 $_{.02}$	**1.00** $_{.00}$	.99 $_{.02}$	.99 $_{.02}$	.71 $_{.32}$	**1.00** $_{.00}$	**1.00** $_{.00}$

In practice, $\eta = 10^{-3}$ is the modal choice across the 14 datasets: $\eta = 10^{-2}$ is rejected on the more sensitive datasets (notably CelebA and CivilComments under TabPFN, where the joint loss oscillates across iterations 30–50) and $\eta = 10^{-4}$ is the fallback on four cells (Adult Income under both models, CelebA-TabICL, and Folktables-TabPFN), where neither 10^{-2} nor 10^{-3} stabilizes; Appendix D.5 reports the census. The dataset-by-dataset selected LRs and the stability margins are produced by `detection/viz/cache_results.py` and consumed by the main-paper table builder.

C.5 Comparison to Drift-Resilient TabPFN

A natural alternative to *detecting* distribution shift is to *adapt* the model so that shift no longer harms predictions. Drift-Resilient TabPFN [17] pursues this route: it retrains the PFN prior on synthetic data sampled from *evolving* structural causal models, producing a tabular foundation model whose predictions are intended to be robust under covariate shift. If this adaptation were sufficient on the benchmark of §C.3, the failure mode that motivates **CALD** would not arise, and a detector would be redundant. We therefore evaluate whether Drift-Resilient TabPFN closes the ID/OOD accuracy gap on our benchmark; closing the gap is a *necessary* condition for the claim that adaptation removes the need for a detector.

Table 9 repeats the deterioration survey of Table 6 using Drift-Resilient TabPFN (single-estimator configuration, matching standard TabPFN) on the same ID/OOD splits, support sizes, and round structure. The right-most column reports $\delta\mathrm{OOD} = \mathrm{acc}^{\mathrm{drift}}_{\mathrm{OOD}} - \mathrm{acc}^{\mathrm{std}}_{\mathrm{OOD}}$, which isolates the effect of drift adaptation on OOD accuracy. Across all 15 datasets of Table 9 (including the Camelyon17 ResNet18 variant) the mean $\delta\mathrm{OOD}$ is -1.4 pp; on the eight datasets classified as harmful in Table 6 the mean is -2.2 pp. Drift adaptation strictly improves OOD accuracy on only two harmful datasets (Covertype, $+3.8$ pp; Camelyon17 ResNet18, $+3.0$ pp), and *degrades* OOD accuracy by more than the Hoeffding noise floor ($\Delta^* \approx 4.8$ pp, Eq. 52) on three (PACS -8.0 pp, CIFAR-10 -7.8 pp, Waterbirds -4.8 pp). The dataset with the largest standard-TabPFN deterioration after Covertype, PACS at $\Delta = +27.9$ pp, is made measurably *worse* by adaptation rather than better, and ID accuracy is also slightly degraded on a majority of datasets (mean -1.0 pp). The harm classification of Table 6 is essentially unchanged: every dataset that was harmful under standard TabPFN remains harmful under Drift-Resilient TabPFN by the same Hoeffding criterion. The headline finding is thus that drift adaptation does *not* eliminate the deterioration that **CALD** is designed to detect; on the present benchmark it slightly aggravates it on average.

The result is consistent with, and indeed predicted by, the design of Drift-Resilient TabPFN. Its prior is constructed by sampling structural causal models and then perturbing their parameters along a temporal axis to simulate *gradual evolution* of the data-generating process. The resulting model is therefore tuned to a specific, structured family of shifts (smoothly drifting SCMs, of the kind found in the Wild-Time benchmarks for which the method was developed) and not to the heterogeneous covariate shifts in our benchmark, which include image-domain transfer (Camelyon17, Waterbirds, PACS), feature-space stratification (Covertype's elevation cut, UCI Heart's hospital split), and large textual-distribution gaps (CivilComments). On several datasets (PACS, CIFAR-10, Waterbirds) the perturbed prior actively misallocates predictive mass, assuming the wrong drift mechanism, which manifests directly as degraded OOD accuracy. More fundamentally, no fixed-prior model can be uniformly robust to arbitrary covariate shift: by a no-free-lunch argument, any inductive bias added to handle one family of shifts must be paid for on another. This is the structural reason adaptation cannot *replace* detection: a deployed practitioner has no way to verify, a priori, that the shift at hand falls within the family the adapted prior was trained for. **CALD** and Drift-Resilient TabPFN are therefore complementary in their roles, with the detector indicating *whether* a shift has occurred while the adapter (when applicable) indicates *how* to respond. The empirical evidence in Table 9 suggests that on a heterogeneous covariate-shift benchmark, detection is the more dependable of the two.

C.6 Per-Dataset Detection Performance of $\hat{\lambda}$

Table 10 reports AUROC and TPR@5% FPR for the post-finetune anchor-loss statistic $\hat{\lambda}$ alone (without baselines) across all datasets and both models, complementing the baseline comparison in the main paper. The protocol matches Table 1: seeds 42–51, $K = 50$ calibration rounds, per-seed 5%-quantile thresholds, learning rate selected per cell by the tail-stability rule.

Table 9: **TabPFN vs. Drift-Resilient TabPFN accuracy under distribution shift.** Standard TabPFN (left) and Drift-Resilient TabPFN [17] (right) on the same ID/OOD splits. $\Delta = \mathrm{acc}_{\mathrm{ID}} - \mathrm{acc}_{\mathrm{OOD}}$ (positive = harmful drop). $\delta\mathrm{OOD} = \mathrm{acc}^{\mathrm{drift}}_{\mathrm{OOD}} - \mathrm{acc}^{\mathrm{std}}_{\mathrm{OOD}}$ (positive = drift model improves OOD accuracy). Harm threshold $\Delta^* \approx 4.8\%$ (Hoeffding, $\delta = 0.05$, $N = 1{,}280$); benign datasets marked †.

	Standard TabPFN			**Drift-Resilient TabPFN**			
Dataset	ID	OOD	Δ (%)	ID	OOD	Δ (%)	δOOD (%)
Benign ($\Delta \leq \Delta^*$)							
CelebA†	0.952	0.908	+4.5	0.946	0.873	+7.3	−3.4
Bank Marketing†	0.902	0.859	+4.4	0.901	0.859	+4.1	+0.1
Folktables†	0.799	0.768	+3.1	0.800	0.766	+3.4	−0.2
Credit Default†	0.847	0.819	+2.8	0.842	0.815	+2.7	−0.4
Amazon†	0.645	0.630	+1.5	0.633	0.644	−1.1	+1.3
Diabetes Readm.†	0.886	0.909	−2.3	0.886	0.909	−2.3	+0.0
Adult Income†	0.820	0.914	−9.5	0.800	0.912	−11.2	−0.2
Harmful ($\Delta > \Delta^*$)							
Covertype	0.904	0.069	+83.5	0.777	0.106	+67.1	**+3.8**
PACS	0.998	0.720	+27.9	1.000	0.640	+36.0	−8.0
UCI Heart	0.824	0.668	+15.5	0.818	0.656	+16.2	−1.2
CIFAR-10→10.1	0.795	0.670	+12.4	0.741	0.592	+14.9	−7.8
Waterbirds	0.988	0.871	+11.6	0.987	0.823	+16.4	−4.8
Camelyon17 (DINOv2)	0.953	0.864	+8.9	0.941	0.847	+9.4	−1.7
Camelyon17 (ResNet18)	0.923	0.840	+8.3	0.921	0.870	+5.2	**+3.0**
CivilComments	0.918	0.844	+7.4	0.915	0.836	+7.9	−0.8

D Ablations

This appendix collects the hyperparameter sweeps, design-choice ablations, and the further ablations deferred from §5: the disagreement weight, the finetuning budget and its deployment cost, the frozen deployment configuration, anchor robustness, feature collapse, the theorem's condition and premise, the two-signal detector, and detectability across shift types.

D.1 Hyperparameter Sweeps

We sweep four hyperparameters of **CALD** on three representative datasets covering catastrophic, moderate, and structured-image shift: Covertype, Credit Default, and CIFAR-10. The point of the sweeps is twofold: to check that the detector is not living on a knife-edge of its compute budget, and to verify that the operating regime under which the theory applies (small-step finetuning, modest support window) is the regime under which the practical detector actually performs well.

Number of finetune steps T. The natural worry is that we have tuned T to a sweet spot beyond which the signal collapses; if so, the theorem's first-order analysis would be tracking a transient that the deployed detector exits at $T = T_{\mathrm{deploy}}$. AUROC is in fact largely insensitive to T on strong shifts (Covertype, Credit Default: 1.000 at every $T \in \{1, 3, 5, 10, 20, 50\}$). On the subtler CIFAR-10 shift more steps help, with AUROC rising from 0.522 at $T = 1$ to 0.992 at $T = 50$. The default $T = 50$ sits well within the small-step regime that Theorem 4.1's remainder bound requires, and the ID/OOD separation does not depend on a finely tuned step count.

Learning rate η. Theorem 4.1 gives a first-order guarantee that becomes vacuous if the second-order remainder $O(T^2\eta^2)$ catches up with the leading $T\eta$ term; in plain words, the analysis only applies when η is small. Sweeping over a three-decade range $\eta \in \{10^{-4}, 10^{-3}, 5\times10^{-3}, 10^{-2}, 5\times10^{-2}, 10^{-1}\}$, Credit Default and Covertype keep AUROC at 1.000 throughout, and CIFAR-10 degrades only at very low η ($< 5\times10^{-3}$) where the steps are too small to move the decoder into the disagreement regime at all. The detector therefore inhabits the small-step regime under which the theory applies, and the Algorithm 1 LR-stability heuristic can pick within that regime with substantial slack.

Table 10: **Per-dataset detection performance of the post-finetune anchor loss $\hat{\lambda}$.** AUROC and TPR@5% FPR for distinguishing ID from OOD eval batches, under the same protocol as Table 1 (seeds 42–51, $K = 50$ calibration rounds, per-seed 5%-quantile thresholds, learning rate selected per cell by the tail-stability rule). Direction ✓ : OOD < ID as expected; **!**: reversed for the majority of seeds. All four reversed cells are benign shifts and coincide with the cells where the learning-rate stability rule falls back to $\eta = 10^{-4}$. The per-cell selection lies outside the conformal guarantee of Proposition 4.2, which covers a configuration fixed in advance; the certified frozen-$\eta = 10^{-3}$ configuration is reported alongside these results in Appendix D.5.

Dataset	Model	AUROC	TPR@5%	Dir.
Image				
Camelyon17 (DINOv2)	TabPFN	0.675	0.192	✓
	TabICL	**1.000**	1.000	✓
Camelyon17 (ResNet18)	TabPFN	0.977	0.914	✓
	TabICL	0.990	0.932	✓
CIFAR-10→10.1	TabPFN	**1.000**	0.998	✓
	TabICL	0.996	0.990	✓
PACS	TabPFN	**1.000**	1.000	✓
	TabICL	**1.000**	1.000	✓
Tabular				
UCI Heart	TabPFN	0.908	0.713	✓
	TabICL	0.930	0.714	✓
Covertype	TabPFN	**1.000**	1.000	✓
	TabICL	**1.000**	1.000	✓
Credit Default	TabPFN	**1.000**	1.000	✓
	TabICL	0.698	0.242	✓
Folktables	TabPFN	0.624	0.020	**!**
	TabICL	0.899	0.578	✓
Bank Marketing	TabPFN	0.997	0.978	✓
	TabICL	**1.000**	1.000	✓
Diabetes Readm.	TabPFN	0.976	0.918	✓
	TabICL	**1.000**	1.000	✓
Adult Income	TabPFN	**1.000**	0.000	**!**
	TabICL	**1.000**	0.000	**!**
Subpop.				
Waterbirds	TabPFN	0.921	0.462	✓
	TabICL	0.993	0.986	✓
CelebA	TabPFN	0.919	0.730	✓
	TabICL	0.845	0.030	**!**
CivilComments	TabPFN	0.999	0.998	✓
	TabICL	0.997	0.987	✓
Amazon	TabPFN	0.550	0.082	✓
	TabICL	0.600	0.127	✓

Support size n. The PFN's in-context support set is the conditioning information the entire detector reads against, and using more of it costs both quadratic attention compute and a longer prompt. We sweep n to see how small the in-context window can be before the detection signal degrades. Performance falls gracefully at $n = 100$ (AUROC 0.762–0.914) but recovers fully by $n = 250$, so a modest in-context window suffices to anchor the detector. This rules out the alternative that **CALD**'s separation is a side-effect of the PFN being asked to solve a particularly large in-context task.

Anchor size $|X_{\text{anc}}|$. The anchor batch is the calibration substrate against which the post-finetune λ is scored, so its size enters the variance of the calibration record and the conformal threshold from Proposition 4.2. **CALD** is robust to anchor size across $|X_{\text{anc}}| \in \{16, 32, 64, 128, 256\}$, with AUROC > 0.99 in all but one cell. The anchor set therefore does not need to be large to serve as a reliable ID reference, and the calibration cost of the method is bounded by the eval-batch geometry rather than by the held-out anchor.

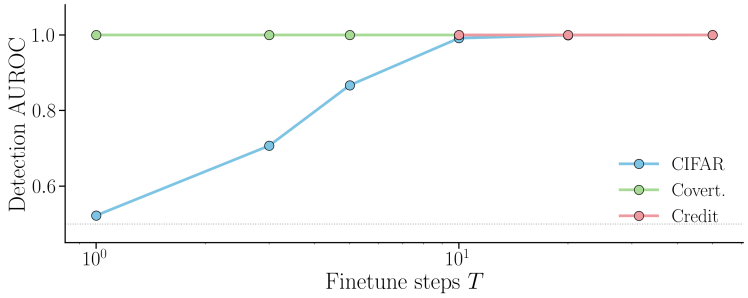


Figure 14: **Ablation: number of gradient steps T.** Detection AUROC vs. T on three representative datasets. Covertype and Credit Default saturate immediately; CIFAR-10 improves with more steps.

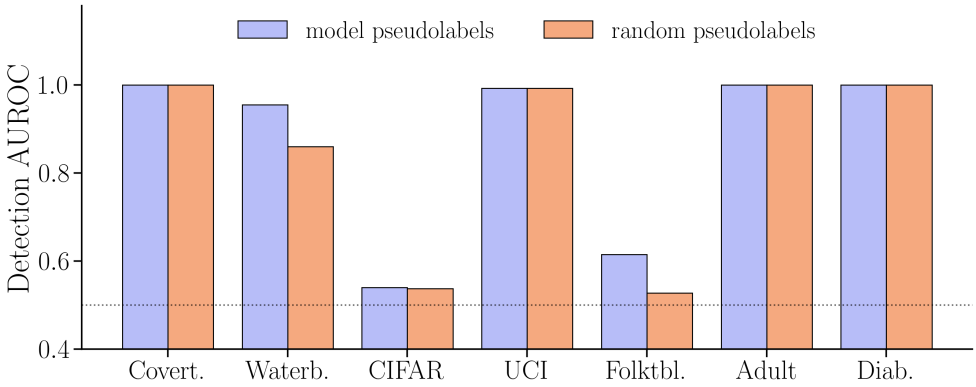


Figure 15: **Detection AUROC with model-pseudolabels vs. random pseudolabels.** Harm-adjusted; dotted line marks chance (0.5). Random pseudolabels leave detection unchanged on strong-shift datasets but cost ∼10 points of AUROC on moderate-shift datasets (Waterbirds, Folktables) where the PFN is less confident on its own pseudolabels.

D.2 Design-Choice Ablations

Pseudolabel choice. **CALD**'s disagree term pushes the eval predictions away from the model's own pseudolabels $\hat{y} = \arg\max\, q_{\theta_0}(y \mid X_{\text{eval}}, \mathcal{D}^{\text{sup}})$. Why use the model's pseudolabels rather than any other arbitrary target? If *any* target produced the same ID/OOD separation, the signal would be coming from the disagreement objective in general rather than from the specific direction in label space that a confident model points to. Replacing $\hat{y}$ with random labels drawn uniformly over the C classes (Figure 15) leaves AUROC unchanged on the datasets with large covariate shift (Covertype, UCI Heart, Adult, Diabetes Readmission), where any non-degenerate target is far enough from the posterior to drive the disagree gradient. On the moderate-shift datasets where the pretrained PFN is less confident on its own pseudolabels (Waterbirds and Folktables), random pseudolabels cost 9–10 AUROC points. The theoretical residual $r_e = p - u_{\hat{y}}$ depends on $\hat{y}$ matching the modal class prediction; the degradation is therefore concentrated in exactly the regime where the modal-class structure is the key component of the signal, consistent with the gradient form derived in §4.2.

Pseudolabel scope. A natural alternative to freezing $\hat{y}$ at θ_0 is to recompute pseudolabels each step, so the disagree term always points away from the *current* model's predictions. The theoretical analysis in §4.2 assumes $\hat{y}$ is fixed; if recomputing it preserved or improved the signal, the assumption would be a pure analytical convenience. We ablate per-batch vs. per-sample pseudolabels and frozen-at-θ_0 vs. recomputed-at-step. Freezing at θ_0 is necessary: recomputing destabilizes the target and the anchor-loss signal becomes noisy. The fixed-$\hat{y}$ choice is therefore both an analytical and an operational requirement, not a tractability shortcut.

Last-layer-only finetuning. Theorem 4.1 proves a clean rank-one form for the per-sample gradients g_a and g_e only when updates are restricted to the final linear layer: with the full two-layer GELU decoder unfrozen, the gradient mixes contributions across layers and the feature-orthogonality

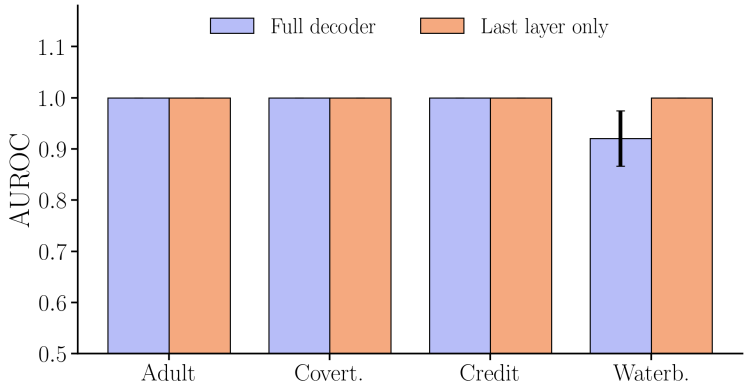


Figure 16: **Last-layer-only vs. full-decoder finetuning.** Detection AUROC matches to within 0.01 on every dataset, confirming that Theorem 4.1's last-layer assumption is benign in practice.

Table 11: **Ablation of the disagreement weight β.** Mean detection AUROC (TabPFN, fixed $\eta = 10^{-3}$, 30 rounds, 3 seeds). Within-configuration results under a reduced protocol; absolute values are not comparable to Table 10, which selects the learning rate per cell. "–": not run.

Dataset	β=0	0.25	0.5	1	2	4	8	16
Adult Income	0.500	1.000	1.000	1.000	1.000	1.000	0.663	0.990
Bank Marketing	0.500	0.913	0.937	0.957	0.975	0.983	–	–
Covertype	0.500	1.000	1.000	1.000	1.000	1.000	1.000	1.000
UCI Heart	0.500	0.808	0.814	0.745	0.856	0.963	0.994	0.976
Waterbirds	0.500	0.980	0.999	1.000	1.000	1.000	–	–

argument in Corollary A.4 no longer applies. The deployed detector finetunes the whole decoder, so the theory could in principle be analyzing a regime that has nothing to do with how **CALD** is actually run. Re-running with only the last linear layer trainable matches full-decoder AUROC to within 0.01 on every dataset (Figure 16). The two regimes are therefore empirically indistinguishable, which means the analytical restriction is not a tractability tax: the theory describes the regime under which the detector operates, and a practitioner who wants the cheaper last-layer-only protocol pays nothing in detection power.

D.3 Disagreement Weight β

The joint objective weights the disagreement term by a coefficient β that the main experiments fix at 1, and a natural concern is that this choice is key: if detection degraded away from $\beta = 1$, the reported results would hinge on an untuned constant, and if detection survived $\beta = 0$, the disagreement term would not be the mechanism at all. We are therefore interested in ablating β across several orders of magnitude, including the degenerate $\beta = 0$. At first order the question is already constrained: the separation gap obeys $\Delta\lambda \approx \eta\beta\,(\mathbb{E}_{\mathrm{ID}}\langle g_a, g_e\rangle - \mathbb{E}_{\mathrm{OOD}}\langle g_a, g_e\rangle)$, so η and β enter only through their product and the learning-rate sweep of Appendix D.1 already spans this axis at $\beta = 1$; the sweep below tests whether the first-order picture survives beyond it.

We set up the experiment as follows: TabPFN with fixed $\eta = 10^{-3}$, 30 calibration rounds, and 3 seeds, sweeping $\beta \in \{0, 0.25, 0.5, 1, 2, 4\}$ on five datasets spanning far-OOD, moderate tabular, and subpopulation shift, extended to $\beta \in \{8, 16\}$ on three of them. Figure 17 plots mean detection AUROC with per-seed spread; Table 11 reports the values.

These results show two things. First, at $\beta = 0$ the AUROC collapses to exactly 0.500 on every run of every dataset: with the disagreement term removed, the post-finetune anchor loss carries no ID/OOD information whatsoever, confirming that the eval-batch disagreement gradient is the mechanism rather than an auxiliary regularizer. Second, detection is flat across $\beta \in [0.25, 4]$, a sixteen-fold range around the default, so $\beta = 1$ sits on a wide plateau rather than a tuned peak. At very large β the product $\eta\beta$ exits the small-step regime of Assumption A.7 and the first-order picture no longer applies: Adult Income becomes unstable at $\beta = 8$ before recovering at 16, while UCI Heart, the

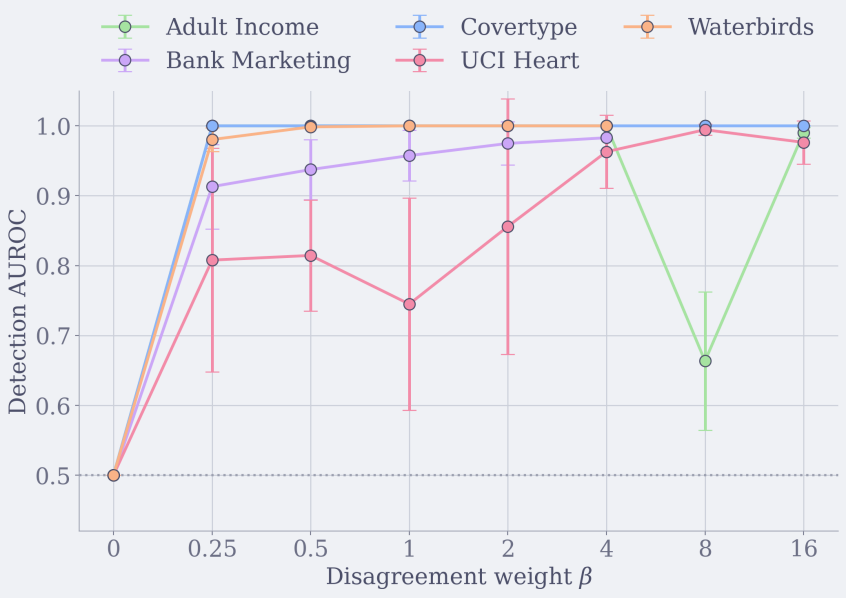


Figure 17: **Ablation of the disagreement weight β.** Mean detection AUROC over 3 seeds (error bars: ± 1 std; TabPFN, fixed $\eta = 10^{-3}$, 30 rounds; dotted line: chance). At $\beta = 0$ every dataset collapses exactly to chance; detection is flat across $\beta \in [0.25, 4]$ and the first-order picture degrades only at very large $\eta\beta$.

dataset with the weakest first-order signal (cf. Table 18), actually benefits from the stronger push. $\beta = 1$ is principled at first order, where only the product $\eta\beta$ matters, and empirically non-critical over two orders of magnitude.

D.4 Deployment Cost and the Finetuning Budget T

Unlike post-hoc scores, **CALD** must run T gradient steps on every monitored batch, so its deployment cost is a practical concern: a detector that is accurate but orders of magnitude more expensive than the alternatives needs to state exactly what the budget buys. We seek to determine (i) the wall-clock cost per monitored batch and (ii) how detection accuracy depends on the finetuning budget T, i.e. how much of the paper's $T = 50$ budget is actually needed.

We set up the experiment as follows. We time one finetuning step per monitored batch of 128 points on an NVIDIA H100: 0.168 s for TabPFN and 0.056 s for TabICL, so the full $T = 50$ budget costs 8.4 s and 2.8 s per batch respectively, against 0.0003–0.67 s *total* for the 15 post-hoc baselines. We then truncate the detector at $T \in \{5, 10, 20, 50\}$ on all 15 datasets and both models (10 seeds, per-cell selected learning rate as in Table 10) and record T_{95}, the smallest T whose mean AUROC reaches 95% of the $T = 50$ value. Table 12 reports the full per-dataset cost–accuracy profile.

The timing results show a trade-off. In raw wall-clock terms **CALD** is one to four orders of magnitude more expensive than any post-hoc baseline, and no amount of truncation closes that gap. But the budget needed to reach the reported accuracy is a small fraction of $T = 50$: the median T_{95} is 1 step for TabPFN and 2 for TabICL, and $T \leq 10$ (about 1.7 s and 0.6 s per monitored batch) recovers 95% of the final AUROC on 14 of 15 datasets for TabPFN and 12 of 15 for TabICL. The exceptions are the slow-converging cells (Folktables/TabPFN at $T_{95} = 38$; CelebA, UCI Heart, and Diabetes Readmission under TabICL at 36, 21, and 14), which are also among the weaker cells at full budget. One cell is non-monotone: Camelyon17 (DINOv2) under TabPFN peaks early (0.992 at T=5) and degrades with further finetuning, which is the budget-resolved view of its weak full-budget value in Table 10. $T = 10$ is the practical operating point for deployment, with $T = 50$ retained in the main results as the conservative budget at which the theory's small-step premises were validated.

Table 12: **Detection AUROC as a function of the finetuning budget T.** Mean over 10 seeds at the per-cell selected learning rate (same protocol as Table 10). T_{95} is the smallest T reaching 95% of the T=50 AUROC; s@T_{95} is the corresponding wall-clock per monitored batch (H100, batch 128). Aggregates: TabPFN reaches 0.845 ± 0.159 at T=10 vs. 0.904 ± 0.154 at T=50 with median $T_{95} = 1$; TabICL reaches 0.851 ± 0.161 vs. 0.930 ± 0.124 with median $T_{95} = 2$.

Dataset	Model	@T=5	@10	@20	@50	T_{95}	s@T_{95}
Adult Income	TabPFN	1.000	1.000	1.000	1.000	1	0.17
	TabICL	0.978	1.000	1.000	1.000	2	0.11
Amazon	TabPFN	0.538	0.544	0.543	0.550	1	0.17
	TabICL	0.592	0.612	0.624	0.600	1	0.06
Bank Marketing	TabPFN	0.890	0.981	0.985	0.997	2	0.34
	TabICL	1.000	0.957	1.000	1.000	1	0.06
Camelyon17 (DINOv2)	TabPFN	0.992	0.639	0.678	0.675	1	0.17
	TabICL	0.578	0.999	1.000	1.000	2	0.11
Camelyon17 (ResNet18)	TabPFN	0.994	0.811	0.882	0.977	3	0.50
	TabICL	0.608	0.977	0.988	0.990	2	0.11
CelebA	TabPFN	0.819	0.734	0.778	0.919	1	0.17
	TabICL	0.711	0.718	0.698	0.845	36	2.02
CIFAR-10→10.1	TabPFN	0.912	0.984	0.999	1.000	7	1.17
	TabICL	0.758	0.964	0.996	0.996	10	0.56
CivilComments	TabPFN	0.998	0.821	0.913	0.999	1	0.17
	TabICL	0.994	0.832	0.976	0.997	1	0.06
Covertype	TabPFN	1.000	1.000	1.000	1.000	1	0.17
	TabICL	1.000	1.000	1.000	1.000	1	0.06
Credit Default	TabPFN	1.000	0.960	1.000	1.000	1	0.17
	TabICL	0.771	0.611	0.615	0.698	1	0.06
Diabetes Readmission	TabPFN	0.993	0.956	0.962	0.976	1	0.17
	TabICL	0.835	0.913	0.997	1.000	14	0.78
Folktables	TabPFN	0.573	0.574	0.577	0.624	38	6.37
	TabICL	0.771	0.588	0.752	0.899	9	0.50
PACS	TabPFN	0.943	0.959	1.000	1.000	10	1.68
	TabICL	0.713	1.000	1.000	1.000	8	0.45
UCI Heart	TabPFN	0.859	0.792	0.727	0.915	7	1.17
	TabICL	0.715	0.693	0.777	0.930	21	1.18
Waterbirds	TabPFN	0.726	0.919	0.920	0.921	1	0.17
	TabICL	0.844	0.896	0.918	0.993	3	0.17

D.5 Frozen Deployment Configuration and Stability of the Learning-Rate Selection

Proposition 4.2 bounds the false-alarm rate only when the anchor, support, and finetuning hyperparameters are fixed *before* deployment, yet Table 10 selects the learning rate per dataset–model cell with the tail-stability rule, i.e. with a procedure that looks at the trajectories being certified. The guarantee therefore covers a fixed configuration, not the selection protocol, and two questions follow: what does it cost to actually deploy the fixed configuration the guarantee speaks about, and how fragile is the selection rule that produces the headline numbers? We measure three things: the accuracy cost of freezing $\eta = 10^{-3}$ everywhere, a census of the selection rule's fallback behavior across all cells and tail windows, and the empirical false-positive rate of the frozen configuration under stress.

We set up the experiment as follows. (a) We recompute all 30 method×dataset cells with the single frozen configuration $\eta = 10^{-3}$ (no selection of any kind) and compare against the selected-LR pipeline of Table 10. (b) We re-run the tail-stability selection over all 30 cells and three tail windows (steps 25–45, 30–50, 35–50). (c) We estimate the false-positive rate of the frozen configuration under the no-shift null over 240 calibration/deployment rounds on six datasets, in four stress regimes: deterministic replay, nondeterministic GPU kernels, re-drawing the anchor every round, and time-correlated eval streams.

The certified configuration is deployable at essentially no average cost: freezing $\eta = 10^{-3}$ changes AUROC by -0.005 on average over the 30 cells. The average hides real per-cell movement in both directions, which we disclose rather than average away: CIFAR-10 is the worst case (-0.103

Table 13: **Frozen configuration ($\eta = 10^{-3}$, no selection) vs. the per-cell selected learning rate of Table 10.** Detection AUROC; Δ = frozen $-$ selected. Mean Δ over all 30 cells is -0.005 (AUROC) and $+0.008$ (TPR@5%, not shown). The frozen column is the configuration covered by the conformal guarantee.

	TabPFN			TabICL		
Dataset	sel.	frozen	Δ	sel.	frozen	Δ
Adult Income	1.000	1.000	+0.000	1.000	1.000	+0.000
Amazon	0.550	0.606	+0.056	0.600	0.624	+0.024
Bank Marketing	0.997	0.935	−0.062	1.000	1.000	+0.000
Camelyon17 (DINOv2)	0.675	0.999	+0.324	1.000	1.000	+0.000
Camelyon17 (ResNet18)	0.977	0.998	+0.021	0.990	0.998	+0.008
CelebA	0.919	0.919	+0.000	0.845	0.806	−0.039
CIFAR-10→10.1	1.000	0.896	−0.103	0.996	0.769	−0.227
CivilComments	0.999	0.990	−0.009	0.997	0.993	−0.005
Covertype	1.000	1.000	+0.000	1.000	1.000	+0.000
Credit Default	1.000	1.000	+0.000	0.698	0.766	+0.068
Diabetes Readmission	0.976	0.984	+0.008	1.000	0.895	−0.105
Folktables	0.624	0.572	−0.052	0.899	0.859	−0.040
PACS	1.000	0.940	−0.060	1.000	0.992	−0.008
UCI Heart	0.908	0.885	−0.024	0.930	0.903	−0.027
Waterbirds	0.921	1.000	+0.079	0.993	1.000	+0.007

Table 14: **Empirical false-positive rate of the frozen configuration under the no-shift null.** Six datasets $\times$ 240 null rounds per regime; conformal bound $\alpha + 1/(K+1) = 0.070$. Regime means sit at or below the bound; four individual dataset–regime cells marginally exceed it (max 0.075), consistent with finite-sample fluctuation at $n = 240$.

Stress regime	Mean FPR	Per-dataset range
Deterministic replay	0.065	0.057–0.072
Nondeterministic GPU kernels	0.065	0.057–0.072
Anchor re-drawn every round	0.064	0.052–0.069
Time-correlated eval stream	0.068	0.063–0.075

TabPFN, -0.227 TabICL; it needs $\eta = 10^{-2}$), while Camelyon17 (DINOv2)/TabPFN *improves* by $+0.324$, meaning the selection rule itself mis-serves that cell. The census in (b) shows the selection rule is otherwise stable: 4 of 30 cells trigger the fallback to $\eta = 10^{-4}$ (Adult Income on both models at AUROC 1.000 despite the fallback, CelebA/TabICL at 0.845, and Folktables/TabPFN at 0.624), all four on benign shifts, while every severely shifted dataset selects its learning rate stably; across the three tail windows the selected rate changes in 1 of 60 comparisons (Folktables/TabPFN under the 25–45 window, ΔAUROC -0.052). Finally, Table 14 shows the frozen configuration's false-alarm rate stays controlled under all four stress regimes. The fixed $\eta = 10^{-3}$ configuration is the certified deployment mode of **CALD** — the mode Proposition 4.2 actually speaks about — and per-cell selection is a reporting-time diagnostic that requires labeled OOD trajectories and should be read as such in Table 10.

D.6 Anchor Construction Robustness

The anchor set is the reference that every statistic in the method is scored against, so the entire pipeline could in principle hinge on drawing a lucky, perfectly clean anchor — a serious concern in deployment, where held-out labels may be scarce, noisy, or quietly contaminated by the very shift being monitored. Appendix D.1 already varies the anchor's *size*; here we seek to determine the sensitivity to its *quality*, stressing the anchor along the remaining axes: label noise, contamination by OOD points, and re-drawing the anchor itself.

We set up the experiment as follows: TabPFN, 3 seeds per cell. For label noise we flip a fraction $\rho \in \{0, 0.05, 0.1, 0.2\}$ of anchor labels uniformly at random; for contamination we replace a fraction $\rho \in \{0, 0.05, 0.1, 0.25\}$ of anchor points with OOD samples; for re-drawing we resample the anchor from the held-out ID pool at every calibration round and measure the false-positive rate under the

Table 15: **Anchor-quality stress.** Detection AUROC (TabPFN, 3 seeds) under anchor label noise and anchor OOD contamination. Within-configuration results under a reduced protocol; absolute values are not comparable to Table 10. "–": stress level not completed. All movements are within the per-seed standard deviation of the corresponding cell (up to ± 0.18 on PACS and ± 0.16 on CIFAR-10).

	Label noise ρ				OOD contamination ρ			
Dataset	0	.05	.10	.20	0	.05	.10	.25
Adult Income	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Bank Marketing	0.934	0.935	0.933	0.946	0.932	0.932	0.929	0.925
CIFAR-10→10.1	0.920	0.757	0.835	0.955	0.963	–	–	–
CivilComments	0.991	0.989	0.988	1.000	0.991	0.991	0.988	–
Covertype	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Credit Default	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Diabetes Readmission	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Folktables	0.516	0.518	0.526	0.526	0.515	0.511	0.517	0.529
PACS	0.906	0.790	0.903	0.852	0.932	–	–	–
UCI Heart	0.751	0.704	0.691	0.839	0.751	0.723	0.776	0.735
Waterbirds	1.000	0.997	0.989	–	1.000	1.000	1.000	0.999

no-shift null (the "anchor re-drawn every round" regime of Table 14). Table 15 reports detection AUROC on the datasets with complete stress grids.

Anchor quality is a low-sensitivity component of the method. Under up to 20% label noise, no dataset shows a movement exceeding its own per-seed standard deviation: the largest swings (PACS, about ± 0.1 across noise levels against a per-seed std of up to ± 0.18, and CIFAR-10's dip at $\rho = 0.05$ against a std of ± 0.16) are seed noise, not trends, and neither is monotone in ρ as genuine degradation would be. OOD contamination up to 25% moves no dataset by more than 0.03. Re-drawing the anchor every round keeps the false-positive rate at 0.052–0.069, within the 0.070 bound. This robustness is also what the theory predicts: in Theorem 4.1, γ is a population-level constant in which the anchor sample never appears (conditioning the proof on a fixed anchor merely replaces γ by $\gamma(X_{\mathrm{anc}})$), and in Proposition 4.2 the exchangeable unit of the rank argument is the calibration *round*, not the anchor point, so anchor size and composition never enter either guarantee. Empirically the anchor contributes little even to the *variance* of the statistic: the spread of the per-round calibration scores is essentially flat from anchor size 16 to 256 (median std ratio 0.97), because round-to-round eval-batch resampling, not anchor averaging, dominates the variance of λ. The anchor needs to be a modest, roughly clean sample of ID data and nothing more.

D.7 Feature Collapse in Deployed Multi-Layer PFNs

The feature-orthogonality assumption behind Theorem 4.1 is motivated by Nagler's tilting argument (Appendix A.8), which is proven for single-layer attention, while the deployed TabPFN and TabICL are multi-layer transformers — so the assumption could in principle describe a model class that the experiments never touch. The multi-layer extension is a hard open problem, but the argument's central *prediction* is directly measurable: as an evaluation batch becomes more OOD, its features should collapse toward a common representation, becoming more orthogonal to the ID anchor features. We seek to determine whether this predicted collapse actually occurs in the deployed multi-layer models, under both far-OOD and near-OOD conditions.

We set up the experiment as follows. On Covertype (far-OOD) and Waterbirds (near-OOD) we construct evaluation batches containing a fraction $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ of OOD examples and $1 - \alpha$ ID examples. For each α we extract the anchor and evaluation features of the deployed models with no finetuning (TabPFN: the pre-decoder representation; TabICL: the input to the ICL predictor's decoder, i.e. the post-LayerNorm output of the dataset-wise ICL transformer) and measure two statistics: the 95th percentile of the absolute cosine similarity across all anchor–evaluation feature pairs, after centering by the pooled $\alpha = 1$ evaluation mean as a proxy for the collapse direction ϕ_∞ (this is the empirical proxy for ϵ in Assumption A.3, and should *decrease* under tilting), and the evaluation-feature dispersion, the mean Euclidean distance of each evaluation feature to the mean evaluation feature (which should also *decrease* under collapse). We additionally report detection

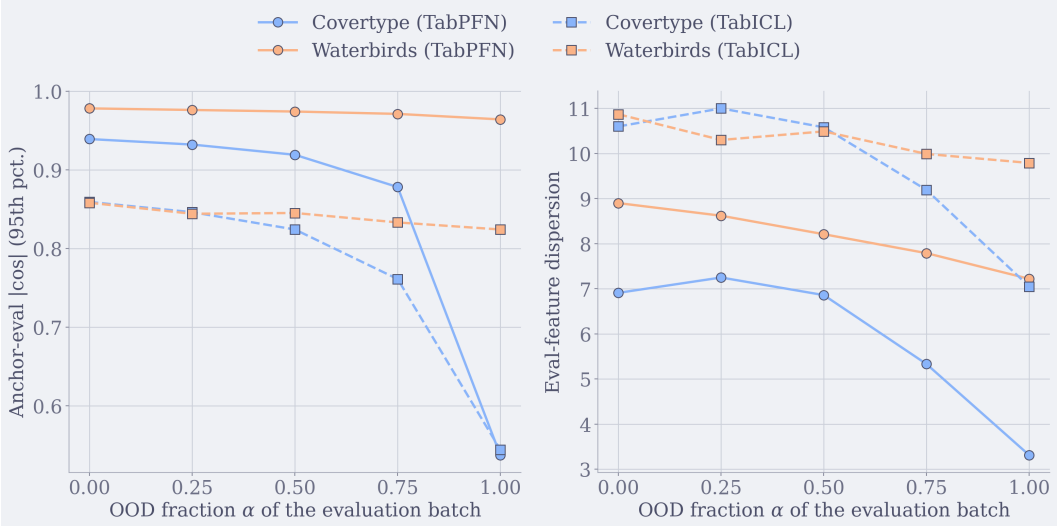


Figure 18: **Feature collapse under controlled OOD mixing.** 95th-percentile absolute anchor–eval cosine (left; the empirical proxy for ϵ) and evaluation-feature dispersion (right) as the OOD fraction α of the evaluation batch grows; solid lines are TabPFN, dashed are TabICL. Both collapse statistics fall sharply on the far-OOD Covertype and mildly on the near-OOD Waterbirds, as the tilting argument predicts.

Table 16: **Feature collapse under controlled OOD mixing.** 95th-percentile absolute anchor–eval cosine (the empirical proxy for ϵ), evaluation-feature dispersion, and detection AUROC vs. the α=0 reference, as the OOD fraction α of the evaluation batch grows. Both statistics decrease with α, strongly on the far-OOD Covertype and mildly on the near-OOD Waterbirds. AUROC here is within the mixing protocol and not comparable to Table 10.

		Covertype (far-OOD), $\alpha =$					Waterbirds (near-OOD), $\alpha =$				
Model	Metric	0	.25	.5	.75	1	0	.25	.5	.75	1
TabPFN	95% $\lvert\cos\rvert$	0.939	0.932	0.919	0.878	0.537	0.978	0.976	0.974	0.971	0.964
	dispersion	6.91	7.25	6.86	5.33	3.31	8.90	8.62	8.21	7.79	7.22
	AUROC	–	0.750	0.883	1.000	1.000	–	0.709	0.936	0.973	1.000
TabICL	95% $\lvert\cos\rvert$	0.859	0.846	0.824	0.761	0.544	0.858	0.844	0.845	0.833	0.824
	dispersion	10.60	11.00	10.58	9.19	7.05	10.87	10.30	10.49	9.99	9.79
	AUROC	–	1.000	1.000	1.000	1.000	–	0.772	0.932	1.000	1.000

AUROC of the mixed batch against the $\alpha = 0$ reference. Figure 18 plots the two collapse statistics; Table 16 reports all values.

The tilting prediction holds in the deployed multi-layer models, with exactly the severity profile the theory assigns it. On Covertype, both collapse statistics fall sharply and monotonically with the OOD fraction: the cosine proxy drops from 0.939 to 0.537 (TabPFN) and from 0.859 to 0.544 (TabICL), and dispersion contracts by factors of 2.1 and 1.5. On Waterbirds the same decreases appear but are mild (cosine ratios 0.99 and 0.96), consistent with Theorem 4.1's scope: the collapse is a far-OOD limit phenomenon, and a near-OOD subpopulation shift should, and does, produce only a weak echo of it. Detection AUROC rises with α in step with the collapse statistics in both models. Although the tilting argument is unproven for general transformer stacks, its measurable prediction is what the deployed multi-layer models actually do, precisely where Theorem 4.1 invokes it; the multi-layer proof itself remains open (Appendix A.8).

D.8 Theorem 4.1's Condition and Premise at Finite Severity

Section 5.4 of the main paper gives an abridged version of this analysis (Figure 3b); the full estimates and the per-dataset λ-gap are collected here.

Table 17: **Theorem 4.1's condition at finite severity.** Empirical estimates of both sides of $\epsilon < \gamma/(B_\phi^2 B_C)$ (TabPFN). The condition fails on every dataset: $\hat{\epsilon}$ is $O(1)$ while the right-hand side is near zero or negative. The AUROC column repeats Table 10's selected-LR values for reference.

Dataset	$\hat{\epsilon}$ (q95)	$\hat{\gamma}/(\hat{B}_\phi^2 \hat{B}_C)$	AUROC (Table 10)
Covertype	0.497	-0.0033	1.000
CIFAR-10⟶10.1	0.771	-0.0010	1.000
Camelyon17 (DINOv2)	0.920	-0.0003	0.675
Camelyon17 (ResNet18)	0.928	$+0.0006$	0.977
CelebA	0.951	$+0.0083$	0.919
Credit Default	0.951	-0.0951	1.000
Waterbirds	0.967	-0.0047	0.921
UCI Heart	0.971	-0.1744	0.908
Bank Marketing	0.973	-0.1138	0.997
Diabetes Readmission	0.991	-0.0791	0.976

Table 18: **The proof's premise, measured directly on the finetuning trajectories.** $\Delta_{\mathrm{traj}} > 0$ iff the gradient-interference mechanism of Theorem 4.1 operates; the λ-gap is the raw ID/OOD separation of the statistic. Both columns are computed under fixed $\eta = 10^{-3}$ (within-configuration); they are not comparable to Table 10 or to the selected-LR AUROC column of Table 17.

Dataset	Δ_{traj}	λ-gap
Waterbirds	$+87.2$	$+0.255$
Credit Default	$+40.3$	$+0.198$
Diabetes Readmission	$+22.6$	$+0.133$
Bank Marketing	$+10.8$	$+0.044$
Covertype	$+8.2$	$+0.158$
Camelyon17 (ResNet18)	$+7.6$	$+0.228$
CelebA	$+3.0$	$+0.080$
Camelyon17 (DINOv2)	$+1.5$	$+0.234$
CIFAR-10⟶10.1	$+0.11$	$+0.006$
UCI Heart	-4.4	$+0.008$

Theorem 4.1 certifies the signed separation only when $\epsilon < \gamma/(B_\phi^2 B_C)$, a condition stated in worst-case constants, and the question is whether it is ever satisfied on the benchmark at finite shift severity — or whether the theory certifies a regime the experiments never visit. We seek to determine, per dataset, (i) whether the theorem's condition holds as stated, and (ii) when it does not, whether the gradient mechanism the proof is built on is nonetheless present, by measuring the proof's premise directly instead of routing it through the worst-case constants.

We set up the experiment as follows. On ten datasets (TabPFN) we estimate every quantity in the condition from the factorization of Lemma A.1: $\hat{\epsilon}$ is the 95th percentile of the absolute anchor–eval feature cosines (centered by the pooled-OOD-mean proxy for ϕ_∞; the 95th percentile is a *lenient* estimate, since the assumption formally requires a uniform bound), and $\hat{\gamma}$, $\hat{B}_\phi$, $\hat{B}_C$ are the empirical versions of the ID-interference and boundedness constants. Separately, we measure the proof's premise on the finetuning trajectories themselves: $\Delta_{\mathrm{traj}} := \mathbb{E}_{\mathrm{OOD}}[\langle \bar{g}_a, \bar{g}_e \rangle] - \mathbb{E}_{\mathrm{ID}}[\langle \bar{g}_a, \bar{g}_e \rangle]$, the gap in the time-averaged anchor–eval gradient alignment of Proposition A.8 computed by exact full-decoder autograd, together with the λ-gap $\mathbb{E}[\lambda \mid \mathrm{ID}] - \mathbb{E}[\lambda \mid \mathrm{OOD}]$, the raw separation available to the detector (here and below we write $\mathbb{E}[\lambda \mid \mathrm{ID}]$ for the main paper's $\mathbb{E}[\lambda_{\mathrm{ID}}]$, and likewise for OOD). The proof's mechanism operates iff $\Delta_{\mathrm{traj}} > 0$: the disagreement gradient interferes with anchor descent more strongly under ID than under OOD inputs.

These results show a sharp split between the condition and the mechanism. The condition fails everywhere: $\hat{\epsilon} \geq 0.497$ on every dataset while the right-hand side never exceeds $+0.0083$ and is negative on eight of ten, so at finite severity every benchmark is near-OOD in the theorem's sense and the formal guarantee certifies none of them (Covertype, the most separated, comes closest). The failure, however, lives in the worst-case constants, not in the mechanism: the premise $\Delta_{\mathrm{traj}} > 0$ holds on nine of ten datasets, including the moderate-overlap benchmarks where the static condition is vacuous. More tellingly, the two datasets where the premise is weakest or fails — CIFAR-10

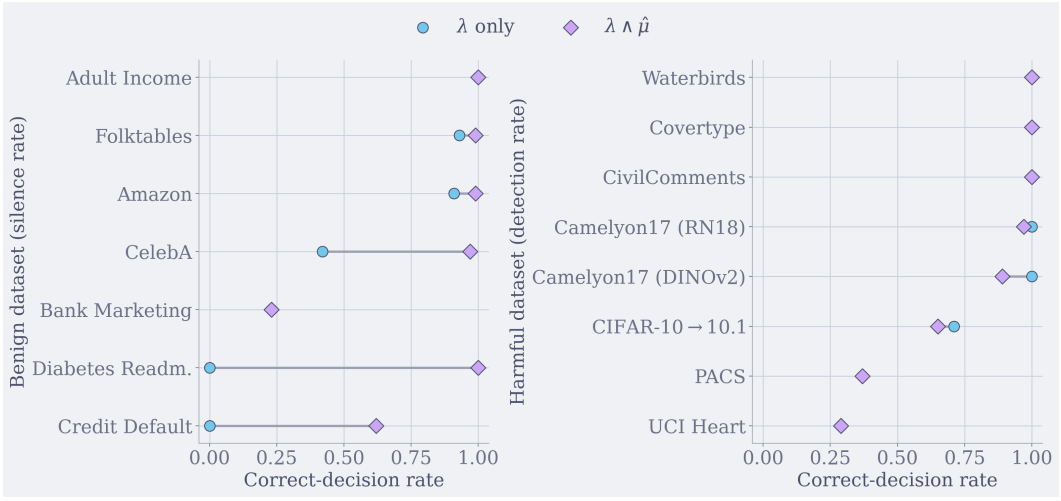


Figure 19: **Two-signal detector vs. λ alone.** Correct-decision rate per dataset (benign: silence; harmful: detection), TabPFN with fixed $\eta = 10^{-3}$. A diamond to the right of its circle is an improvement from the $\hat{\mu}$ gate. The gate removes the benign false alarms on CelebA, Credit Default, and Diabetes Readmission, cannot help Bank Marketing (whose confidence collapses), and costs little on harmful shifts.

($+0.11$) and UCI Heart (-4.4) — are exactly the two with near-zero λ-gap ($+0.006$, $+0.008$): the measurement predicts the detector's weakest cells in advance rather than rationalizing its successes after the fact. We conclude that Theorem 4.1 certifies detection in the far-OOD regime only, that the gradient-interference mechanism it is built on is empirically present across the near-OOD benchmarks, and that a finite-ϵ bound with measurable constants in place of worst-case ones is the right target for future theory.

D.9 Two-Signal Harm Detection: Empirical Evaluation

The deployed statistic λ detects distributional difference, not harm: Lemma A.12 shows the accuracy gap depends only on mass moved onto the model's mistake region, which no label-free statistic can see, and empirically λ false-alarms on the benign Bank Marketing, Diabetes Readmission, Credit Default, and CelebA shifts. Appendix A.13 proposes the confidence-drop statistic $\hat{\mu}$ (Proposition A.21) as a harm-magnitude estimator, with λ as a shift-confirmation gate; here we are interested in whether the composed detector actually restores benign-shift silence, and at what cost in harmful-shift detection. Two properties hold before any measurement: the conjunction fires only when both tests fire, so its false-alarm rate under the null is at most the smaller of the two individual rates, under arbitrary dependence and with no Bonferroni correction (the conformal bound $\alpha + 1/(K{+}1)$ carries over unchanged); and under *any* shifted distribution the $\hat{\mu}$ gate can only remove alerts produced by the λ-test, never add them, so it cannot increase benign false alarms.

We set up the experiment as follows: all 15 datasets, TabPFN, fixed $\eta = 10^{-3}$, with $\hat{\mu} = \hat{\mathbb{E}}_{\mathrm{ID}}[p_x(y)] - \hat{\mathbb{E}}_{\mathrm{eval}}[\max_c p_x(c)]$ estimated per Definition A.17. The composed detector fires iff the λ-test fires and $\hat{\mu}$ exceeds its threshold. Figure 19 shows the per-dataset movement from λ alone to the conjunction; Table 19 reports the correct-decision rate of each configuration: the silence rate on benign datasets and the fire rate on harmful ones.

The gate works where the theory says it should. Mean benign silence rises from 0.50 to 0.83: the gate removes the false alarms on three of the four problematic benign datasets (CelebA $0.42 \longrightarrow 0.97$, Credit Default $0.00 \longrightarrow 0.62$, Diabetes Readmission $0.00 \longrightarrow 1.00$). Bank Marketing remains a failure ($0.23 \longrightarrow 0.23$) because the model's confidence collapses under that benign shift, so both signals fire — the one benign case where the gate's own premise is violated. On harmful shifts the mean detection rate drops only from 0.80 to 0.77, with the losses concentrated where the λ-test is already weak under this fixed configuration (PACS, UCI Heart) rather than created by the gate. The construction has a real limitation that bounds how far it should be trusted: $\hat{\mu}$ reads the model's own confidence,

Table 19: **Two-signal detector: correct-decision rates.** Benign rows report the silence rate (correct = no alarm); harmful rows report the detection rate (correct = alarm). All 15 datasets, TabPFN, fixed $\eta = 10^{-3}$ (within-configuration; not comparable to Table 10). Means: benign silence $0.50 \to 0.83$, harmful detection $0.80 \to 0.77$.

Dataset	Shift	λ only	$\hat{\mu}$ only	$\lambda \wedge \hat{\mu}$
Adult Income	benign	1.00	1.00	1.00
Amazon	benign	0.91	0.97	0.99
Bank Marketing	benign	0.23	0.00	0.23
CelebA	benign	0.42	0.97	0.97
Credit Default	benign	0.00	0.62	0.62
Diabetes Readmission	benign	0.00	1.00	1.00
Folktables	benign	0.93	0.75	0.99
Camelyon17 (DINOv2)	harmful	1.00	0.89	0.89
Camelyon17 (ResNet18)	harmful	1.00	0.97	0.97
CIFAR-10→10.1	harmful	0.71	0.90	0.65
CivilComments	harmful	1.00	1.00	1.00
Covertype	harmful	1.00	1.00	1.00
PACS	harmful	0.37	1.00	0.37
UCI Heart	harmful	0.29	0.98	0.29
Waterbirds	harmful	1.00	1.00	1.00

so an overconfident or poorly calibrated model would silence true alarms, and Proposition A.21's error terms are exactly calibration gaps. A formal treatment of when the confidence signal can be trusted is beyond the scope of this work. The two-signal construction is an evaluated proposal that demonstrably recovers benign-shift silence at modest cost, while λ remains the primary, conformally certified statistic of **CALD**.

D.10 Detectability Across Shift Types

CALD monitors unlabeled covariates, so its reach is bounded by a structural fact: a shift is visible only insofar as it changes the covariate marginal $p(X)$. The PFN's pretraining prior suggests organizing shifts by where they enter the data-generating process — TabPFN's synthetic tasks are drawn from randomly sampled structural causal models, so one can ask which interventions on the SCM a label-free monitor can see. A shift in the label prior $p(y)$ changes the mixture of class-conditional covariate distributions and therefore moves $p(X)$; a concept shift that rewrites $p(y \mid X)$ while leaving $p(X)$ fixed is invisible in principle, however harmful. We seek to determine whether the deployed detector respects this taxonomy, using controlled shift injections where the benchmark's natural shifts cannot separate the two.

We set up the experiment as follows: TabPFN, fixed $\eta = 10^{-3}$, 2 seeds. For label-prior shift we resample the evaluation batch's class proportions toward a skewed target prior at full strength, keeping each class-conditional covariate distribution intact; for concept shift we relabel the evaluation batch through a permuted class map at full strength, changing $p(y \mid X)$ while leaving the covariates untouched. Figure 20 plots detection against the harm each injected shift inflicts; Table 20 reports detection AUROC at full shift strength against the no-shift baseline, alongside the accuracy drop the concept shift inflicts.

The detector obeys the taxonomy almost exactly. Concept shift is invisible: on nine of ten datasets, AUROC under a full-strength concept shift stays within 0.03 of the no-shift baseline even as accuracy falls by 52 to 98 percentage points — this is the impossibility of Lemma A.12 realized in its purest form, since no label-free monitor of any kind could detect these shifts. Covertype's 0.79 is the one visible exception, and it is instructive: its class-stratified evaluation sampling slightly perturbs the covariate distribution as a side effect of the relabeling, and the detector picks up precisely that covariate residue, not the concept change itself. Label-prior shift, by contrast, is detected strongly wherever the resampled prior substantially moves the covariate mixture (seven of ten datasets at 0.92–1.00); it is missed on Bank Marketing and Diabetes Readmission, whose class-conditional covariate distributions overlap heavily, so that even a full prior swap barely moves $p(X)$ — consistent with the mechanism, since the detector sees only the covariate marginal. UCI Heart is confounded in

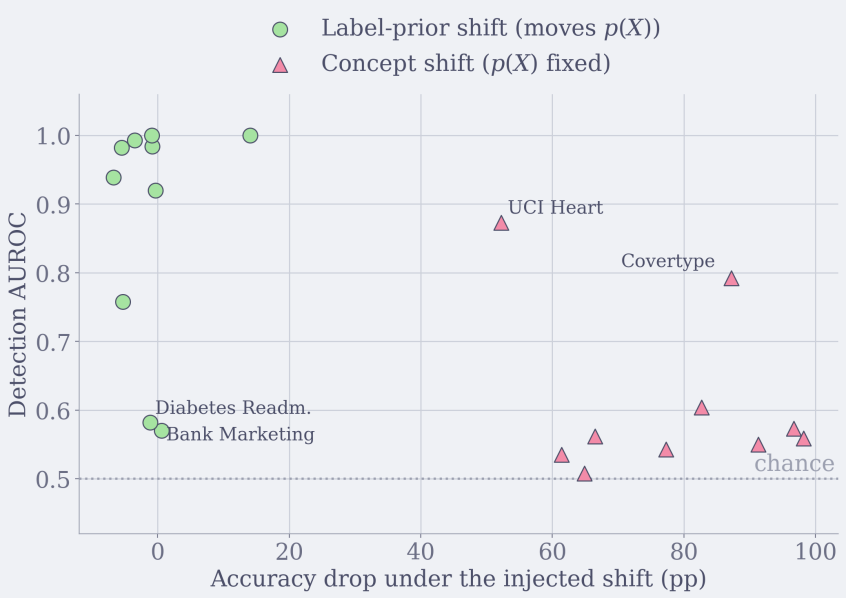


Figure 20: **Label-prior vs. concept shift at full strength.** Each point is one dataset under one injected shift (TabPFN, fixed $\eta = 10^{-3}$, 2 seeds; dotted line: chance). Concept shifts (triangles) destroy accuracy yet sit at their no-shift detection baseline; label-prior shifts (circles) barely move accuracy yet are detected wherever they move $p(X)$. The annotated points are the exceptions discussed in the text.

Table 20: **Label-prior vs. concept shift at full strength.** Detection AUROC (TabPFN, fixed $\eta = 10^{-3}$, 2 seeds; within-configuration, not comparable to Table 10). The no-shift column is the same protocol at strength 0 and marks each dataset's chance baseline. Concept shift destroys accuracy while leaving the covariates untouched, and detection stays at its baseline; label-prior shift moves the covariate marginal and is detected wherever that movement is substantial.

Dataset	No shift	Label-prior	Concept	Concept Δacc (pp)
Adult Income	0.51	0.98	0.51	−64.9
Bank Marketing	0.61	0.57	0.60	−82.7
Camelyon17 (DINOv2)	0.55	0.98	0.55	−91.3
Covertype	0.57	1.00	0.79	−87.2
Credit Default	0.57	0.94	0.56	−66.5
Diabetes Readmission	0.55	0.58	0.54	−77.3
Folktables	0.56	0.99	0.54	−61.4
PACS	0.60	1.00	0.57	−96.7
UCI Heart	0.88	0.76	0.87	−52.2
Waterbirds	0.56	0.92	0.56	−98.2

both directions: a residual covariate quirk of its ID split elevates its no-shift baseline to 0.88, and neither injected shift moves the detector far from that baseline. **CALD** detects exactly the SCM interventions that reach $p(X)$ and is structurally blind to those that do not; shifts expressible as re-sampling the SCM's exogenous inputs are the detector's natural domain, while mechanism changes with fixed covariates require labels and are out of scope (Appendix B).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Each claim in the abstract and introduction is substantiated: the algorithm (Section 4), calibration-to-harm and benchmark results (Section 5, Table 1), and the theoretical guarantees (Theorem 4.1, Proposition 4.2).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix B catalogues out-of-scope failure modes (label shift, concept shift, adversarial inputs) and the regimes where Theorem A.9's guarantee becomes vacuous (near-OOD with overlapping supports).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Section 4.2 presents simplified statements of the main results; complete assumptions and proofs are in Appendix A, with all assumptions consolidated in Appendix A.14.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Algorithm 1 gives the full pseudocode; Section 5.1 and Appendix C list all datasets, baselines, hyperparameters, and per-dataset preprocessing.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code with reproduction scripts is released at `https://github.com/teivng/cald`; all datasets are publicly available and their sources are described in Appendix C.1.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5.1 specifies the setup; Appendix C lists all hyperparameter values ($T = 50$, $\eta \in \{10^{-2}, 10^{-3}, 10^{-4}\}$ selected per cell by the tail-stability rule, $K = 50$, $\alpha = 0.05$) and per-dataset configurations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results are reported as means with standard deviations across 10 seeds (Appendix C.4); Type-I error is closed-form-bounded by $\alpha + 1/(K+1)$ via Proposition 4.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments run on a single NVIDIA H100 GPU; total compute is on the order of tens of GPU-hours, with per-step runtimes in Appendix D.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and adhered to the NeurIPS Code of Ethics; the work involves no human subjects or scraped personal data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Positive impact is safer deployment of tabular foundation models in high-stakes domains; the main risk is over-trusting the test outside its proven scope, which Appendix B explicitly catalogues to mitigate.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no high-misuse-risk artifacts; **CALD** is a deployment-time monitor over existing pretrained PFNs.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All models (TabPFN, TabICL) and datasets are cited and used under their original licenses; dataset sources are described in Appendix C.1.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The **CALD** implementation is released at `https://github.com/teivng/cald` with a README documenting the API and run commands.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as any component of the core methodology. The PFNs studied (TabPFN, TabICL) are tabular in-context learners, not LLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.