ALIGNED TREE-STATE TRANSFER FOR ONE-MESSAGE CROSS-SILO ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-silo personalization often works from one model transfer and few receiver labels. We ask when the internal state a shipped tree ensemble computes for a record adds information beyond the score it reports. SHIFT (SHipped-model Internal-state Feature Transfer) has the receiver run the immutable teacher on its own records, compute a TreeSHAP vector for each, and train a student on that aligned state with the teacher score and its permitted raw features; nothing is uploaded and the source never participates again. Under a restricted interface, where the student reads a subset of features plus the teacher's declared outputs, aligned state improved AUC over an otherwise identical score-only student by $+.0058$ on 18 of 22 tasks after correction, surviving controls that break record-state pairing or match the state's width. A within-protocol intervention locates the cause in access itself: the same state is worth $+.0004$ AUC when the student reads every column and $+.0058$ when it does not, an interaction of $+.0054$ on 19/22 datasets at Holm $p < .001$ in that experiment's two-test family. State also improves what a receiver *selects*: across 7 tasks whose permitted and withheld fields follow documented semantics, against a scalar-only candidate set free to decline adaptation, selected test NLL improved by $+.0128$ on 7/7 tasks (CI $[+.0046, +.0219]$, exact $p = .016$, a single uncorrected test). A label-free leaf reduction matches the score-residualized form in our tests at a fraction of the extraction cost: the contribution is the interface and its characterization under restricted access, not a unique advantage for Shapley attribution. State is useful because it carries information from the teacher's complete input, and recoverable for the same reason.

1 INTRODUCTION

Consider two organizations with separate data governance. A data-rich source has trained a tree ensemble, which we call the *teacher* throughout. A *receiver* serves a different population with fewer labels. Records cannot be pooled and repeated coordination is costly, so the teacher ships once and nothing comes back; this is the one-way limit of one-shot personalized federated learning (Huang & Shu, 2025). Training on receiver labels alone discards the teacher, and distilling it transfers only what its output already states, which is one number per record. Using a trained tree ensemble to produce features for another learner is not new (He et al., 2014; Moosmann et al., 2006); the open question is narrower: *what does a frozen teacher's aligned internal state add beyond its predictions when the downstream learner has a restricted input interface?*

That question cannot be answered by a system benchmark. A workflow that ships a teacher and trains a student differs from its comparators in several things at once: the teacher, how its probabilities are rescaled, how they are mixed with the student's, and the representation, so an advantage it wins is not attributable to any of them. We therefore hold everything else fixed, the teacher, the student learner and the rule that combines their outputs, and vary only what the student reads, and we add controls that break the pairing between a record and its own state and that match the state's width with a random projection, since added dimensions and an unaligned feature block are the two explanations a raw gain cannot rule out.

The source sends one immutable tree artifact and exits. No receiver record, label, explanation or update crosses back, and the receiver's own model may not read part of what that artifact reads. Three questions decide whether the teacher's score is enough.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Why would the score not be enough? A tree ensemble computes far more per record than it reports. The prediction is a sum of contributions from the features the ensemble split on, and those splits can involve fields the receiver's model is not permitted to read; the score collapses that account into one number, so records scored alike need not be records the teacher treats alike.

Why would state help only under restriction? Because state is computed from the input. A receiver that already reads the whole record can in principle derive it, so no population information is added, and Equation 1 makes that precise. The interesting case is the one where the receiver may run the teacher but may not read part of its input.

Is this attribution as features? No. The effect survives replacing TreeSHAP with a label-free leaf reduction, so what matters is that the state is aligned to the record and comes from the teacher, not which attribution semantics produced it.

We report what the interface buys and what it costs, including the settings where the fields the student may not read carry little the receiver can use, and it does not pay at all. We formalise frozen state transfer under asymmetric downstream feature access and give the identity that says when it can pay (Equation 1). On 22 tabular tasks aligned state improves AUC over an otherwise matched score-only student and over controls that shuffle which record each state vector belongs to, or replace the state with a random projection of the same width, and the gain survives giving both students every scalar the teacher *bag*, the three-model ensemble supplying the probability the student is mixed with, already computes. A prospectively specified access-by-state intervention attributes the gain to the restriction itself rather than to the extra columns. The effect is representation-general: a label-free leaf encoding is not separated from the attribution channels, so the interface, not Shapley attribution, is what matters.

2 Related Work

Representation and explanation transfer. Distillation can transfer hidden representations or match a student's explanations to its teacher's (Srinivas & Fleuret, 2018; Aguilar et al., 2020; Chowdhury et al., 2023; Parchami-Araghi et al., 2024; Gao et al., 2024). We use a fixed teacher's state as an *input* to receiver adaptation rather than as a target to match. Since TreeSHAP's additive decomposition duplicates the score (Lundberg et al., 2018; 2020), one variant removes its fitted affine score component. Consuming aligned tree state as downstream features has precedent in terminal-leaf indicators (Moosmann et al., 2006; Vens & Costa, 2011; He et al., 2014; Wolpert, 1992; Feng & Zhou, 2018; Shen et al., 2025). SHIFT builds on these encodings in a different setting, where the source encoder is already trained, remains immutable and leaves the source after one transfer. Our contribution is to isolate what its state adds beyond scores when the receiver's trainable component has a restricted feature interface.

Tree-model transfer and label-shift correction. Boosted-tree transfer reweights source instances (Dai et al., 2007) or edits source trees on target data (Segev et al., 2017). SHIFT leaves the source trees untouched. Label-shift correction divides by the source prior (Lipton et al., 2018; Saerens et al., 2002; Garg et al., 2020); our map divides by the teacher's mean prediction on receiver inputs instead, making it an empirical prevalence-alignment heuristic (Appendix D). Shapley variants differ in semantics (Sundararajan & Najmi, 2020; Janzing et al., 2020), so we fix exact path-dependent TreeSHAP as one encoder and compare it against label-free leaf encodings rather than adjudicate between them. Inference follows Demšar (2006).

Restricted interfaces and privileged information. Learning using privileged information trains with features absent at test time (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016), and vertical FL keeps feature holders available (Wu et al., 2024). Here the teacher stays executable on $A \cup B$ while the receiver-owned component is restricted to A and exposed teacher outputs.

One-shot personalization and source-model adaptation. One-shot federated learning aggregates client models, statistics or representations into a global predictor (Guha et al., 2019; Allouah et al., 2024; Jhunjhunwala et al., 2024; Guan et al., 2025; Yang et al., 2026), and its personalized variants target local predictors by distilling peer models (Huang & Shu, 2025). Both presume peers and a coordinator. We have one frozen artifact and no return channel, so the question is what that

Table 1: **The three interfaces this paper compares.** X_A: fields the student may read; X_B: fields only the teacher may read; s: its authorized scalars; R: its per-record state. The teacher always reads the complete record; the last column is what Equation 1 permits, not what a learner achieves.

Interface	Teacher reads	Student reads	Room for state to help
Complete access	X_A, X_B	X_A, X_B, s	none: R is a function of what it holds
Restricted, score-only	X_A, X_B	X_A, s	(the baseline)
Restricted, with state	X_A, X_B	X_A, s, R	up to $I(Y; R \mid X_A, s)$

artifact may expose, not how to aggregate several. Federated tree systems exchange gradients, histograms or updates (Ma et al., 2023; Li et al., 2023), and source-hypothesis transfer (Liang et al., 2020) adapts a frozen neural source model without target labels, updating the source hypothesis by entropy and pseudo-label objectives. SHIFT shares its defining constraint, that the source data never reappear, and differs on the three axes that decide the method: the receiver is *labeled*; the tree artifact is never updated, so adaptation lives entirely in a separate receiver model; and that receiver model is restricted in what raw features it may read, the condition our evidence turns on and one source-free adaptation does not impose.

3 SHIFT

3.1 Problem setting

Let $X = (X_A, X_B)$ and let Y take values in a finite label set. A data-rich source fits a tree ensemble $\mathcal{M}$ on labeled data over $A \cup B$ and sends the serialized artifact once. The receiver holds complete records for its own population and executes $\mathcal{M}(X_A, X_B)$ locally. Its adaptation model consumes X_A and declared outputs of $\mathcal{M}$. The *complete-feature interface* sets $B = \varnothing$. The *restricted interface* keeps raw X_B inside the teacher. This receiver-local schema is distinct from the cross-silo record boundary, and neither interface needs a return message or source records. A score-only interface exposes only the teacher margin. SHIFT adds a fixed-width description of the same teacher's aligned tree state.

What state can add, and when it cannot. Write $R = g(X)$ for the frozen state the teacher exposes and $m = f(X)$ for its score, each a deterministic function of the complete record. We do not assume the score is recoverable from the exposed state, which fails for Leaf-SVD and for the score-residualized channel; the identity needs only that both are fixed functions of X. For a predictor class rich enough to reach the conditional distribution, the lowest expected log loss an interface permits is the entropy of Y given what it exposes. A receiver restricted to X_A thus faces $H(Y \mid X_A, m)$ with the score alone and $H(Y \mid X_A, m, R)$ with the state, so admitting state lowers the attainable loss by exactly

$$I(Y; R \mid X_A, m), \tag{1}$$

the label-relevant information the score discarded. Under complete access R is a function of what the receiver already holds, $H(R \mid X) = 0$, so $I(Y; R \mid X, m) = 0$ and no population gain exists; an improvement measured there comes from the representation and the finite sample, not from new information. This is an identity, not a promise the gain is reachable: the model class and label budget decide how much is realised. It says where to look, and §5.2 tests that contrast.

The restricted interface models a teacher governed separately from its downstream student (Table 1). This is an output-access constraint, not missing-data imputation: the teacher executes as a sealed component inside the receiver's environment. The receiver may invoke it on records it holds and read the outputs it is entitled to, while X_B stays under a separate permission, as with fields from a vendor feed, a device stream, or a clinical system the analytics team is not cleared to query. The boundary constrains what the receiver's own modelling may read, not what exists: a receiver that does control X_B can train on it directly. We instantiate B as high-gain tabular features, physical sensor blocks, or clinical fields separated from administrative ones (Appendix B). These were experimental interface rules, not evaluations of a deployed access policy. Derived state can reveal information about X_B, so permission to expose it is a separate requirement from keeping records within the receiver (Section 6).

3.2 ALIGNED TREE-STATE REPRESENTATION

SHIFT represents each receiver record by the teacher's aligned tree state in one fixed coordinate system. The artifact plays two roles and we separate them: a three-member bag $\mathcal{M} = (M_0, M_1, M_2)$ supplies the probability the student is mixed with, while M_0 alone supplies state and margin, so every representation control shares one encoder and one extraction pass. Leaf identities are model-specific, and aggregating feature-indexed attributions across members is a separate option we did not evaluate. A later control gave the student every scalar the whole bag computes and did not change the conclusion (§5.1). For receiver example x, let $m(x)$ be this reference teacher's logit and $\phi(x) \in \mathbb{R}^d$ its exact path-dependent TreeSHAP vector. The interface is defined by what it exposes rather than by one encoder: $R(x)$ denotes any fixed-width aligned-state channel read from the same reference teacher, and we evaluate three. *Leaf-SVD* is a truncated SVD, fitted on receiver-fit rows only and held fixed thereafter, of the sparse one-hot terminal-leaf identities across that teacher's trees, matched to the attribution channels' width; terminal-leaf encodings are established (He et al., 2014; Moosmann et al., 2006), and we claim the setting and the measurement rather than the encoder. *SHIFT-Raw* uses $R_{\text{raw}}(x) = \phi(x)$ directly and keeps the already exposed margin as a separate input. *SHIFT-Q* is a score-residualized form that estimates

$$\widehat{W} = \arg\min_{W} \|\Phi - [\mathbf{1}, m]W\|_F^2, \qquad \Phi^{\perp} = \Phi - [\mathbf{1}, m]\widehat{W}, \tag{2}$$

and applies this map unchanged to validation and test rows. TreeSHAP coordinates sum to the teacher margin minus its bias term, so raw state already contains the score the student is given; Equation 2, fitted on receiver-fit rows only, removes each coordinate's fitted affine dependence on it but not every nonlinear one. At a fixed $\widehat{W}$ this is a reparameterization, not a change of information: the student receives m either way, so (m, Φ) and $(m, \Phi^{\perp})$ determine each other. They can still score differently, and do, because a finite tree learner is not invariant to how its inputs are rotated: what changes is the geometry, not what the interface exposes. Both are measured against the same score-only comparison. SHIFT-Q was declared primary before the benchmark ran and stays primary although SHIFT-Raw ends with the higher mean: promoting whichever channel wins afterwards is what a declared primary prevents. That status governs which hypothesis we test, not which channel to deploy; §6 gives the operational recommendation, which is conditional and is not SHIFT-Q everywhere. Leaf-SVD plays two roles. It was declared as one of the four label-free controls corrected together in the 16-contrast family and is reported that way; later experiments also run it as a channel in its own right. Presenting it as both changes no statistical declaration.

Multiclass extension. For the K-class DSADS validation we replace ϕ in Equation 2 by the predicted-class contrast against the class mean, which keeps the state width at d rather than Kd. Both constructions are stated in full in the supplement's pipeline section.

3.3 RECEIVER-LOCAL ADAPTATION

The receiver student is an XGBoost (Chen & Guestrin, 2016) classifier on $z_R(x) = [x_A, m(x), R(x)]$. The defining operation is row alignment: the receiver computes $R(x)$ locally for the same record it uses to fit and serve the student. The deployed predictor is the fixed mixture

$$p_R(x) = (1 - \lambda)p_{\text{align}}(x) + \lambda p_S(y \mid z_R(x)), \tag{3}$$

where p_{align} reweights the teacher's probability toward the receiver's class mix, an empirical heuristic rather than source-prior correction (Appendix D). We fix $\lambda = \frac{1}{2}$ so that every representation relies on the anchor equally: the comparison of interest is between students, and a tuned blend would let one arm borrow more from the teacher than another. Choosing λ per cell by validation NLL instead moves it off $\frac{1}{2}$ in 69% of cells and leaves the result standing ($+.0036$ AUC, Appendix H), so the fixed weight is not carrying it. The deployed guarded form, which also decides whether to fit a student at all, is specified in Appendix D and used only where a table says so.

Representation choice and hypothesis. We keep all coordinates, not only those of the withheld block, because what a tree assigns to a visible feature changes when a hidden one alters the path. A gain could mean two things: that state carries information about the withheld block itself, or only that it reveals which path a record took through the teacher. Shuffle, leaf-state and projection controls, a full-feature comparison, coordinate-block ablations and the reference test of Section 5.1

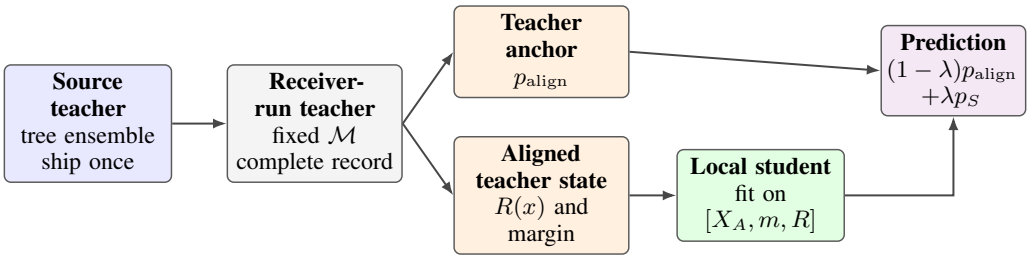


Figure 1: SHIFT uses the shipped teacher twice: a prior-aligned anchor supplies its probability signal, an aligned-state channel $R(x)$ supplies record-aligned state to the student. Communication is one-way and occurs once; all receiver records stay local. Setting $B = \varnothing$ gives the complete-feature interface.

point to the first under the restricted interface. The complete-feature published-method comparison did not resolve an additional state AUC gain.

4 Experimental Setup

One question organises the experiments: what aligned state adds to an otherwise matched receiver, and under what access conditions. the supplement names every benchmark, its data, interface and correction family, and the baselines it compares SHIFT against. The representation benchmark answers the question directly, holding the teacher, receiver, anchor and blend fixed and varying only what the student reads; it, the 12-task extension, the sensor study and the reference test are restricted-interface mechanism studies run against the score-only student, alignment-breaking and leaf-state controls, a direct-access reference and two or three teachers. The matched access-by-state experiment of Section 5.2 then varies the access condition itself inside one protocol.

4.1 Benchmarks and silo partitions

Controlled tabular benchmark. The 22 tasks came from three established OpenML suites (Vanschoren et al., 2014): eight TabZilla-Hard (McElfresh et al., 2023), eight TabArena (Erickson et al., 2025) and six from Grinsztajn et al. (Grinsztajn et al., 2022), selected by metadata and dataset identity rather than observed performance and capped at 20,000 rows (exact tasks and exclusions in the supplement). For each task and seed a per-class Dirichlet split with randomly drawn class proportions sent 85% of rows to the source and 15% to the receiver. The teacher saw every feature. Ranking features by *gain*, the total loss reduction their splits produce in the fitted teacher, the top 50% became the withheld block B, and receiver-trained arms saw only A and declared teacher outputs. All preprocessing was fitted on source rows alone.

Shift-condition sweeps. Two 20-task sweeps tested state-augmented students with complete features, one source teacher and no teacher–student blend, so they are auxiliary representation tests rather than the main SHIFT protocol. The covariate sweep resampled receiver rows along a random feature projection at four sharpness levels. The concept sweep flipped labels in feature-defined regions at four target fractions while preserving the receiver's class counts, and included a row-permuted state control. Each severity used five seeds. Appendix J gives the construction and states the outcome; the per-severity table is in the supplement.

Natural sensor benchmark. We specified the evaluation on the DSADS activity-recognition dataset (Altun et al., 2010), 19 activities over eight subjects and five physical sensor units, before downloading it. Teachers saw seven subjects and every unit. Each receiver saw one unit. Fixed indices separated fit, validation and test, and all 360 cells were run.

4.2 Compared methods and controls

Each of three source teachers had 400 depth-10 trees and receiver students 200 depth-8 trees, sharing the artifact and a fixed 50/50 blend. The shuffle uses clusters fitted on $[X_A, m]$, preserving local state

distributions while breaking record alignment, and the random-projected leaf state is dimension-matched to SHIFT-Q. The supplement's experimental details define the remaining systems controls.

4.3 Metrics and statistical analysis

Status of each analysis. Claims rest on studies whose endpoints, contrasts, correction family and prediction were written down and hashed before their runner existed: the matched state-versus-score comparison declared with the benchmark (Table 2), the richer scalar interface (Table 13), the access interaction (Table 12), the access-by-encoding study (Table 11) and the semantic deployment comparison (Table 10). These reuse the benchmark's datasets and splits, and fixing a plan before a new run does not make previously observed arms unseen; each specification records which arms already existed. Controls, the direct-access reference and the cost and exposure measurements support those claims without having been specified in advance. Everything else is retrospective analysis of completed runs and is labelled where it appears.

AUC is the primary endpoint except in the deployment comparison, which uses NLL because that is what its selector optimises, and in the two sensor studies, whose declared endpoint is also NLL. The declared mechanism and strong-control families (Tables 5, 6) correct across all four metrics jointly, as registered; elsewhere a secondary metric is reported beside the primary one and not corrected with it. Validation selected the macro-F1 and balanced-accuracy thresholds. We average seeds within dataset, use datasets as inference units, and report paired differences, dataset-bootstrap 95% intervals and two-sided Wilcoxon tests. Tables say which test each contrast used, exact or approximate; zero differences and ties force the normal approximation in places. A contrast is *resolved* when it passed the stated .05 test and *unresolved* when it did not establish a difference, which is not evidence the arms are equal. Counts such as $16/22$ denote units favouring the first arm. Holm correction covered the stated teacher and mechanism families. For DSADS, validation NLL selected the proposed method and strongest comparator before test evaluation, with Holm correction over the two planned contrasts.

5 Results

5.1 What does aligned state add beyond the score?

Table 2: **Representation benchmark** at the *restricted* interface (22 tasks, five seeds). Every rich representation used the same receiver, anchor and blend; teacher-only controls used no student. Bold marks the best mean per column. Students read only permitted features and macro-F1 is threshold-tuned, so these columns are not comparable with the complete-feature table.

System	AUC	NLL ↓	tuned mF1	tuned BalAcc
SHIFT-Raw	**.8460**	**.3293**	.7528	.7838
SHIFT-Q	.8443	.3312	.7506	.7837
Score-only student	.8385	.3355	.7452	.7776
Prior-aligned teacher	.8454	.3590	**.7536**	.7820
Neighborhood-shuffled SHAP	.8400	.3345	.7443	.7750
Random-projected leaf state	.8396	.3351	.7459	.7771
Supervised leaf target	.8437	.3308	.7506	.7814
Leaf-SVD	.8442	.3336	.7516	**.7840**
Direct-access reference (X_B)	.8447	.3306	.7509	.7769

On the primary 22-task restricted-interface benchmark, SHIFT-Raw achieved the highest mean AUC and lowest NLL among common-anchor adaptation arms, improving on the score-only student by $+.0075$ on 20/22 datasets (Table 2). SHIFT-Q improved AUC by $.0058$ over the score-only student (18/22), $.0043$ over neighborhood-shuffled SHAP (18/22) and $.0047$ over random-projected leaf state (19/22), at Holm $p = .033$, $.018$ and $.009$ respectively in the 16-contrast mechanism family that corrects the four label-free state controls together. Those are adjusted values, not raw ones (Table 5). These three contrasts are on AUC; within the same 16-contrast family the score-only contrasts on NLL, tuned macro-F1 and tuned balanced accuracy point the same way without passing it (Holm $p = .083$, $.140$, $.230$ in that family), so the corrected mechanism evidence is a ranking result. The shuffle comparison supports pairing each record with its own state rather than adding

Table 3: **Published-method comparison**, complete-feature interface, 22 tasks, five seeds; macro-F1 in the supplement. All 21 published references, receiver-side controls above them. Bold marks the best value, * a corrected SHIFT-Q win in the joint 63-test Holm family: 18 of 21, but not the two strongest, where Holm $p = 1$ in that same family. The score-only student clears every reference too, so this is evidence for the one-message transfer, not aligned state.

Method	AUC	Acc	NLL ↓	Method	AUC	Acc	NLL ↓
SHIFT-Raw	**.8473**	**.8569**	**.3286**	FedProx*	.8230	.7500	.4971
SHIFT-Q	.8452	.8546	.3306	FedCGS*	.8230	.7323	.5429
Score-only student	.8448	.8527	.3311	FedXgbBagging (Flower)*	.8227	.8155	.4277
Prior-aligned teacher	.8454	.8406	.3590	MOON*	.8218	.7479	.4949
Uncalibrated teacher	.8454	.7440	.5395	FedBEns-Kron*	.8205	.8214	.3949
Local only	.8019	.8347	.4261	FedXgbCyclic (Flower)*	.8205	.8396	.4046
HFedXGBoost / FedXGBllr	.8433	.7193	.4598	FedDF*	.8116	.8197	.3886
FedTree	.8424	.7747	.4710	GH-OFL-QDA*	.8094	.7043	2.8716
FedFisher-Diag	.8283	.7484	.4961	GH-OFL-NB*	.8073	.7454	3.1310
FedRFE-RP*	.8282	.7541	.5866	FedCGS-Personalized*	.7991	.8357	.3653
FedNova*	.8268	.8246	.3861	SCAFFOLD*	.7746	.7041	.5282
GH-OFL-ProtoHyper*	.8240	.7609	.5265	FENS*	.7270	.7172	.6034
GH-OFL-LDA*	.8240	.7481	.5511	GH-OFL-FisherMix*	.6389	.6196	.6903
FedAvg*	.8232	.7499	.4969				

an arbitrary feature block. One objection is that the score-only student is starved of inputs. A pre-registered control answers it by giving *both* students every scalar the bag computes: those scalars are worth $-.0001$ on their own, so they are not the missing ingredient, and the state gain is undiminished at $+.0061$ AUC (19/22, Holm $p = .0005$ in that study's three-test family; Appendix G). The AUC difference from Leaf-SVD was unresolved ($+.0001$, raw $p = .75$, Holm $p = 1$), as was every difference from the supervised leaf target, which reads receiver labels and so is corrected with the prior-aligned teacher in the 8-contrast strong-control family (Holm $p = 1$, Table 6). These controls support record-aligned tree state without establishing a unique advantage for TreeSHAP over other aligned encodings.

The gains persisted under alternative adaptation choices. Validation-NLL blend selection reproduced the result for both SHIFT variants in a post-hoc re-analysis that was not part of any pre-declared family (Appendix H). SHIFT-Q's gains also persisted when either the teacher or receiver was replaced with LightGBM. Both variants improved AUC and NLL at 25%, 50% and 100% of the receiver label pool (Appendix H). The 100% cell is the primary comparison's own rows re-corrected inside the smaller label-budget family, not an independent replication of it; only the reduced budgets add new evidence. The state gains reported here were measured against matched adaptation arms sharing the teacher, anchor and blend, and are not asserted as a uniform improvement over teacher recalibration.

What the state carried. A reference student reading the withheld block X_B directly, under the same teacher, learner, anchor and blend, gained $+.0062$ AUC over the score-only student against $+.0058$ for SHIFT-Q and $+.0011$ for a matched-width random projection: state reaches the same order as direct access to the block it describes, and the projection does not. The $-.0004$ difference from direct access is unresolved, which is not equivalence. Both numbers describe one learner on these tasks and neither bounds what X_B contains (Appendix I).

At the complete-feature interface, where every arm reads all features (Table 3), SHIFT-Q wins 18 of 21 published federated and one-shot systems after correction but not the two strongest. The score-only student clears every reference *there* too, which is why that table speaks for the one-message transfer and not for the state channel: given complete access it needs no state to match them.

5.2 Does restricted access change what state adds?

The restricted interface is the setting the paper is about, but the same representations can be given to a student that reads every raw column. On a separate complete-access protocol SHIFT-Q's AUC difference from the matched score-only student was unresolved ($+.0004$, raw $p = .388$), while the raw attribution channel exceeded the same student by +.0025 AUC (16/22, raw $p = .013$), a secondary contrast in no declared correction family. Read across protocols that pair measures nothing about restriction. We test the interaction inside one protocol, specified before the runner was written. For each dataset and seed we hold the teacher bag, the splits, the anchor, the blend and the student family fixed and vary only the student's raw columns and whether it receives state.

The teacher executes on complete records throughout, so the state channel is identical across access levels and only the student's entitlement changes.

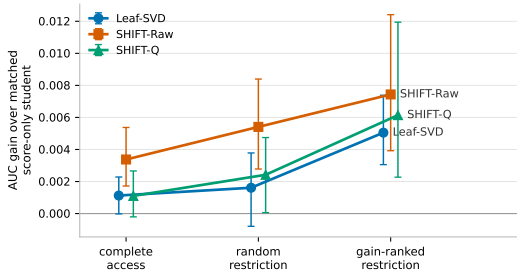


Figure 2: **State pays when the student cannot read the block itself.** Gain over a matched score-only student, one protocol per point; only the student's raw columns change. Bars are dataset-bootstrap 95% intervals over 22 tasks. Detail in §5.2.

A student reading every raw column gains $+.0004$ AUC from state; under the gain-ranked restriction the same addition is worth $+.0058$. The paired interaction is $+.0054$ (19/22 datasets, Holm $p = <.001$ over the two policies, the declared family here), and $+.0018$ (15/22, Holm $p = .010$ in that same two-test family) when a uniformly random half is withheld instead. We predicted that ordering: a random block withholds less of what the teacher relies on. Equation 1 says why restriction *creates room* for state to carry label-relevant information, not that it guarantees a finite learner will exploit it; whether a given channel does is the empirical question these cells answer (Table 12).

Every channel, at every access level. That experiment ran one channel against a single-margin scalar interface. A second, also specified in advance, gives every arm every bag scalar and runs all three channels at all three levels (Figure 2, Table 11). Under the gain-ranked restriction each channel improves on that richer arm, Leaf-SVD by $+.0051$ AUC on 22 of 22 datasets, SHIFT-Raw by $+.0074$ on 21 of 22 and SHIFT-Q by $+.0061$ on 19 of 22, each interaction resolving at Holm $p \leq .002$ in that experiment's six-test family. No random-restriction interaction resolves in that same six-test family ($p = .083$ to $.100$), though the single-margin protocol does, so the effect tracks whether the withheld block matters rather than restriction as such. The channels are not interchangeable. SHIFT-Raw has the highest mean AUC at every access level, ahead of Leaf-SVD by $+.0023$, $+.0024$ and $+.0038$ under complete, gain-ranked and random access. The interval excludes zero for the first and third ($[+.0010, +.0036]$, $[+.0007, +.0082]$) and includes it for the second ($[-.0003, +.0059]$); SHIFT-Q and Leaf-SVD are not separated anywhere.

5.3 Replication and physical sensor interfaces

On the 12-task extension, SHIFT-Q improved AUC by .0151 over the score-only student (10/12) and by .0162 over neighborhood-shuffled SHAP (11/12). Both gains passed the three-test Holm correction (adjusted $p = .0098$ and $.0029$), while the width-matched random-leaf comparison was unresolved. The extension's broader pre-registered criterion also required the best mean rank. That criterion was not met because Leaf-SVD ranked first. The separate pre-registered 20-dataset reference test did not confirm a resolved state gain (Table 17).

DSADS tested a physical interface: one sensor unit supplied raw features to the student while the teacher used all five. SHIFT-Q reduced NLL by .0101 relative to the score-only student (7/8 subjects) and by .0103 relative to the shuffle (8/8). Both contrasts passed the study's two-test Holm correction ($p = .039$ and $.016$). A second sensor study, GasDrift, did not confirm the score-residualized channel: the teacher is much the weaker predictor there, so little teacher information remains to transfer (Appendix H; all arms in Tables 15 and 16).

5.4 Naturally specified interfaces

The restrictions above are constructed. The question a practitioner asks is whether state pays when the boundary is one somebody wrote down. We assembled 7 tasks whose permitted and withheld

groups follow documented field semantics and nothing else, for instance medical specialties splitting administrative from clinical fields and cabin classes splitting what is known at booking from what is known only after the flight; the appendix lists all seven. Partitions were fixed from documentation before any of these cells ran. The teacher still reads the withheld fields at inference, so these are component-level permissions at one decision time, not claims about predicting earlier: the airline split is a post-flight decision, not a booking-time forecast.

Both candidate sets share the label budget, model families, calibration opportunity and blend grid, and both select by validation NLL on rows disjoint from fitting. The scalar-only set of seven holds three teacher calibrations, so the receiver may decline to adapt at all, plus a permitted-feature student on every bag scalar; the all-state set adds the three channels at the same blends. Admitting state improved the selected decision on every task: $+.0128$ NLL at the task level (7/7, CI $[+.0046, +.0219]$, exact $p = .016$, a single uncorrected test), which is the smallest two-sided p seven paired units permit, and $+.0113$ across the 24 contexts individually. We take the task as the unit because contexts inside a task share a dataset and a source population. The scalar-only set used its teacher-only option in 35 of 120 cells, so it was choosing freely (Table 10).

The gain is heterogeneous and tracks how much the withheld block matters, not the interface: $+.0307$ where post-flight service ratings are withheld and $+.0280$ where offer context is, against $+.0009$ on medical specialties and $+.0004$ on online shoppers. Four of 24 contexts degraded, by at most .0013.

5.5 Generalization and limitations

A direct probe of the compression argument failed: within narrow teacher-score bins, a learner given only the bin and the state did worse than the bin's own label rate on 20 of 22 tasks. Stripped of permitted features and given few rows per bin, it measures what that learner can exploit rather than what the state holds, but the test was specified in advance and is reported as it came out.

Restricted access is not distribution shift: two complete-feature sweeps did not support their original hypothesis that state gains grow with severity (Appendix J), and an earlier four-rule masking study reached the same qualitative result the access-by-encoding experiment now establishes under one protocol. The naturally specified result of §5.4 is about what a receiver *selects* when state candidates are admitted. A single fixed state arm is a weaker claim and we do not establish it there: on the three tasks of that suite that we tested this way, a fixed channel gained $+.0045$ NLL over the score-only student (8/12, raw $p = .147$) and the study's gate did not pass. Admitting state to a candidate set is the operation these experiments support; imposing one channel in advance is not.

6 Discussion

Receiver-side cost and state exposure. Extraction is paid locally at both adaptation and inference, and what it costs is a property of the implementation rather than of the interface: GPU-parallel exact TreeSHAP (Mitchell et al., 2022) computes the same values far faster than the CPU implementation we timed, and fitting the teacher, the leaf reduction and the student is paid once per receiver, not per record. Per-stage extraction costs are in Appendix E. Our evidence supports the aligned-state interface, not a particular encoder of it. Leaf-SVD is not separated from SHIFT-Q in AUC here ($+.0001$, Holm $p = 1$), which is not equivalence, and extracts 20 to 42 times more cheaply per record on the two tasks we timed. That ratio is about extraction in batch and does not survive end-to-end (Table 9): Leaf-SVD leads on 1000-record throughput yet is the slowest of the three on a single record, its one-hot and SVD transforms paying a fixed per-call cost the attribution channels avoid. Prefer the leaf channel for batch scoring, not for online single requests.

What contract authorises this. The entitlement attaches to the audited component, not to the information: a shipped teacher may read fields a receiver's own new model may not be trained on, because only the teacher passed the review authorising it. That is provenance, not privacy, and admitting $R(x)$ authorises derived influence from the withheld block, which ridge probes recover at pooled median $R^2 = .710$ against .120 from permitted features and score (Shokri et al., 2021; Luo et al., 2022). Appendix K works a clinical instance.

STATEMENTS

Appendix and supplementary material. The appendix is self-contained: it carries every table the main text cites. A further supplement, not required to read this paper, retains the per-protocol results and the recorded development and confirmation outcomes.

AI use. The research idea, the questions this paper asks, the study design and every claim it makes are the authors'. Generative AI tools assisted with drafting and revising sections of the text. The authors verified every reported number against source code and raw results, and take full responsibility for the paper.

Ethics. All experiments used public benchmark data and collected no human-subject data. The state channel is treated as a governed component entitlement; environments that permit derived influence from restricted features can expose rich state, and deployments requiring formal privacy need a separately analyzed mechanism.

Reproducibility. The supplementary archive `shift_supplement.zip`, submitted with this paper, carries the analysis and table-generation code, twelve specification documents (nine pre-registrations and three frozen sensor protocols), and the per-seed result tables from which every reported number is computed, including the canonical value file behind the representation benchmark. Its `README_SUPPLEMENT.md` maps each table to its generator and its result file; running a generator from the archive root rewrites that table without re-running any experiment, and five of them were checked to reproduce output byte-identical to the tables in this PDF. The exception is the cost table, which measures wall-clock time and therefore reports the timings of whichever machine runs it. The per-seed CSV metrics are the record. `PREREGISTRATION_NOTE.md` states what the pre-registration trail does and does not establish, including the eight files whose drafting-time digests no longer reproduce against the shipped text. We have not back-filled those digests, because a record rewritten to verify would mean nothing. What the repository history does establish is order: for seven of the eight studies whose pre-registration and runner are separate commits, the pre-registration was committed first, by 11 to 35 minutes in six cases and by more than two days in the seventh. The exception is the score-collision study, whose pre-registration and runner share a single commit, so priority is not established there; its prediction was not confirmed and is reported as a failure. Readers who need the confirmatory labels to carry full weight should check that history rather than the digests. The full repository, with the experiment runners and the dataset preparation code, will be released publicly on acceptance.

REFERENCES

Gustavo Aguilar, Yuan Ling, Yu Zhang, Benjamin Yao, Xing Fan, and Chenlei Guo. Knowledge distillation from internal representations. In *AAAI*, 2020. arXiv:1910.03723.

Youssef Allouah, Akash Dhasade, Rachid Guerraoui, Nirupam Gupta, Anne-Marie Kermarrec, Rafael Pinot, Rafael Pires, and Rishi Sharma. Revisiting ensembling in one-shot federated learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7ea46207ec9bda974b140fe11d8dd727-Abstract-Conference.html.

Kerem Altun, Billur Barshan, and Orkun Tunçel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43(10):3605–3620, 2010. doi: 10.1016/j.patcog.2010.04.019.

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*, pp. 785–794, 2016.

Sayantan Chowdhury, Ben Liang, Ali Tizghadam, and Ilijc Albanese. Improving knowledge distillation with teacher's explanation. *arXiv preprint arXiv:2310.02572*, 2023.

Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pp. 193–200, 2007.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. Tabarena: A living benchmark for machine learning on tabular data. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Ji Feng and Zhi-Hua Zhou. AutoEncoder by forest. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. Going beyond XAI: A systematic survey for explanation-guided learning. *ACM Computing Surveys*, 56(7):1–39, 2024.

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary C. Lipton. A unified view of label shift estimation. In *NeurIPS*, 2020. arXiv:2003.07554.

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

Zenghao Guan, Yucan Zhou, and Xiaoyan Gu. Capture global feature statistics for one-shot federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16942–16950, 2025. doi: 10.1609/aaai.v39i16.33862. URL https://ojs.aaai.org/index.php/AAAI/article/view/33862.

Neel Guha, Ameet Talwalkar, and Virginia Smith. One-shot federated learning. In *NeurIPS Workshop on Machine Learning on the Phone and other Consumer Devices*, 2019.

Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, and Joaquin Quiñonero Candela. Practical lessons from predicting clicks on ads at Facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, pp. 1–9, 2014. doi: 10.1145/2648584.2648589.

Guan Huang and Tao Shu. Federated oriented learning: A practical one-shot personalized federated learning framework. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25762–25790, 2025. URL https://proceedings.mlr.press/v267/huang25ae.html.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable AI: A causal problem. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2907–2916, 2020.

Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. FedFisher: Leveraging Fisher information for one-shot federated learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1612–1620, 2024. URL https://proceedings.mlr.press/v238/jhunjhunwala24a.html.

Qinbin Li, Yiqun Cai, Yuxuan Han, Ching Man Yung, Tianyuan Fu, and Bingsheng He. FedTree: A federated learning system for trees. In *Proceedings of Machine Learning and Systems (MLSys)*, 2023.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6028–6039, 2020. URL https://proceedings.mlr.press/v119/liang20a.html.

Zachary C. Lipton, Yu-Xiang Wang, and Alex Smola. Detecting and correcting for label shift with black box predictors. In *ICML*, 2018. arXiv:1802.03916.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016.

Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2:56–67, 2020. doi: 10.1038/s42256-019-0138-9.

Xinjian Luo, Yangfan Jiang, and Xiaokui Xiao. Feature inference attack on shapley values. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2022. doi: 10.1145/3548606.3560573.

Chenyang Ma, Xinchi Qiu, Daniel J. Beutel, and Nicholas D. Lane. Gradient-less federated gradient boosting trees with learnable learning rates. *arXiv preprint arXiv:2304.07537*, 2023.

Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad, Benjamin Feuer, Chinmay Hegde, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? In *Advances in Neural Information Processing Systems*, 2023.

Rory Mitchell, Eibe Frank, and Geoffrey Holmes. GPUTreeShap: massively parallel exact calculation of SHAP scores for tree ensembles. *PeerJ Computer Science*, 8:e880, 2022. doi: 10.7717/peerj-cs.880.

Frank Moosmann, Bill Triggs, and Frederic Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

Amin Parchami-Araghi, Moritz Böhle, Sukrut Rao, and Bernt Schiele. Good teachers explain: Explanation-enhanced knowledge distillation. In *European Conference on Computer Vision (ECCV)*, 2024.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14(1):21–41, 2002. doi: 10.1162/089976602753284446.

Noam Segev, Maayan Harel, Shie Mannor, Koby Crammer, and Ran El-Yaniv. Learn on source, refine on target: A model transfer learning framework with random forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1811–1824, 2017.

Cencheng Shen, Yuexiao Dong, and Carey E. Priebe. Decision tree embedding by leaf-means, 2025.

Reza Shokri, Martin Strobel, and Yair Zick. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 231–241, 2021. doi: 10.1145/3461702.3462533.

Suraj Srinivas and Francois Fleuret. Knowledge transfer with jacobian matching. In *ICML*, 2018. arXiv:1803.00443.

Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 9269–9278, 2020.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014. doi: 10.1145/2641190.2641198.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009.

Celine Vens and Fabrizio Costa. Random forest based feature induction. In *IEEE International Conference on Data Mining*, pp. 744–753, 2011.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

Zhaomin Wu, Junyi Hou, and Bingsheng He. Vertibench: Advancing feature distribution diversity in vertical federated learning benchmarks. In *International Conference on Learning Representations*, 2024.

Luyuan Yang, Shayan Shafaei, Yiming Liu, Naeem Shahabi Sani, Yu Cai, Jun Huan, and Chao Lan. One-shot federated learning based on random feature extractor. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 7737–7750, 2026. URL https://proceedings.mlr.press/v337/yang26a.html.

A Notation and reading conventions

A contrast is *resolved* when the two-sided Wilcoxon test stated for that experiment separates the two arms at .05, after whatever multiplicity adjustment that experiment declares, and *unresolved* when it does not. Each table says whether its test is exact or a normal approximation; the approximation is used where zero differences or ties make the exact test unavailable. Unresolved means no separation was established, not that the arms are equal. A figure written $16/22$ counts the inference units, meaning datasets, subjects or receivers as each table states, on which the first arm named is the better of the pair, out of all units compared. Higher is better for AUC, accuracy, macro-F1 and macro-AUPRC. Lower is better for NLL, Brier score and calibration error, so a negative NLL difference favours the first arm. Every p is raw unless a correction family is named. *Teacher* is the shipped source model, *student* the receiver's own model, and *aligned state* the teacher's per-record internal read-out.

B Access contract

Table 4 states, for each asset and stage, which party may read it under each interface, and which single item crosses the silo boundary. The restricted interface is an *output-access* constraint: the teacher still reads X_B at training and at inference, so this is not missing-data imputation and it does not show that a deployed system needs only the permitted block. The receiver also *retains* X_B locally throughout; what the interface restricts is which component may read it, not whether it exists. Enforcement is assumed to come from the receiver's own governance rather than from the algorithm: SHIFT specifies what a compliant student consumes, it does not police the boundary, and the rules we instantiated were experimental rather than transcriptions of a deployed policy.

A worked instance. A vendor ships a readmission model trained on full encounters; a department adapts it to its own population. Every field exists at the receiver at decision time, and the shipped teacher may read all of them because that artifact passed the review authorising it. What the department may not do is train a *new* model on the clinical block, which has passed no such review. The permission therefore tracks which component was audited, not which information is involved, and a deployment admitting state is authorising derived influence from the withheld fields with its eyes open.

Table 4: **Access contract.** What each party may read at each stage, under the two interfaces. The only item that crosses the silo boundary is the artifact itself, once. Scope and enforcement are stated in the text above.

Asset	Complete-feature interface	Restricted interface	Crosses silos?
Source training records	source only	source only	no
Teacher artifact $\mathcal{M}$	shipped once, immutable	shipped once, immutable	**yes**, one message
Complete receiver records	held by receiver	held by receiver	no
raw X_A	teacher and student	teacher and student	no
raw X_B	*empty by definition*	**teacher only**	no
Teacher execution $\mathcal{M}(X_A, X_B)$	receiver-local	receiver-local	no
Teacher margin m	student may read	student may read	no
Aligned state ϕ	student may read	student may read	no
Student fitting and inference	$[X_A, m, \phi]$	$[X_A, m, \phi]$	no
Receiver labels	fit, validation, selection	fit, validation, selection	no
Reconstruction probes on X_B	evaluation only	evaluation only	no

C Benchmarks and comparators

the supplement names every benchmark, the data and interface it uses, and the baselines it compares SHIFT against.

Published-method comparison. The references of Table 3 ran in their published configuration; no calibration or prior correction from this paper was applied to any of them. Accuracy and macro-F1 for every row are in the supplement.

D THE GUARDED ADAPTATION PIPELINE

The pipeline decides *whether* the receiver should use a student at all, and if so how much weight to give it. It is called guarded because every step can decline: when no student earns its place the output is exactly the recalibrated teacher, so the pipeline can never do worse than that fallback by its own selection criteria. Two disjoint sets of receiver rows do the work. The ***fit*** rows train each candidate and, through out-of-fold predictions (each row scored by a model that did not see it), also serve as the *selection* set on which candidates are compared. The separate *validation* rows are used only for the final guards. Test rows are never consulted.

1. **Anchor.** The *anchor* is the shipped teacher's probability after prevalence alignment on the receiver's fit rows. The map is an empirical prevalence-alignment heuristic in the spirit of label-shift correction (Lipton et al., 2018), not that correction itself: for teacher probability $p(y \mid x)$,
$$p_{\text{align}}(y \mid x) \;\propto\; p(y \mid x)\,\frac{\hat{\pi}_R(y)+\epsilon}{\hat{\mu}_T(y)+\epsilon}, \qquad \epsilon = 10^{-6}, \tag{4}$$
renormalised over classes per record. Here $\hat{\pi}_R$ is the receiver class prior estimated from the *fit-row labels* and $\hat{\mu}_T$ is the teacher's mean probability over the *same fit rows*; neither uses validation or test rows. The standard correction divides by the *source* prior π_S. We divide instead by the teacher's mean prediction on receiver *inputs*. That expectation is taken under the receiver's covariate distribution, so it does not equal π_S in general, and a well specified teacher does not make it so. We therefore state the map as implemented, an empirical prevalence-alignment heuristic, and claim none of the guarantees of source-prior correction. The matched state-versus-score contrasts are unaffected either way, since every arm shares this anchor. The map is applied unchanged thereafter. Being a per-class rescaling it moves argmax decisions toward $\hat{\pi}_R$ and leaves binary ranking unchanged up to ties. Alignment is skipped when the two already agree, $\|\hat{\pi}_R - \hat{\mu}_T\|_1 \le .05$. The anchor is the fallback output.
2. **Candidate students.** Fit a class-balanced XGBoost student on the permitted features, the teacher score and the state, $[x_A, m, R]$, in three modes: *residual*, which predicts what the anchor's logit gets wrong, *independent*, which predicts the label directly. And *local-only*, which ignores the teacher, where the interface permits it.
3. **Sanity gate.** A mode may compete only if the selection set holds at least 300 rows and the student's selection AUC reaches at least 0.90 times the anchor's, so a student far worse than the teacher is never considered.
4. **Weight cap.** The blend weight w is the share of the final prediction taken from the student. Its ceiling $w_{\max}$ falls with how little the student's validation AUC exceeds the anchor's, and with how few receiver rows there are, so a weak or small-sample student cannot receive a large weight.
5. **Guarded selection.** Over a fixed grid $0 < w \le w_{\max}$, a pair (mode, w) is adopted only if its blended selection AUC beats the current best by more than 8×10^{-4} (3×10^{-3} for local-only), or ties within that tolerance while validation accuracy improves by more than 5×10^{-3}. If nothing qualifies, $w = 0$ and the output is exactly the anchor.
6. **Refinements, each with a no-regression guard.** A per-class blend weight replaces the single w only if validation AUC and accuracy both improve. A per-class macro-F1 reweighting is adopted only with at least 400 validation rows and a validation macro-F1 gain above .01.
7. **Second-pass alignment.** When $w > 0$ the blended output is prevalence-aligned once more on the fit rows, because blending disturbs the calibration the anchor already had. When $w = 0$ the pass is skipped.

E ALIGNED-STATE CONTROLS AND FEATURE ACCESS

Table 12 reports the matched access-by-state interaction of Section 5.2, whose four arms share one teacher bag, one anchor and one set of splits per cell. Tables 5 and 6 report the complete 16-contrast mechanism and 8-contrast strong-control families, with correction applied separately to each family.

An earlier study compared four feature-withholding rules: mean state gains were positive under each, without resolved differences from gain-ranked withholding. The access-by-encoding experiment supersedes it by testing two policies within one protocol, and its table is in the artifact. The per-protocol teacher/state contrasts, the encoding summary and the regime-indicator summary are in the supplement and the released artifact; they are a selected descriptive collection rather than a declared comparison family, so this paper does not present their tally as evidence. Table 7 reports validation selection over three nested candidate sets, each containing three teacher calibrations. the supplement reports channel-extraction and transformation time, excluding student prediction and blending.

The nested selection audit. Each candidate set shares its label budget, validation split, teacher bag and blend grid, and all three teacher calibrations are candidates in every set. Panel A is the audit as originally run, in which only the teacher candidates had a calibration fitted on the receiver fit rows while the blended student candidates were never recalibrated. Panel B gives every candidate one affine map fitted on the fit rows, using 5-fold cross-fitted student predictions so that a student's calibrator is not fitted on its own in-sample output. The panels differ only in calibration opportunity, so Panel B is the equal-opportunity comparison. Paired dataset-level contrasts between the Panel B sets use a two-sided Wilcoxon with dataset-bootstrap 95% intervals, exact where the differences admit it and a normal approximation otherwise: scalar+leaf − scalar-only $-.0018$ $[-.0029, -.0008]$ (17/22, Holm $p = .009$ over the three nested contrasts); all-state − scalar+leaf $-.0004$ $[-.0013, +.0005]$ (9/22, Holm $p = .601$ over those three nested contrasts); all-state − scalar-only $-.0022$ $[-.0036, -.0008]$ (17/22, Holm $p = .015$ over those three nested contrasts). This paired analysis is exploratory; the earlier report compared these sets by their means alone.

The four label-free controls of Table 5 are the score-only student, neighborhood-shuffled SHAP, Leaf-SVD state and random-projected leaf state, each on AUC, NLL, tuned macro-F1 and tuned balanced accuracy, with NLL signed so positive favours SHIFT-Q. Tests were exact except for Leaf-SVD on tuned balanced accuracy, where a zero difference triggered the normal approximation. The supervised leaf target reads receiver labels, so it is corrected in the strong-control family of Table 6 instead.

Table 5: **Representation benchmark.** The declared 16-contrast mechanism family in full: SHIFT-Q minus each of four label-free state controls on four metrics (22 tasks, five seeds, two-sided Wilcoxon, Holm over all 16). Positive favours SHIFT-Q. Controls and tests are described in the text above. 4 of the 16 contrasts pass at Holm $p < .05$; the Holm column gives each.

Control	Metric	Δ	wins/22	raw p	Holm p
score-only student	AUC	$+0.0058$	18/22	0.003	0.033
score-only student	NLL	$+0.0043$	17/22	0.008	0.083
score-only student	macro-F1 (tuned)	$+0.0054$	15/22	0.017	0.140
score-only student	bal. acc. (tuned)	$+0.0061$	15/22	0.033	0.230
neighborhood-shuffled SHAP	AUC	$+0.0043$	18/22	0.001	0.018
neighborhood-shuffled SHAP	NLL	$+0.0033$	17/22	0.016	0.140
neighborhood-shuffled SHAP	macro-F1 (tuned)	$+0.0063$	16/22	0.054	0.271
neighborhood-shuffled SHAP	bal. acc. (tuned)	$+0.0087$	16/22	0.007	0.082
Leaf-SVD state	AUC	$+0.0001$	13/22	0.750	1.000
Leaf-SVD state	NLL	$+0.0024$	15/22	0.036	0.230
Leaf-SVD state	macro-F1 (tuned)	-0.0010	11/22	1.000	1.000
Leaf-SVD state	bal. acc. (tuned)	-0.0003	12/22	0.986	1.000
random-projected leaf state	AUC	$+0.0047$	19/22	$< .001$	0.009
random-projected leaf state	NLL	$+0.0039$	16/22	0.007	0.080
random-projected leaf state	macro-F1 (tuned)	$+0.0047$	15/22	0.176	0.705
random-projected leaf state	bal. acc. (tuned)	$+0.0066$	18/22	$< .001$	0.012

Table 6: **Representation benchmark.** The declared 8-contrast strong-control family in full: SHIFT-Q minus the prior-aligned teacher and the supervised leaf target on the same four metrics (22 tasks, five seeds, exact two-sided Wilcoxon, Holm over all 8). Positive favours SHIFT-Q. Both controls read receiver labels, so they are corrected separately from the label-free family. 1 of the 8 contrasts pass at Holm $p < .05$; the Holm column gives each.

Control	Metric	Δ	wins/22	raw p	Holm p
prior-aligned teacher	AUC	-0.0011	9/22	0.371	1.000
prior-aligned teacher	NLL	$+0.0278$	20/22	$< .001$	$< .001$
prior-aligned teacher	macro-F1 (tuned)	-0.0030	6/22	0.092	0.642
prior-aligned teacher	bal. acc. (tuned)	$+0.0017$	10/22	0.656	1.000
supervised leaf target	AUC	$+0.0006$	12/22	0.726	1.000
supervised leaf target	NLL	-0.0005	13/22	0.750	1.000
supervised leaf target	macro-F1 (tuned)	-0.0000	8/22	0.443	1.000
supervised leaf target	bal. acc. (tuned)	$+0.0023$	9/22	0.799	1.000

Table 7: **Selection over nested candidate sets, without and with calibration parity.** Each set adds representations to the one above; selection is by lowest validation NLL on rows disjoint from fitting. **Panel A** is the audit as originally run, **Panel B** gives every candidate equal calibration opportunity. Protocol and the three paired contrasts are stated in the text above.

Candidate set	candidates	test NLL	teacher chosen
Panel A: calibration for teachers only, as originally run			
scalar-only	7	.3289	26%
scalar+leaf	11	.3283	21%
all-state	19	.3273	13%
Panel B: equal calibration opportunity			
scalar-only	7	.3136	17%
scalar+leaf	11	.3118	14%
all-state	19	.3114	7%

Automatic representation selection. A receiver might hope to decide by validation which representation to use. Allowing leaf candidates alongside scalar ones lowered selected NLL on the suite where the rule was developed, by $-.0018$ (17/22, Holm $p = .009$ over three nested contrasts, Table 7), but the rule did not carry to new tasks (Table 8).

Prospective test of the selection procedure. Table 7 is retrospective: the candidate sets and the selection rule were developed on the same 22 tasks it reports. This study asks whether that procedure helps a receiver on tasks it has never seen. The procedure was frozen and pre-registered, then run unchanged on the 20-dataset fresh suite, which shares no dataset with the 22-task benchmark on which the procedure was built. That suite is the same one used for the reference test of Table 17, so these two results are not independent collections of new data: they are two questions asked of one extra suite. Neither is a set of datasets untouched everywhere else in this project. The implementation calls the same functions as the original audit, so the procedure is identical by construction rather than by description. The primary contrast did not replicate, and that outcome is reported in §6 rather than confined here. The panels are those of Table 7: Panel A repeats the original calibration asymmetry, Panel B gives equal calibration opportunity. Paired dataset-level contrasts on Panel B use a two-sided Wilcoxon with dataset-bootstrap 95% intervals, the draws and seed fixed in the pre-registration, exact where the differences admit it and a normal approximation otherwise: scalar+leaf $-$ scalar-only $+.0000$ $[-.0008, +.0009]$ (11/20, Holm $p = .999$ over the three prospective contrasts); all-state $-$ scalar+leaf $-.0003$ $[-.0013, +.0007]$ (11/20, Holm $p = .798$ over those three prospective contrasts); all-state $-$ scalar-only $-.0003$ $[-.0016, +.0011]$ (11/20, Holm $p = .999$ over those three prospective contrasts). Wins count datasets on which the wider set has the lower selected test NLL.

Table 8: **Prospective test of the frozen selection procedure**, pre-registered. The procedure of Table 7, taken unchanged and run on the 20 fresh datasets of Table 17, which played no part in building it. Panels as in Table 7. Protocol and the three paired contrasts are stated in the text above.

Candidate set	candidates	test NLL	AUC	teacher chosen
Panel A: calibration for teachers only, as originally run				
scalar-only	7	.1352	.9622	32%
scalar+leaf	11	.1345	.9624	28%
all-state	19	.1343	.9619	24%
Panel B: equal calibration opportunity				
scalar-only	7	.1256	.9631	27%
scalar+leaf	11	.1256	.9628	25%
all-state	19	.1253	.9620	23%

Exact CPU TreeSHAP took 0.86–23.66 seconds per 1,000 rows across 400–5,000 trees. Per 1,000 records, extracting the score alone took 3.4–6.3 ms, the leaf reduction 56.6–80.6 ms and TreeSHAP 1240.6–3066.5 ms, while the complete path a deployed receiver pays took 1231.2–3395.5 ms. In Table 9, *batch-1* is the latency an online caller waits for on a single record, *rows/s* is throughput on a 1000-record batch, *fit* is the one-off receiver cost that channel imposes including the leaf reducer where one is needed, and *peak* is peak transient Python allocation on the inference path, which is allocation rather than resident set size and is comparable across channels because each is measured the same way. Earlier cost figures in this paper are extraction ratios; these are end to end.

Table 9: **What each channel costs a receiver, end to end.** The whole inference path per channel: teacher probability, aligned anchor, extraction, student probability and blend. One machine, arm64, 2 threads, same teacher throughout. Column definitions are in the text above.

	jasmine				MiniBooNE			
Channel	batch-1	rows/s	fit (s)	peak (MiB)	batch-1	rows/s	fit (s)	peak (MiB)
score-only	0.82	121243	0.08	1.1	1.43	87751	0.43	0.4
Leaf-SVD	19.04	15050	0.40	14.0	22.33	12033	0.82	14.1
SHIFT-Raw	3.97	929	0.22	2.8	6.99	302	0.42	1.0
SHIFT-Q	2.97	1117	0.34	5.0	7.49	312	0.40	1.8

The seven splits are medical specialties separating administrative from clinical fields, cabin classes separating what is known at booking from what is known only after the flight, regions separating visitor profile from in-session behaviour, education categories separating application profile from repayment history, marital statuses separating person attributes from offer context, countries separating profile from account activity, and education levels separating profile from employment history. Diabetes130US keeps the six specialty contexts frozen in an earlier protocol; the other six tasks use the three most frequent values of their grouping column, fixed by count. In Table 10 the source for each context is every other context of its task, and the teacher still reads the withheld fields at inference. Both candidate sets share the label budget, model families, calibration opportunity and blend grid; every candidate fits one affine map on the fit rows with 5-fold cross-fitted student predictions, and selection is by validation NLL on disjoint rows. The *scalar-only* set is 7 candidates, three teacher calibrations including the teacher-only fallback plus a permitted-feature student on every bag scalar at four blends; *all-state* adds the three state channels at the same blends, 19 in all.

Table 10: **Does admitting state improve what the receiver selects?** 24 receiver contexts across 7 tasks whose permitted and withheld field groups follow documented semantics. Positive favours the state set; inference is at the task level, since contexts within a task share a dataset and a source population. Splits, candidate sets and selection are described in the text above.

Task	Receiver	scalar-only	all-state	Δ NLL
AirlineSatisfaction	Business	.1278	.0687	+.0590
AirlineSatisfaction	Eco	.1691	.1431	+.0260
AirlineSatisfaction	Eco Plus	.1418	.1347	+.0071
BankChurn	France	.3260	.3165	+.0095
BankChurn	Germany	.5050	.4596	+.0454
BankChurn	Spain	.3431	.3389	+.0042
CouponOffer	Married partner	.5772	.5475	+.0297
CouponOffer	Single	.5562	.5294	+.0268
CouponOffer	Unmarried partner	.5732	.5459	+.0273
CreditDefault	1	.4086	.4038	+.0048
CreditDefault	2	.4573	.4508	+.0065
CreditDefault	3	.4755	.4753	+.0002
Diabetes130US	Cardiology	.2701	.2695	+.0007
Diabetes130US	Emergency/Trauma	.3285	.3297	−.0012
Diabetes130US	Family/GeneralPractice	.3498	.3503	−.0004
Diabetes130US	InternalMedicine	.3354	.3349	+.0005
Diabetes130US	Nephrology	.4163	.4136	+.0027
Diabetes130US	Surgery-General	.3313	.3281	+.0033
HRJobChange	Graduate	.4888	.4798	+.0090
HRJobChange	High School	.4455	.4368	+.0087
HRJobChange	Masters	.4148	.4134	+.0013
OnlineShoppers	1	.2430	.2403	+.0027
OnlineShoppers	3	.2442	.2444	−.0002
OnlineShoppers	4	.2209	.2222	−.0013

Semantic access contracts. Every restriction used above is constructed: gain-ranked, random, or a withheld sensor block. One suite in this project instead follows documented application semantics, frozen in a written protocol before its extension tasks were run, and it is reported here because it bears on whether the benefit reaches a governance-defined interface rather than because it is favourable. It is not: in that earlier fixed-channel study the state-versus-score contrast is unresolved. The later candidate-set comparison on seven semantically split tasks, reported in §5.4, is a different experiment and does resolve; the two should not be read together. Every arm in Table 11 receives m_0, m_1, m_2 and the bag logit, so no channel is credited with information the scalars already carry. This richer scalar interface differs on purpose from the single-margin protocol of Table 2, which stands unchanged. Intervals are dataset bootstraps with the draws and seed fixed in the pre-registration, and the interquartile range shows dataset-level spread rather than the mean alone.

Table 11: **Access by encoding**, pre-registered. Every channel against its matched score-only arm at every access level, in one protocol where every arm receives every scalar the bag computes. The interaction is the state benefit under restriction minus the benefit under complete access, per dataset; exact two-sided Wilcoxon, every p Holm-corrected over this experiment's declared six-test family. Protocol detail is in the text above.

Channel	restriction	interaction	95% CI	wins	IQR	Holm p
Leaf-SVD	gain	+.0039	[+.0023, +.0057]	20/22	[+.0006, +.0063]	<.001
Leaf-SVD	random	+.0005	[−.0025, +.0028]	15/22	[−.0000, +.0031]	.100
SHIFT-Raw	gain	+.0041	[+.0016, +.0073]	20/22	[+.0005, +.0059]	.002
SHIFT-Raw	random	+.0020	[+.0007, +.0034]	14/22	[−.0003, +.0035]	.083
SHIFT-Q	gain	+.0050	[+.0014, +.0109]	20/22	[+.0003, +.0044]	<.001
SHIFT-Q	random	+.0013	[−.0003, +.0029]	17/22	[+.0000, +.0035]	.100

Each cell of Table 12 uses one protocol, one teacher bag, one anchor and one set of splits, so its four arms are matched and the interaction is measured within the cell rather than across protocols. The teacher always executes on complete records, so the state channel is identical across access levels and only the student's raw columns change. Wins count datasets with a positive interaction. The *gain* policy withholds the top half of features by teacher gain, this paper's primary rule; *random* withholds a uniformly random half, drawn independently of the split.

Table 12: **Does restricting the student's raw access change what state adds?** Pre-registered. *Complete* and *restricted* are the simple effects of adding state at each access level, uncorrected; the interaction is their per-dataset difference, exact two-sided Wilcoxon, every p Holm-corrected over this experiment's declared two-test family, one test per restriction policy. Protocol and policies are stated in the text above.

Restriction	complete	restricted	interaction	wins	Holm p
gain	+.0004	+.0058	$+.0054$	19/22	$<.001$
random	+.0004	+.0022	$+.0018$	15/22	.010

F CALIBRATION SPECIFICATION

The selection audit of Table 7 is stated here in full, so that the Panel B comparison can be read without reference to the code.

Transform. Calibration is an affine map on the *logit*, not on the probability. For a candidate probability q the calibrated output is $\sigma\left(a \log \frac{q}{1-q} + b\right)$ with $a = e^{\theta_0} > 0$ and b free, fitted by L-BFGS-B on the mean binary cross-entropy of the fit rows. Probabilities are clipped to $[10^{-7}, 1 - 10^{-7}]$ and logits to ± 35 before the map is applied. Because $a > 0$ the map is strictly increasing, so it cannot change any ranking: AUC is invariant under it and only the probabilistic endpoints move. A positive-slope affine map applied to probabilities instead is also rank-preserving; the reason we fit the logit form is that it maps into $(0, 1)$ by construction, whereas an affine map on probabilities can leave the interval and need clipping, which creates ties.

Candidates. Three teacher calibrations (raw, prevalence-aligned, temperature-scaled) applied to the same shipped artifact and bag as the representation benchmark, plus blended candidates $(1 - w)$ anchor $+ w$ student at $w \in \{.25, .5, .75, 1\}$ over four representations (score-only, Leaf-SVD, SHIFT-Raw, SHIFT-Q). The nested sets of Table 7 differ only in which representations are admitted; every set contains all three teacher calibrations.

Rows. Each candidate fits its one affine map on the receiver *fit* rows. For a teacher those rows are already out of sample. For a student they are its training data, so the blend used to fit the calibrator is built from 5-fold stratified *cross-fitted* student predictions, which is what makes the two sides comparable; without that step a student's calibrator would be fitted on its own in-sample output. Selection is by lowest NLL on the disjoint *validation* rows. Test rows are never consulted, by either the calibrator or the selector.

What NLL measures here. NLL is the quality of the predicted probabilities, which combines discrimination and calibration; it is not a calibration metric alone. A candidate can improve NLL by ranking better, by being better calibrated, or both. Panel A and Panel B differ only in calibration opportunity, so the movement between them is attributable to calibration; the differences *within* Panel B are not decomposable that way.

G RICHER SCALAR INTERFACE

Pre-registered in the released artifact before the arms were implemented; the three contrasts, the Holm family and the prediction were fixed there. The prediction was that the state gain would survive, with a point estimate within .002 of the published value; it did, at $+.0061$. The one part of the prediction that missed was its sign for the scalar-only secondary, which we expected to be small and positive and which came out small and negative. The deployed system builds its anchor from a three-member teacher bag, but the student's scalar input in the primary protocol is M_0's margin alone. Everything else in Table 13 is that primary protocol, and the two m_0 arms reproduce it exactly, so the four arms are matched. The state channel is unchanged between the SHIFT-Q arms: $\phi^{\perp}$ is residualized on $[1, m_0]$ and deliberately not re-residualized against the added scalars. The three contrasts were fixed in the pre-registration before the run and are reported here whichever way they landed.

Table 13: **Richer scalar interface**, pre-registered. The *all bag scalars* arms additionally receive the other two bag members' margins and the bag logit, quantities the deployed system already computes. Unit is the dataset, means over five seeds; exact two-sided Wilcoxon, Holm over the three pre-registered contrasts. Protocol detail is in the text above.

Student scalar interface	AUC	NLL ↓
Score-only student (m_0)	.8385	.3355
Score-only student, all bag scalars	.8384	.3371
SHIFT-Q (m_0)	.8443	.3312
SHIFT-Q, all bag scalars	.8446	.3323

Contrast (AUC)	Δ	wins	raw p	Holm p
SHIFT-Q (all scalars) − Score-only (all scalars)	$+.0061$	19/22	.0002	.0005
Score-only (all scalars) − Score-only (m_0)	$-.0001$	11/22	.7502	.9652
SHIFT-Q (all scalars) − SHIFT-Q (m_0)	$+.0003$	12/22	.4826	.9652

H ROBUSTNESS CHECKS

The checks behind the main text's report that the gains persisted under alternative adaptation choices, each on the same 22 tasks and five seeds as the representation benchmark: the label-budget analysis (Table 14) and the blend sweep, the conditional split and blend-policy re-analysis that follow, and the LightGBM teacher and receiver swaps (the supplement), which test whether the effect is specific to one gradient-boosting library.

Table 14: Receiver-label fraction audit on the same 22 tasks and five seeds. The receiver test split remained fixed; only the labeled fit/validation pool was reduced by nested class-stratified sampling. Positive Δ favors SHIFT-Raw. Dataset-level two-sided Wilcoxon tests were Holm-corrected jointly over the six displayed AUC/NLL contrasts.

		AUC			NLL ↓		
Labels	Median n	Margin	SHIFT-Raw	Δ	Margin	SHIFT-Raw	Δ
25%	343.5	0.8346	0.8396	+0.0050	0.3419	0.3366	+0.0053
50%	687	0.8417	0.8450	+0.0033	0.3357	0.3304	+0.0053
100%	1374	0.8385	0.8460	+0.0075	0.3355	0.3293	+0.0061

Where the gain sits, and where that pattern breaks. A natural conjecture is that state pays only while the shipped teacher still out-predicts what the receiver can build alone. Splitting each suite by whether the matched score-only student out-discriminates the prior-aligned teacher tests it. We rank by AUC rather than NLL, because NLL mixes calibration with discrimination and a teacher can be well ordered yet badly calibrated.

On the primary 22 tasks the split is sharp. Where the teacher still leads (19 tasks) state adds $+.0071$ AUC over the score-only student (18/19); where the student already leads (3 tasks) the mean change is $-.0030$ and none improved (0/3). GasDrift points the same way: on its 8 batches where the student leads, the mean NLL change is $-.0008$ (4/8).

The conjecture does not survive DSADS. There state helps on both sides of the split, $+.0097$ NLL on the 4 subjects where the teacher leads (4/4) and $+.0104$ on the 4 where the student does (3/4). So this is not an operational rule for deciding when to extract state. It describes where the measured gain sits on two of the three suites and fails on the third, the subgroups are small and post-hoc, and we draw no selection guidance from it.

Why GasDrift did not confirm. The non-confirmation is consistent with the split above rather than unexplained. On GasDrift a student reading only the permitted features already reaches .9966 macro-AUC, while the shipped prior-aligned teacher reaches .9551 at .1007 expected calibration error and .5463 NLL. The teacher is therefore the weaker predictor by a wide margin, and has little to tell a receiver whose own features already separate the classes. SHIFT-Q duly ties the score-only student there ($-.0005$ NLL, 18/36 cells), which matches the split above on the primary suite. DSADS shows that this is a description of these two suites and not a rule, since state helps there

on both sides of the same split. The leaf channel still helped there ($-.0089$ NLL, 26/36), consistent with the absence of an established attribution advantage elsewhere in this paper. This reading is descriptive: it was formed after seeing the result and was not tested.

Does the fixed blend carry the result? The primary study fixes $\lambda = \frac{1}{2}$ so that representations are compared under equal reliance on the anchor. Choosing λ per dataset, seed and arm by validation NLL instead moves it off $\frac{1}{2}$ in 69% of cells, and the state-versus-score conclusion survives that change: SHIFT-Q gains $+.0036$ AUC over the matched score-only student (19/22, Holm $p = .0009$ in the two-test family below) and SHIFT-Raw gains $+.0054$ (21/22, Holm $p < .0001$ in that family), against $+.0058$ and $+.0075$ at the fixed weight. The gain is smaller when the weight is selected and holds on one more dataset in each case. These contrasts were not part of any family declared before the original runs; the Holm adjustment over the two of them is made post-hoc in this revision and is reported as such, not as an original declaration. This is a re-analysis of the records already in the supplement, not a new run.

I Recovery fraction, equivalence and teacher-family follow-ups

These analyses are *exploratory*. None was preregistered, and the equivalence region was chosen after the initial run; they qualify the matched direct-access comparison of Section 5.1 rather than support a claim of their own.

SHIFT-Q's $-.0004$ AUC difference from direct access is unresolved, which does not establish that the two arms are equivalent. We report the difference rather than an equivalence margin chosen after seeing results.

A frozen-MLP-teacher follow-up found higher mean AUC with four state readouts than with scores alone (all raw $p \leq .03$), but did not establish equivalence to direct feature access. Full-feature and MLP-receiver controls are in the released artifact. These follow-ups indicate that the interface is not specific to a tree teacher in the cases we tried; they do not license a claim about arbitrary teachers.

J Distribution shift and the natural-shift studies

In the unblended covariate-shift diagnostic, the state-augmented student gained +.0011 to +.0073 AUC over the score-only student across the four settings, while the teacher remained ahead. The pooled correlation between measured severity and the state gain was $\rho = +.06$. It did not establish a severity-dependent trend. In the separate concept-shift diagnostic, the state–score-only difference decreased from +.0020 to -.0082 as the target label-flip fraction increased. At the highest setting, the state student still exceeded the teacher by +.310 AUC, but the score-only student performed better. The sweeps did not support their original hypothesis that state gains would increase with severity. These results bound the tested diagnostic pipelines, not every form of shift or the blended SHIFT method. The per-severity table is in the supplement; these sweeps test unblended diagnostic pipelines, not the method this paper reports, which is why the table is not carried here.

the supplement summarizes the conditions. The covariate- and concept-shift sweeps, including the predictions each runner fixed before running and did not meet, are reported per severity in the supplement. Table 15 is the sensor study and Table 16 the sensor confirmation that did not pass. Table 17 is the pre-registered reference test whose secondary prediction did not hold. The GasDrift protocol was frozen on 2026-09-01, the same day as the sensor study of supplementary Table S19 and before any GasDrift file was read. Its frozen gate required SHIFT-Q to have the lowest mean NLL and to beat SHIFT-Raw and Leaf-SVD after correction. Batch is the inference unit, four views averaged within batch; exact paired two-sided Wilcoxon over nine batches, Holm over the two contrasts of gate item 4 only, while the supervised-leaf-target contrast belongs to gate item 3 and its p is uncorrected. The teachers are the reading: a prior-aligned teacher at .546 NLL and an uncalibrated one at .795 are far worse than a local model at .130, so the teacher's *score* carries little the receiver can use, and the dose–response account of §5.1 predicts a small state gain at best. The defensible conclusion is the narrow one: SHIFT-Q did not confirm the registered benefit under this protocol. It is not that the withheld sensors carry nothing transferable. Four aligned-state arms have lower mean NLL than the score-only student in Table 16, Leaf-SVD and SHIFT-Raw among them. No contrast resolves in

this study's declared family, so those means establish nothing by themselves; equally they do not establish an absence of transferable information. A weak teacher score does not imply uninformative teacher state, which is the distinction this paper turns on.

Table 15: Natural multiclass validation on DSADS. Means first averaged five physical sensor views within each of eight receiver subjects.

System	NLL ↓	Accuracy	Macro-F1	mAUPRC
SHIFT-Q	**.1564**	.9733	.9705	.9951
SHIFT-Raw†	.1629	**.9781**	**.9763**	.9942
Score-only student	.1665	.9714	.9683	.9949
Neighborhood-shuffled SHAP	.1666	.9711	.9677	.9955
Random-projected leaf state	.1670	.9747	.9737	**.9956**
Supervised leaf target	.1769	.9757	.9748	.9890
Leaf-SVD	.1785	.9738	.9711	.9955
Prior-aligned teacher	.2475	.9324	.9237	.9767
Uncalibrated teacher	.3409	.8922	.8708	.9729
Local only	.4125	.9082	.8933	.9523

†Matched secondary control.

Table 16: **Pre-registered confirmation that did not pass**, GasDrift: chronological shift across ten batches over 36 months, receiver restricted to one of four physical sensor types (32 of 128 columns). All $9 \times 4 \times 10 = 360$ cells completed. **It failed**: SHIFT-Q is fifth of ten on NLL and two of the three gate contrasts are resolved *against* it. The frozen gate, the tests and the diagnosis are in the text above.

System	NLL ↓	Accuracy	Macro-F1	
supervised leaf target	0.0778	0.9858	0.9605	
Leaf-SVD	0.0860	0.9799	0.9544	
SHIFT-Raw	0.0863	0.9812	0.9520	
random-projected leaf state	0.0907	0.9799	0.9526	
SHIFT-Q	0.0944	0.9812	0.9562	
score-only student	0.0949	0.9768	0.9521	
neighborhood-shuffled SHAP	0.1090	0.9721	0.9465	
local only	0.1298	0.9671	0.9418	
prior-aligned teacher	0.5463	0.8067	0.7629	
uncalibrated teacher	0.7954	0.7773	0.7274	
Gate contrasts on NLL, where positive means SHIFT-Q is WORSE				
Contrast	ΔNLL	wins/9	raw p	adjusted p
SHIFT-Q − supervised leaf target	+0.0166	1/9	0.020	uncorrected (gate 3) 0.020
SHIFT-Q − SHIFT-Raw	+0.0081	2/9	0.020	Holm, $m = 2$ 0.039
SHIFT-Q − Leaf-SVD	+0.0084	1/9	0.098	Holm, $m = 2$ 0.098

Table 17: **Reference test, pre-registered follow-up**, Twenty datasets chosen by a rule fixed in writing, with the SHA-256 of the dataset list recorded before any was loaded. Fifteen are new to this paper, and a defective name filter in the selection pass readmitted five already used in the operating-pattern benchmark (churn, mushroom, musk, jm1, eeg-eye-state), which we record rather than reselect. The pre-registered primary contrast did not resolve: SHIFT-Q − reference −0.0002 (8/20, exact raw $p = 0.68$). The reference gain itself is also unresolved (+0.0005, 11/20, raw $p = 0.207$), so no recovery fraction is defined here. Equivalence at ±.003 passes but has no discriminating power on this suite, three datasets are saturated at AUC 1.0000, the seven transfer-arm means span .0011, and the score-only student passes the same test, so we do not read it as evidence of equivalence. The Holm column corrects over the three contrasts listed here; every other p is raw.

Contrast (AUC)	Δ	wins/20	exact p	Holm p
SHIFT-Q - margin	+0.0003	11/20	0.145	0.327
SHIFT-Raw - margin	+0.0005	12/20	0.109	0.327
reference - margin	+0.0005	11/20	0.207	0.327

K STATE EXPOSURE

Table 18 reports how well held-out ridge probes recover the withheld block from each channel the receiver may read.

Table 18: Held-out recoverability of restricted features from alternative teacher-state channels: median feature R^2, then median over three seeds within each dataset. The pooled row is the median over all 12 dataset–seed values. All ridge probes also received $[X_A, m]$. Larger values mean greater recoverability.

Dataset	Score	SHIFT-Q	Leaf-SVD	Sup. leaf	Full leaf/path
Australian	−.005	.637	.469	.581	.815
Bank churn	.018	.224	.771	.673	.966
HELOC	.638	.861	.879	.822	.949
Pol	.164	.824	.644	.501	.759
Pooled median	.120	.710	.700	.652	.901

Does noise on the state buy a useful trade-off? We specified this question in advance and the answer is no. Independent Gaussian noise on each coordinate of $\phi^{\perp}$, at the smallest scale that halves reconstruction exposure, leaves the gain over a score-only student at $+.0031$ at Holm $p = .057$ in that study's six-test family, short of the threshold used throughout; that was the predicted outcome. We therefore do not offer noise as a remedy, and width compression remains the exposure lever this paper supports. This is not differential privacy: there is no sensitivity bound on the state map, no accountant and no ε. The full frontier is in the released artifact.

L Evidence across settings

Table 19 states, for each study, the comparison it ran, its endpoint and effect, the inferential status it declared at the time, and what it supports. It is organised by question rather than by runner, and it consolidates results the earlier draft reported in more than one place. The complete registry of recorded verdicts, including those on method variants this paper does not report, is retained in the supplement and the released artifact; a count of failed gates is not itself an account of the evidence, because different gates test different claims and a composite criterion can fail while a specific contrast succeeds.

Table 19: **Evidence across settings.** One row per study and comparison, organised by the question it answers. *Status* is the status each study declared at the time, with its correction family; unresolved means no difference was established, not equivalence. Parenthetical codes are the released registry's protocol identifiers.

Study	Comparison	Endpoint, effect	Status	What it supports
Does aligned state add beyond the teacher's score?				
22-task restricted (primary)	SHIFT-Q − score-only student	AUC $+.0058$, 18/22	Resolved; Holm, 16-contrast family	The paper's central claim, on the suite it was declared for
Richer scalars, same suite (P13)	both students get every bag scalar	AUC $+.0061$, 19/22	Resolved; Holm, 3-contrast family	The comparator was not merely thin; same suite, not a replication
12-task extension	SHIFT-Q − score-only student	AUC $+.0151$, 10/12	Resolved; Holm, 3-test family	Supports the contrast; the study's broader best-mean-rank criterion was not met, as Leaf-SVD ranked first
DSADS, 8 subjects	SHIFT-Q − score-only student	NLL $-.0101$, 7/8	Resolved; Holm, 2-test family	Holds on a physical sensor interface
GasDrift	SHIFT-Q among ten arms	NLL, ranked fifth	Criterion not met	Does not confirm SHIFT-Q here; other channels had lower mean NLL, which was not tested
20-task reference suite (P9)	SHIFT-Q − score-only student	AUC, see Table 17	Not resolved	No confirmation; headroom is limited, since direct access to the withheld block was also unresolved
Semantic interfaces, fixed channel	one state channel imposed in advance, 3 tasks	NLL $+.0045$, 8/12	Not resolved; raw p, no family	Imposing a channel up front is not established on these contracts

Study	Comparison	Endpoint, effect	Status	What it supports
Semantic interfaces, candidate set	admitting state candidates, 7 tasks	NLL $+.0128$, 7/7 tasks	Resolved; exact test, task level	Admitting state improves what the receiver selects under documented contracts
Complete access, separate protocol	SHIFT-Q $-$ score-only student	AUC $+.0004$	Unresolved; raw p, no family	No claim under complete raw access
Does restricting the student's access change what state adds?				
Matched 2×2 (P16), gain policy	interaction of access and state, one protocol	AUC $+.0054$, 19/22	Resolved; Holm, 2-test family	Restriction is what makes state useful here; with complete access state adds $+.0004$
Matched 2×2 (P16), random policy	the same, withholding a random half	AUC $+.0018$, 15/22	Resolved; Holm, 2-test family	Smaller, as predicted: a random block withholds less of what the teacher uses
Is the gain just extra columns, or an unaligned block?				
22-task controls	vs neighborhood-shuffled state and width-matched random projection	AUC, both favour SHIFT-Q	Holm, 16-contrast family	Row alignment and width do not account for the gain
Direct-access reference	student reading X_B directly $-$ score-only	AUC $+.0062$ vs $+.0058$	Difference $-.0004$, unresolved	State reaches the same order as direct access; not equivalence
Does the interface need an attribution encoding?				
22-task primary	Leaf-SVD $-$ SHIFT-Q	AUC	Unresolved	No attribution encoding is required; Leaf-SVD extracts 20 to 42 times cheaper per record
Does it give a transferable configuration-selection rule?				
Development suite	scalar+leaf $-$ scalar-only candidate set	Selected NLL $-.0018$	Resolved; Holm, 3-contrast family	A gain on the suite the procedure was built on
Prospective, 20 tasks (P14)	the same frozen procedure	Selected NLL $+.0000$, CI $[-.0008, +.0009]$	Not resolved; Holm, 3-contrast family	It does not transfer; we claim no cross-dataset selection improvement
Does noise on the channel buy an exposure reduction worth its cost?				
Noise frontier (P15)	state $-$ score-only at the half-exposure level	AUC $+.0031$	Not resolved; Holm, six-test family	No remedy claimed; width compression remains the lever the paper supports