AMORTIZING CROSS-EXAMPLE RESPONSES FROM LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

An example placed in context can change how a large language model (LLM) responds to other inputs. We show that this cross-example response can provide useful supervision when outcome labels are scarce. For each tabular record, we place the same record in context with each possible label and measure the resulting responses on a set of unlabeled probes. A small student learns to predict this response from the original features without outcome labels; the scarce labels enter only through the final classifier. The predictor distilled from the signed response reaches 63.61% balanced accuracy and outperforms raw logistic regression, Extra Trees, TabPFN-3.5, and TabICLv2. Matched controls show that the gain comes from the signed change induced by the hypothetical label. Altering the learned predictor while leaving its predictions on every labeled example unchanged substantially reduces held-out accuracy. The response data therefore supply predictive structure that the labeled examples alone do not determine. On a distinct low-label task with a different LLM backbone, the signed-response student improves balanced accuracy by 7.11 points over raw logistic regression and attains the highest point estimate among the tested methods, including XGBoost. In network-flow classification, useful information instead remains distributed across individual probe responses and is largely lost by scalar averaging. The LLM is used only offline to produce the response supervision; once the student is trained, new records require no LLM calls.

1 INTRODUCTION

Large language models (LLMs) can use a small number of examples placed in context to change how they answer a new query. But the effect of an example is not confined to that query: placing one labeled example in context can also change how the model responds to other inputs. We study this cross-example response as a source of supervision when outcome labels are scarce. Rather than asking the LLM to predict the outcome of a record, we measure the effect that the record induces elsewhere and learn that effect with an ordinary predictor.

For a tabular record x, we place the same record in context under each possible class label and query the LLM on a fixed set of unlabeled probes. Only the hypothetical label attached to x changes. The paired measurement lets us separate what is shared across the two hypothetical assignments from what changes with the assigned label; neither the true outcome of x nor the outcomes of the probes are needed to construct the measurement. This object is different from a direct LLM prediction, a hidden-state embedding, or a generated feature rule. It is a response induced by one record on separate unlabeled inputs under a controlled contextual intervention.

The response is acquired offline and then amortized into a small student. Records for which LLM responses have been measured train the student to predict the prescribed response from the original attributes. Task outcomes do not enter this response-regression loss. The scarce labeled examples are used only to fit the final task head. At deployment, a new record passes through the raw-feature preprocessing, the student, and the fitted head; the LLM is no longer part of inference. We use behavioral prior to refer to the response-derived restriction supplied by the student. Figure 1 summarizes this separation between offline response measurement and deployment.

This separation is particularly useful in the low-label regime. A small labeled support can leave many predictors indistinguishable on the observed examples even though they behave differently elsewhere.

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.

Response fitting can constrain this ambiguity before the outcome head is fitted. In our clearest scalar setting, the signed-response student reaches 63.61% balanced accuracy and outperforms raw logistic regression, Extra Trees, TabPFN-3.5, and TabICLv2. Matched response controls show that the useful signal is specifically associated with the signed change induced by the hypothetical label, rather than the shared response level, unsigned activity, a direct LLM margin, shuffled supervision, or fixed projections of the response.

More importantly, the response changes what can be learned from the scarce labels. We can remove or reverse a component of the response-selected predictor while preserving its predictions on every labeled support example to numerical precision. Held-out balanced accuracy nevertheless falls by 6.63 and 8.23 points. The same intervention applied to prespecified raw-random and PCA directions does not reproduce this loss; none of 256 random directions or ten PCA directions matches the corresponding behavioral effect. Thus two predictors can agree on every labeled example yet differ sharply on unseen data, and the response-fitting data distinguish between them. In this sense, the behavioral supervision constrains degrees of freedom that the labeled support leaves unresolved.

The signed response construction also transfers unchanged to a distinct low-label task with a different LLM backbone: its student improves balanced accuracy by 7.11 points over raw logistic regression and has the highest point estimate among the tested same-panel methods, including XGBoost. This experiment asks whether the response construction itself remains useful outside the setting in which it was first identified, rather than choosing a new response from target performance.

The remaining experiments test what information the student must preserve. In network-flow classification, useful response information remains distributed across individual probes: probe-resolved responses improve the prescribed LR and RBF readouts, whereas scalar summaries lose most of this advantage. Conversely, accurate response reconstruction alone is not sufficient for task utility. In another tabular setting, the student reconstructs its prescribed response with $R^2 = 0.948$, yet the downstream balanced-accuracy gain over raw logistic regression is only 0.86 points. Response fidelity and predictive value are therefore distinct quantities.

Taken together, these results suggest a different role for large language models in low-label prediction. The useful object need not be the model's prediction for the record itself. A controlled intervention on one record can expose a response on other inputs that provides supervision for a smaller learner. That response can be measured offline, distilled from raw attributes, and used to constrain the subsequent outcome-learning problem. Once this has been done, prediction on new records requires no LLM calls.

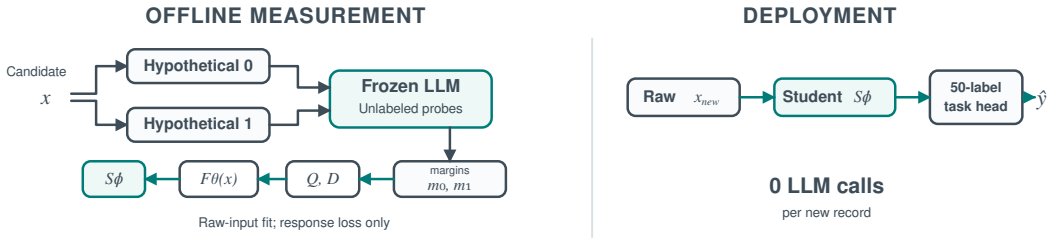


Figure 1: Offline measurement, distillation and deployment. A frozen LLM measures responses on unlabeled probes under both hypothetical candidate labels. The raw-input student learns the prescribed response without outcome labels; a separate task head uses the labeled support. Prediction on a new record requires only preprocessing, the student and the task head, with no LLM call.

2 Related Work

Demonstrations as interventions on model behavior. Work on in-context demonstrations studies what demonstrations do beyond supplying their correct labels. Min et al. (2022) and Wei et al. (2023) show that in-context examples can alter model behavior through properties that are not reducible to ordinary labeled supervision. Chang & Jia (2023) also study examples under alternative labels, using outcome-defined development margins to model demonstration effects. NUGGETS scores

examples by whether they improve the likelihood of known anchor answers and uses those scores for data selection and fine-tuning (Li et al., 2024b). Our probes require no reference answers, and the measured quantity is not used to rank or select demonstrations. The paired hypothetical labels provide a controlled intervention on the context; the resulting response on separate unlabeled probes becomes the supervision target for another learner.

Privileged supervision and amortization. Once the response has been measured, the learning problem is related to knowledge distillation and learning using privileged information. These methods transfer information available during training into a predictor that operates without it at test time (Hinton et al., 2015; Vapnik & Izmailov, 2015; Lopez-Paz et al., 2016). Label-free concept bottlenecks similarly learn an intermediate representation before fitting the task head (Oikarinen et al., 2023), and recent work emphasizes that the usefulness of privileged information depends on the signal being transferred, the student objective, and the downstream learner (Provodin et al., 2025). Our setting follows the same broad separation between training-time supervision and deployment-time inputs. Here, the privileged target is a record-induced response obtained by intervening on the label assigned in context and observing the frozen model elsewhere.

Function vectors, MEND, learned task vectors and Hyper-ICL extract, compress or transfer demonstration information within a language model and retain the language model for prediction (Todd et al., 2024; Li et al., 2024a; Yang et al., 2026; Alipour Talemi et al., 2026). We distill an offline intervention response into a predictor over the original attributes, removing the language model from inference on new records. Table 55 in Appendix J compares selected related supervision mechanisms.

Language models for tabular prediction and representation. TabLLM and TabuLa serialize rows for direct language-model prediction; embedding approaches retain the encoder for downstream inference (Hegselmann et al., 2023; Gardner et al., 2024; Kang et al., 2026; Schindler et al., 2026). FeatLLM, CAAFE, and OCTree construct reusable features, rules, or programs for conventional learners (Han et al., 2024; Hollmann et al., 2023b; Nam et al., 2024). We instead measure how a record changes the frozen model's responses to separate unlabeled inputs and learn that induced response from the original attributes. TabPFN and TabICL learn tabular priors from synthetic tasks and predict from small labeled supports without LLM inference (Hollmann et al., 2023a; 2025; Qu et al., 2025; 2026).

3 Behavioral prior distillation

3.1 The paired response and its structural family

For a binary task, take a candidate record x and a fixed bank of unlabeled probes $u_1, \ldots, u_K$. We measure the effect of x by placing x in context with hypothetical class label $c \in \{0, 1\}$ and querying the frozen model on probe u_j. For each answer mapping, we repeat the measurement, changing only the candidate's assigned label. Neither the true outcome of x nor the probe outcome enters this paired response.

We use the next-token margin between the two answer choices. To reduce A/B-mapping sensitivity, we average both A/B-to-class mappings after orienting each margin toward semantic class 1. Let $g_{c,a}(x, u_j)$ denote this margin for candidate x, hypothetical class c, and answer mapping a. We define

$$m_{cj}(x) = \frac{1}{2}\sum_{a=1}^{2} g_{c,a}(x, u_j), \qquad Q_j = \frac{m_{1j} + m_{0j}}{2}, \qquad D_j = \frac{m_{1j} - m_{0j}}{2}. \tag{1}$$

Q_j captures what is shared between the two hypothetical assignments; D_j captures the part that changes with the candidate's hypothetical label.

Exchanging the two hypothetical labels assigned to the candidate exchanges m_0 and m_1:

$$(m_0, m_1) \mapsto (m_1, m_0), \qquad Q \mapsto Q, \qquad D \mapsto -D. \tag{2}$$

We call Q exchange-symmetric and D exchange-antisymmetric. The transformation is invertible, since $m_0 = Q - D$ and $m_1 = Q + D$; no information in the paired measurement is lost. This symmetry follows from the hypothetical interventions, not classification performance. Q does not compare against a prompt without the candidate.

Retaining all K response coordinates preserves probe-specific structure; a scalar summary necessarily discards it. If no probe index is privileged, a linear scalar summary should be invariant to probe permutation. For $f(D) = w^\top D$, requiring $f(\Pi D) = f(D)$ for every permutation matrix Π forces all entries of w to be equal. Up to an overall scale, the permutation-invariant linear scalar summary is the probe average

$$\bar{D} = \frac{1}{K} \sum_{j=1}^{K} D_j.$$

The same argument gives $\bar{Q}$ for the symmetric component. This invariance concerns only the summary's form; probes need not be semantically interchangeable or drawn from an exchangeable distribution.

We write $P = 2D$. The primary response functionals for Bank, Reading and NetFlow are:

$$F_\theta(x) = \begin{cases} 2\bar{D}(x), & \text{Bank, Reading: signed antisymmetric average,} \\ 2\big(|D_1(x)|, \ldots, |D_K(x)|\big), & \text{NetFlow: magnitude profile.} \end{cases} \tag{3}$$

Bank and Reading use the same signed antisymmetric first moment, fixed before Reading target scoring. NetFlow retains probe identity but removes antisymmetric orientation through $|D|$. The constant factors are absorbed by the corresponding response scalers.

3.2 Response-only student fitting

Let $\mathcal{C} = \{x_i\}_{i=1}^N$ denote the response-bearing donor set, T the raw-feature preprocessing fitted on the permitted donors, and A a scaler fitted to the corresponding response targets. The student is obtained by

$$\begin{aligned} \widehat{\phi} &= \arg\min_{\phi} \frac{1}{N} \sum_{i=1}^{N} \|s_\phi(T(x_i)) - A(F_\theta(x_i))\|_2^2 + \lambda_{\mathrm{reg}} \Omega(\phi), \\ S_\phi(x) &= A^{-1}\{s_\phi(T(x))\}. \end{aligned} \tag{4}$$

No task outcome enters the student regression objective. The numerical interpretation of $\lambda_{\mathrm{reg}}\Omega(\phi)$ depends on the solver's objective normalization. Bank Marketing, Reading and Credit-g use Ridge regression with $\alpha = 1$ under Ridge's sum-of-squares convention. NetFlow uses a multi-output MLP with one hidden layer of 64 ReLU units and K outputs. Any response validation uses only held-out donor records. Appendix B gives the student training settings.

For Bank Marketing, Reading, Credit-g, and canonical NetFlow, donors are support-pool records remaining after the current labeled support is removed. The separate NetFlow evaluation uses the complete response-bearing donor pool. Donors, current support records, and targets are disjoint within each student-fitting condition.

Each probe bank defines a separate fit; coordinates are neither pooled across banks nor selected by task performance. With the behavioral target and donors fixed, student fitting precedes use of current support outcomes, which enter only the downstream task head.

3.3 Fifty-label task calibration and deployment

Each task head receives a labeled support set $\mathcal{S}_B = \{(x_i, y_i)\}_{i=1}^B$, with $B = 50$. The downstream representation is

$$z(x) = \begin{cases} A_B(S_\phi(x)), & \text{Bank Marketing and Reading,} \\ [T_B(x), A_B(S_\phi(x))], & \text{Credit-g and NetFlow.} \end{cases} \tag{5}$$

The response scaler A_B is fitted only on the current labeled support. Downstream raw preprocessing T_B uses permitted response donors for Credit-g and canonical NetFlow, and the current support for the separate NetFlow evaluation. The Bank student's raw preprocessing remains donor-fitted. No target examples are used to fit preprocessing, scaling, classifier parameters, or decision thresholds.

Bank Marketing, Reading and Credit-g use logistic regression for the task head. NetFlow uses logistic and RBF-SVM readouts with fixed hyperparameters. Where teacher responses are available, teacher

comparisons replace $S_\phi(x)$ with the true $F_\theta(x)$ under the same downstream head and evaluation procedure.

At deployment, a new record is processed only by the raw-feature pipeline, the student, and the fitted task head:

$$x_{\text{new}} \longmapsto S_\phi(x_{\text{new}}) \longmapsto h_B(z(x_{\text{new}})), \qquad \boxed{\text{LLMCalls}(x_{\text{new}}) = 0}. \tag{6}$$

LLM use is confined to offline response acquisition.

Response-bearing donors are a separate resource from the 50 labeled support records. The Bank and Reading heads calibrate the predicted scalar through a signed slope and an intercept.

4 Experimental setup

Panels, responses and predictors. We evaluate Bank Marketing (Moro et al., 2014), Reading (Siam et al., 2025), Credit-g (Hofmann, 1994), and the four-source NF-UQ-NIDS NetFlow-v1 benchmark (Sarhan et al., 2021). Bank has 384 chronologically held-out targets; Credit-g has 280 balanced targets. Canonical NetFlow has 768 targets, with 96 records per source–class cell. Its separately allocated evaluation has 192 targets per source, selected without their outcomes; no teacher responses are acquired for these supports or targets. Appendix A gives all roles and counts.

Gemma-3-12B supplies the Bank and Credit-g responses, Llama-3.1-8B-Instruct the NetFlow responses, and Qwen3-14B (Yang et al., 2025), with thinking disabled, the Reading responses. Reading uses the same signed antisymmetric first moment as Bank, with the functional and $K = 64$ specified before target scoring. Its 582 records are allocated without outcomes: 250 response-pool records form five disjoint 50-record supports, 128 form two disjoint 64-probe banks, and 204 are held-out targets. Each support excludes its own records from the student's 200 donors. A prespecified subset of 64 targets has true responses for teacher and fidelity assays; deployment is evaluated on all 204. Appendix I retains the prespecification and prediction provenance.

Bank excludes call duration. Credit-g preserves the mixed-variable interpretation of its twenty fields. NetFlow retains ten measurements, log-transforms counts and duration, and excludes source identity, addresses and attack family. Reading has two numerical and fourteen categorical predictors. Mixed-data preprocessing standardizes numerical variables and one-hot encodes categories, with median/modal imputation where needed; fitting roles follow Section 3.

Readouts and matched controls. The primary task heads use 50 outcomes, separately from response-bearing donor supervision. Logistic heads use $C = 1$, balanced weights and `liblinear`; RBF-SVM uses $C = 1$, $\gamma =$ `scale` and balanced weights. Raw-input references include conventional learners selected by support-only cross-validation and fixed TabPFN-3.5/TabICLv2 settings. Reading's seven comparators share its five supports and 204 targets; three-fold support CV selects booster settings. Donor-matched semi-supervised controls receive the same raw donors without their outcomes or LLM responses (Appendix C).

Response-family controls preserve roles, responses and heads, comparing symmetric and antisymmetric means, activity, RMS and probe-resolved profiles. Scalar controls use Ridge; NetFlow profiles use MLP64. True-response readouts separate response utility from student approximation. The matched second-teacher experiment uses Llama on the same Bank panel, with the same signed construction and student/head recipe, without another functional search (Appendix E.8). The Bank coefficient interventions retain the fitted scaler and head while preserving support logits (Appendix E.5).

A secondary support-only selector chooses among raw, scalar and probe-resolved readouts from donor-fitted Ridge students, jointly selecting logistic regularization by three-fold pooled class-balanced OOF log loss without target outcomes or new LLM responses (Appendix E.9).

Support and resource ladders. Outcome-blind support resampling holds the response constructions and evaluation panels fixed and uses the remaining pool records as donors. Its overlapping draws characterize support dependence conditional on these pools, not independent datasets. The Bank donor ladder varies response-bearing records while keeping the response and downstream procedure prescribed, to track reconstruction and downstream prediction. A separate balanced raw-label ladder

describes raw-LR performance within the existing 250-record pool; it does not equate donors with labels. In prior reuse, the behavioral student and 50-label prior remain fixed while additional labels train downstream boosters; raw and prior-initialized models are tuned independently on support data. This task-dependent comparison also includes separate same-source NetFlow experiments. Appendix F gives these designs and the cached probe-count sensitivity.

Metrics and uncertainty. Balanced accuracy (BA) is primary; AUROC and macro-F1 are secondary. Metrics average supports and banks, not predictions. Support SD first averages banks and is descriptive. Paired target-bootstrap intervals hold fitted models fixed and share target resamples within the prescribed strata; Reading clusters identical predictors. Bank and Reading principal comparisons use 10,000 draws. Prior-reuse intervals additionally resample anchors and augmentation sequences. Neither shared targets nor overlapping supports constitute independent datasets. The appendix retains complete resampling procedures, configurations and secondary metrics.

5 Results

5.1 The same signed response construction improves low-label prediction in two distinct settings

On Bank Marketing, the true signed antisymmetric response reaches 64.53% balanced accuracy (BA) and its Ridge student 63.61%, only 0.92 points lower. The student gains 6.11 points $[3.58, 8.54]$ over raw logistic regression and 8.45 $[4.05, 12.69]$ over Extra Trees. Gains over TabPFN-3.5 and TabICLv2 are 8.77 and 9.24 points; all four paired intervals have positive lower bounds (Appendix C.1).

Using the same unlabeled raw donors without LLM responses, self-training logistic regression, self-training Extra Trees and label spreading reach 54.06%, 53.91% and 50.73% BA, respectively (Appendix C.2).

Reading uses the *same signed response construction*, with the functional and probe count specified before target scoring and a different LLM backbone. The student reaches 73.20% BA versus 66.09% for raw logistic regression, a 7.11-point gain $[1.95, 12.24]$. It has the highest BA point estimate among the tested same-panel methods: XGBoost reaches 70.49%, LightGBM 67.93%, CatBoost 66.53%, TabPFN-3.5 65.87%, Extra Trees 65.61% and TabICLv2 65.35%. The XGBoost difference is 2.72 points $[-0.99, 6.31]$; intervals against the other six comparators have positive lower bounds (Figure 2; Appendix D.1).

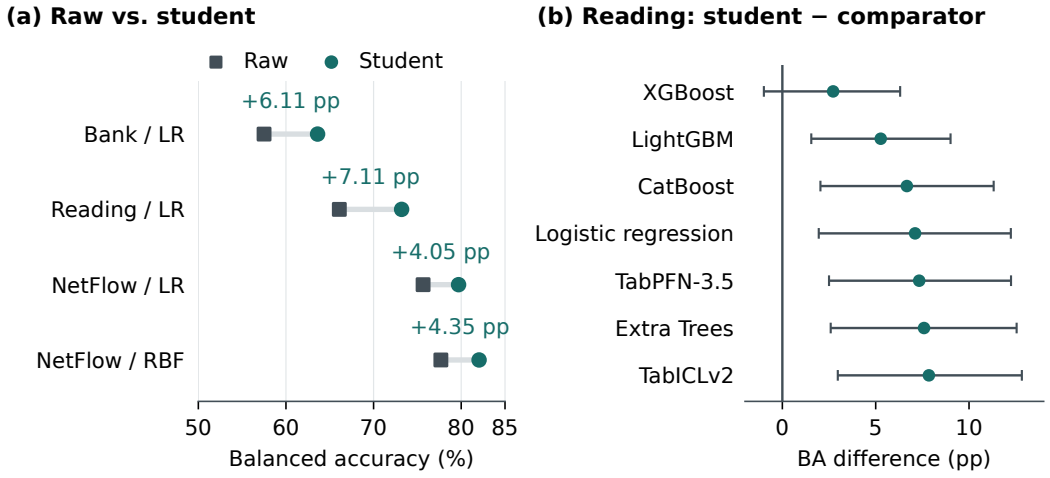


Figure 2: Low-label deployment comparisons with 50 outcome labels per head. (a) Connected raw/student BA pairs on a local 50–85% scale; annotations give gains in percentage points. (b) Reading student-minus-comparator BA with paired 95% target-bootstrap intervals on the same 204 targets, conditional on the five supports and fitted models (Appendix D.1). Banks are separate fits, not a prediction ensemble.

5.2 Matched controls identify the signed label-induced component

Matched Bank donors, Ridge students and scalar heads isolate the signed label-induced component (Figure 3). Against the signed first moment's 63.61% BA, the symmetric mean reaches 50.03%, unsigned activity 50.38%, antisymmetric RMS 51.48%, the candidate's direct LLM margin 50.68%, donor-shuffled supervision 50.84% and a fixed response projection 55.31%. The smallest gap is 8.30 points (Appendix E.4).

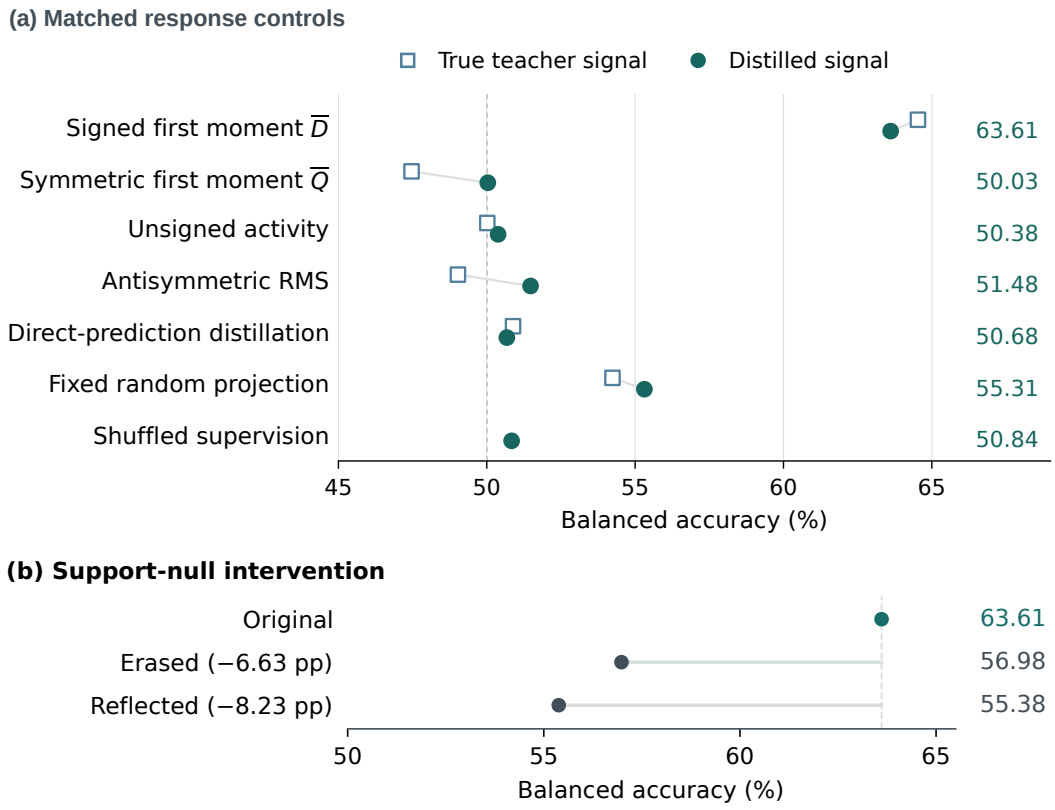


Figure 3: Bank response controls and support-null intervention. (a) Open squares: true-response BA; filled circles: distilled BA. Scalar controls share Ridge and the 50-label logistic head; direct prediction uses the candidate's margin. The dashed line marks 50% BA; shuffled supervision has no target-teacher signal. (b) Erasure/reflection losses are annotated; labeled-support logits remain unchanged to numerical precision (Appendix E.5).

Raw-derived controls do not reproduce the effect: PCA-1 gives 50.00% BA despite $R^2 = 0.99995$; eight raw projections average 50.15%, a one-dimensional autoencoder gives 51.67%, and resource-matched ProbeSim64 and DAE16 give 52.17% and 53.33% (Appendix E.6). These matched comparisons associate the gain with the signed change induced by the hypothetical label.

5.3 Response fitting selects predictive structure left unresolved by the labeled support

For the scalar student, the donor-fitted response predictor is affine,

$$S(x) = w^\top T(x) + a,$$

and the downstream logistic head has coefficient direction

$$\theta_B = \eta_B w. \tag{7}$$

Response fitting determines w before support outcomes enter the head; the support fits its scale, orientation and intercept.

Decompose the coefficient relative to the centered encoded support's row space:

$$w = w_{\parallel} + w_{\perp},$$

where $w_{\perp}$ lies outside that variation. Erasing or reflecting $w_{\perp}$, with an intercept adjustment, preserves support logits to order 10^{-14}. Held-out BA nevertheless falls from 63.61% to 56.98% after erasure and 55.38% after reflection: losses of 6.63 points $[4.57, 8.65]$ and 8.23 $[5.78, 10.63]$ (Appendix E.5).

Under the same intervention, 256 prespecified raw-random directions have maximum erasure/reflection losses of 4.01/5.52 points; the first ten donor-only PCA directions have maxima of 5.21/6.20. None matches the behavioral direction's original BA or either loss, while all preserve support logits to numerical precision (Appendix E.7).

Appending the scalar to raw attributes also loses the restriction: the augmented head reaches 55.47% BA versus 63.61% for the scalar-only head, a difference of -8.14 points $[-11.51, -4.67]$. Its score component aligned with the original scalar alone retains 63.66%. Thus most of the loss arises when the head leaves the response-selected direction, rather than from scalar calibration. Response fitting constrains predictive directions that agreement on the labeled support leaves unresolved.

5.4 Useful response structure can remain probe-resolved

On canonical NetFlow, the distilled magnitude profile improves logistic regression from 75.65% to 79.70% BA and RBF-SVM from 77.68% to 82.04%, gains of 4.05 and 4.35 points, positive on nine and ten of ten supports, respectively.

In the separately allocated evaluation, LR rises from 79.90% to 83.37%, a 3.48-point gain $[1.64, 5.35]$ positive on all five supports; RBF-SVM changes from 83.45% to 84.18%, a 0.73-point difference $[-0.74, 2.21]$ (Appendix D.3).

Probe identity is the main distinction: matched canonical MLP64–RBF signed and magnitude profiles reach 81.93% and 82.04% BA, with a paired interval including zero; the signed profile is slightly higher on the separate evaluation. Scalar collapses remain near raw baselines. Since scalar students use Ridge and profiles use MLP64, the true-response assay checks this distinction before approximation: response-only LR reaches 76.91% for the signed profile and 77.23% for magnitude, versus 61.09% for the signed mean (Appendix E). Useful information can remain distributed across probes without one profile being universally preferred.

5.5 Response utility and amortization fidelity are distinct bottlenecks

Credit-g separates accurate reconstruction from utility. Its diagnostic response is the symmetric mean $\bar{Q}$. Its Ridge student attains $R^2 = 0.948$, Pearson $r = 0.974$ and Spearman $r_s = 0.973$, yet BA rises only from 64.57% for raw LR to 65.43%, a 0.86-point gain. The true response with raw features gives 65.50%: excellent reconstruction of a response with little additional utility (Appendix D.2).

The second Bank teacher gives the opposite failure: its signed response reaches 62.95% BA, 5.45 points above raw LR $[2.19, 8.70]$, but the Ridge student reaches 56.98%, essentially raw-LR performance and 5.97 points below the teacher $[-9.15, -2.74]$. Pearson correlation is 0.864, but $R^2 = -0.279$, with location, orientation and ranking errors: the student loses useful response structure (Appendix E.8).

Amortization data also matter: increasing Bank donors from 16 to 200 raises target R^2 from -0.886 to 0.695 and BA from 50.81% to 63.61%. Against raw LR's 57.50%, the student is below baseline at 16, 32 and 50 donors, roughly level at 100 and positive at 150 and 200 (Appendix F.3).

5.6 Robustness and resource sensitivity

Across 1,000 outcome-blind Bank support draws, student gains have mean $+4.54$ and median $+5.69$ points, positive on 63.6% but with a substantial negative tail. Credit-g averages $+0.50$ over 1,000

draws. Over 100 canonical NetFlow draws, LR and RBF gains average $+4.14$ and $+4.06$ points, positive on 93% and 88%, respectively. These overlapping draws characterize support dependence within fixed pools, not independent datasets; Appendix F.4 gives the distributions and protocol.

Within the tested Bank pool and recipe, raw LR remains below the historical 63.61% behavioral system through 250 labels (58.68%); donors and outcome labels remain distinct resources. Reusing the fixed prior improves all nine independently support-tuned booster/budget comparisons on Bank, with positive hierarchical paired interval lower bounds, but gives no positive tuned mean gain in the same-source NetFlow experiments at 250 or 1000 labels. Appendices F.5–F.6 give the full label ladders and task-dependent reuse analysis.

A support-only selector over the same cached responses raises equal-dataset BA from 65.87% to 67.70% across eight tasks; restricting the identical procedure to raw features gives 65.66%. The analysis requires no additional LLM responses (Appendix E.9).

6 What cross-example response supervision contributes

Cross-example response fitting can supply a learned restriction rather than simply another input feature. In the scalar setting, it selects a direction before support outcomes enter the task head. The coefficient intervention shows that this direction contains predictive structure invisible on the labeled support; ordinary raw-feature augmentation allows the head to leave that restriction.

Response utility, amortization fidelity and low-label calibration are distinct requirements. Credit-g separates faithful reconstruction from downstream utility; the second Bank teacher separates a useful response from its successful amortization. The direct candidate margin contains ranking information but is difficult to orient and calibrate from scarce labels (Appendix E.5). None of these properties alone guarantees a useful deployed predictor. On Adult, changing the teacher from Qwen3-14B to DeepSeek V4 Flash raises distilled BA by 6.47 points under the otherwise fixed protocol (Appendix E.10).

What must be retained depends on the task. A signed first moment suffices in Bank and Reading, whereas NetFlow retains useful variation across probes. The paired response is the underlying measurement; choosing a scalar or probe-resolved functional determines which structure the student must preserve.

Response donors and outcome labels supply different supervision: donors train response reconstruction, while labels fit the downstream task predictor. The support and budget analyses characterize this dependence without making the resources interchangeable. Reuse with a more flexible downstream learner remains task-dependent (Appendices F.3–F.6). The LLM operates only offline: new-record prediction uses preprocessing, the student and the task head, with neither response donors nor LLM calls at deployment.

Scope. The evidence covers binary tabular tasks, a small set of backbones and predetermined probe banks. Historical response choices were developed retrospectively; Reading evaluates a prespecified signed construction with a different backbone. Donors are an additional training resource, and public benchmarks may overlap with LLM pretraining data. A support-only selector reduces the need to hand-specify a functional; selection beyond its fixed family remains open (Appendix E.9).

7 Conclusion

Cross-example responses from large language models can be distilled into useful supervision for low-label prediction. Changing the part of the learned direction that the labeled examples cannot distinguish leaves their fitted scores unchanged but sharply degrades held-out prediction. Response fitting therefore supplies predictive structure that the labeled examples alone do not identify. A signed response is effective in a distinct low-label setting with a different LLM backbone, while in another setting useful information remains distributed across individual probes rather than collapsing to a scalar. The experiments also separate response fidelity from downstream utility. In this formulation, the large language model serves as an offline measurement instrument rather than a deployment-time predictor. Prediction on new records uses only the student and task head, with no LLM calls.

Reproducibility statement. The accompanying source package provides the experimental recipes, role identities or hashes, executed notebooks, numerical traces, plotting code, and vector figures, including the Reading primitive-response cache, frozen deployment objects, prediction locks and same-panel baseline package. Appendix I documents experimental provenance, replay requirements, and the scope of deterministic verification.

AI use statement. We used AI tools for coding, debugging, code documentation, data analysis, literature review, manuscript drafting, editing and compilation, and figure preparation.

Ethics statement. The study uses public tabular benchmarks. Network addresses, source identity, and attack-family labels are excluded from the NetFlow predictors, and call duration is excluded from Bank Marketing. The cybersecurity results are intended for controlled benchmark evaluation and should not be interpreted as evidence of safe operational deployment or robustness to adaptive adversaries (Sommer & Paxson, 2010; Arp et al., 2022). Dataset attribution and cached-input requirements are documented in the supplementary material.

REFERENCES

Niloufar Alipour Talemi, Hossein Kashiani, and Fatemeh Afghah. Hyper-ICL: Attention Calibration with Hyperbolic Anchor Distillation for Multimodal In-Context Learning. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=rDKFflrjZK.

Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. Dos and Don'ts of Machine Learning in Computer Security. In *31st USENIX Security Symposium*, pp. 3971–3988, 2022. URL https://www.usenix.org/conference/usenixsecurity22/presentation/arp.

Ting-Yun Chang and Robin Jia. Data Curation Alone Can Stabilize In-context Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8123–8144, 2023. doi: 10.18653/v1/2023.acl-long.452. URL https://aclanthology.org/2023.acl-long.452/.

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. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

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 38, Datasets and Benchmarks Track*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1697e3fb412da11dc9488249f9e7bbc9-Abstract-Datasets_and_Benchmarks_Track.html.

Josh Gardner, Juan C. Perdomo, and Ludwig Schmidt. Large Scale Transfer Learning for Tabular Data via Language Modeling. In *Advances in Neural Information Processing Systems*, volume 37, pp. 45155–45205, 2024. doi: 10.52202/079017-1435. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4fd5cfd2e31bebbccfa5ffa354c04bdc-Abstract-Conference.html.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 2006. doi: 10.1007/s10994-006-6226-1. URL https://link.springer.com/article/10.1007/s10994-006-6226-1.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting Deep Learning Models for Tabular Data. In *Advances in Neural Information Processing Systems*

34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html.

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 35, Datasets and Benchmarks Track*, pp. 507–520, 2022. doi: 10.52202/068431-0037. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract-Datasets_and_Benchmarks.html.

Sungwon Han, Jinsung Yoon, Sercan O Arik, and Tomas Pfister. Large Language Models Can Automatically Engineer Features for Few-Shot Tabular Learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 17454–17479, 2024. URL https://proceedings.mlr.press/v235/han24f.html.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. TabLLM: Few-shot Classification of Tabular Data with Large Language Models. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 5549–5581, 2023. URL https://proceedings.mlr.press/v206/hegselmann23a.html.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL https://research.google/pubs/distilling-the-knowledge-in-a-neural-network/. Workshop paper; also arXiv:1503.02531.

Hans Hofmann. Statlog (German Credit Data), 1994. URL https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data. Dataset; Creative Commons Attribution 4.0 International.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Noah Hollmann, Samuel Müller, and Frank Hutter. Large Language Models for Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44753–44775, 2023b. doi: 10.52202/075280-1938. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8c2df4c35cdbee764ebb9e9d0acd5197-Abstract-Conference.html.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. doi: 10.1038/s41586-024-08328-6. URL https://www.nature.com/articles/s41586-024-08328-6.

Inwon Kang, Parikshit Ram, Yi Zhou, Horst Samulowitz, and Oshani Seneviratne. Language Model Representations for Efficient Few-Shot Tabular Classification. In *Proceedings of The ACM Web Conference 2026*, 2026. doi: 10.1145/3774904.3792477. URL https://doi.org/10.1145/3774904.3792477.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Yichuan Li, Xiyao Ma, Sixing Lu, Kyumin Lee, Xiaohu Liu, and Chenlei Guo. MEND: Meta demonstration distillation for efficient and effective in-context learning. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=2Y5kBPtU0o.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Ling-Hao Chen, Junhao Liu, Tongliang Liu, Fei Huang, and Yongbin Li. One-Shot Learning as Instruction Data Prospector for Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4586–4601, 2024b. doi: 10.18653/v1/2024.acl-long.252. URL https://aclanthology.org/2024.acl-long.252/.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying Distillation and Privileged Information. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.03643.

Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When Do Neural Nets Outperform Boosted Trees on Tabular Data? In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, pp. 76336–76369, 2023. doi: 10.52202/075280-3337. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f06d5ebd4ff40b40dd97e30cee632123-Abstract-Datasets_and_Benchmarks.html.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

S. Moro, P. Rita, and P. Cortez. Bank Marketing, 2014. URL https://archive.ics.uci.edu/dataset/222/bank+marketing. Dataset; Creative Commons Attribution 4.0 International.

Jaehyun Nam, Kyuyoung Kim, Seunghyuk Oh, Jihoon Tack, Jaehyung Kim, and Jinwoo Shin. Optimized Feature Generation for Tabular Data via LLMs with Decision Tree Reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 92352–92380, 2024. doi: 10.52202/079017-2932. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a7ebe2e8d8cfd2fcec6cd77f9e6fd34d-Abstract-Conference.html.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free Concept Bottleneck Models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FlCg47MNvBA.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html.

Danil Provodin, Bram van den Akker, Christina Katsimerou, Maurits Clemens Kaptein, and Mykola Pechenizkiy. Rethinking Knowledge Transfer in Learning Using Privileged Information. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=dg1tqNIWg3.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 50817–50847, 2025. URL https://proceedings.mlr.press/v267/qu25d.html.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A Better, Faster, Scalable, and Open Tabular Foundation Model. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=SxsyLjIfWB.

Mohanad Sarhan, Siamak Layeghy, Nour Moustafa, and Marius Portmann. NetFlow Datasets for Machine Learning-Based Network Intrusion Detection Systems. In *Big Data Technologies and Applications: BDTA 2020 and WiCON 2020*, pp. 117–135, 2021. doi: 10.1007/978-3-030-72802-1_9. URL https://doi.org/10.1007/978-3-030-72802-1_9.

Günther Schindler, Maximilian Schambach, and Johannes Höhne. Enhancing Tabular Learners with Context-Aware Semantic Embeddings. In *2nd ICML Workshop on Foundation Models for Structured Data*, 2026. URL https://openreview.net/forum?id=QArxQg4U71. Workshop paper.

A. K. M. Fazlul Kobir Siam, Md Asraful Sharker Nirob, and Prayma Bishshash. Reading habits based on gender: Undergraduate students' dataset on book preferences and engagement. Mendeley Data, version 1, 2025. URL https://data.mendeley.com/datasets/24wf5n2gk5/1.

Robin Sommer and Vern Paxson. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. In *2010 IEEE Symposium on Security and Privacy*, pp. 305–316, 2010. doi: 10.1109/SP.2010.25. URL https://doi.org/10.1109/SP.2010.25.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Vladimir Vapnik and Rauf Izmailov. Learning Using Privileged Information: Similarity Control and Knowledge Transfer. *Journal of Machine Learning Research*, 16:2023–2049, 2015. URL https://www.jmlr.org/papers/v16/vapnik15b.html.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently, 2023. URL https://arxiv.org/abs/2303.03846. arXiv preprint.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Haolin Yang, Hakaze Cho, Kaize Ding, and Naoya Inoue. Task vectors, learned not extracted: Performance gains and mechanistic insights. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RGEbVZgf4E.

Appendix guide. Appendix A gives datasets, experimental roles, and response construction. Appendix B gives the student and task-head implementation. Appendix C documents baselines and matched comparators. Appendix D reports complete deployment metrics, teacher assays, and paired comparisons. Appendix E studies response structure, matched controls, and the Bank mechanism. Appendix F reports support, probe, donor, label-budget, and fixed-prior sensitivity analyses. Appendix G gives the source-wise decomposition of the fixed NetFlow systems. Appendix H summarizes amortization fidelity and offline/deployment resource accounting. Appendix I records experimental chronology, provenance, prediction locks, replay coverage, and known limitations. Appendix J compares selected related supervision mechanisms.

A Datasets, roles, and response construction

A.1 Response notation and stored quantities

The semantic-aligned paired margins give $Q = (m_1 + m_0)/2$ and $D = (m_1 - m_0)/2$. Historical notation uses $P = 2D$: Bank polarity and NetFlow MAG64 differ from $\bar{D}$ and $|D|$ only by this positive scaling. Hypothetical labels are controlled interventions, not claims about the candidate's true outcome. Equation 3 gives the primary deployment functionals; Appendix E gives the experimental response family.

A.2 Experimental roles

Table 1 includes response-validation donors in N. Original-task historical main heads use 25 labels per class; Reading uses fifty outcome-blind, unbalanced support records. Support and bank identities remain fixed across matched controls. Banks are repeated measurements, not independent datasets.

Table 1: Experimental roles and response supervision by panel.

Panel	Response target / student	N	B	Targets	Supports	Banks
Bank	$2\bar{D}$ / Ridge	200	50	384	5	3
Reading	$2\bar{D}$ / Ridge	200	50	204	5	2
Credit-g	$\bar{Q}$ / Ridge	200	50	280	5	2
NetFlow, canonical	$2\lvert D\rvert$ / MLP64	450	50	768	10	3
NetFlow, evaluation	$2\lvert D\rvert$ / MLP64	1,268	50	768	5	3

A.3 Bank Marketing

The Bank Marketing data use UCI's bank-additional version. The raw predictors are age, job, marital status, education, default, housing, loan, contact, month, day of week, campaign, pdays, previous, previous outcome, employment variation, consumer price and confidence indices, Euribor and number employed. Call duration is excluded; raw dates and role identities are not added as predictors. The chronological record ranges for candidate supports, probes and targets are $[0, 27690)$, $[27690, 39130)$ and $[39130, 41188)$ in the original ordering. The target has equal class counts within three late chronological strata. This is temporal separation of records, not a verified customer-level split.

Five disjoint supports form a 250-record pool. For each support, the other 200 records are response donors; the 384 targets never enter student fitting. Three banks each contain 64 probes. All fifteen support/bank conditions use Ridge $\alpha = 1$.

A.4 Reading

The workbook *Reading Habits Among Undergraduate Students* is Mendeley Data version 1, DOI `https://doi.org/10.17632/24wf5n2gk5.1`, under CC BY 4.0 (Siam et al., 2025). Its single sheet, `Form Responses 1`, has 582 rows. The target is "Which type of books do you read most of the time": Non-Academic is class 0 and Academic class 1. The sixteen other fields are predictors. Age and "Amount of time spent reading books per day" are numerical; the fourteen other fields, including book-count ranges and frequency descriptions, remain categorical. Literal `None`

responses in the e-book and printed-book count fields (89 and 50 rows) are categories, not missing entries. There are no missing predictor values.

The categorical predictors are “Gender”; “Department”; “Reason for reading books”; “How to manage books for reading”; “Number of e-books you read last two years”; “Number of printed books you read last two years”; “Favorite books for reading to spend leisure time”; “What is the change that you fell after reading books”; “Who influenced you to read books”; “Which language books do you like to read”; “Where do you read printed books”; “Do you read Newspaper”; “The format that you used for reading books”; “Frequency of reading books”. Field spelling is preserved in the encoder and prompt strings.

Exact-predictor groups are shuffled without outcomes to allocate five 50-record supports, two disjoint 64-probe banks and 204 targets. The supports form the 250-record response pool; each student’s donors are the other 200 pool rows. The one duplicate predictor pair stays together in the target role. Pool, probes and targets are disjoint. Supports are not balanced: their class 0/1 counts are 21/29, 19/31, 10/40, 21/29 and 19/31. Targets contain 79/125 records of classes 0/1. A fixed 64-target assay is selected without outcomes and contains 22/42 records after scoring. These 64 belong to the 204-target panel.

Prompts and teacher. Candidate and probe rows use the same deterministic newline-separated `Feature name: value` serializer in workbook order, excluding the target. Actual field names and values are retained. The two system instructions name the target, map A/B to Non-Academic/Academic (then reverse the mapping), and request exactly A or B. Messages are system instruction, candidate user row, hypothetical assistant answer, and unlabeled probe user row. Only the candidate’s hypothetical label changes within a label pair.

Qwen3-14B uses thinking-disabled immediate next-token logits, without text generation. Both A/B mappings are measured and their semantic-aligned margins are averaged into m_0, m_1. The prescribed response is the same signed first moment as Bank, $2\bar{D} = K^{-1}\sum_j (m_{1j} - m_{0j})$, at $K = 64$.

A.5 CREDIT-G

For the diagnostic experiment, Credit-g uses the symmetric mean $\bar{Q}$ and its corrected mixed-variable interpretation on a fixed 280-record target panel. Numerical variables are duration, amount and age. The remaining fields retain their categorical coding, including installment rate, present residence, number of credits and people liable; no new semantic meanings are inferred for the codes. Five disjoint fifty-label supports form the response pool, with 200 donors per condition and two 64-probe banks. Bank and Credit-g raw subsets retain UCI CC BY 4.0 attribution.

A.6 NETFLOW

The canonical dataset is `NF-UQ-NIDS.csv`, using NetFlow-v1 measurements. Its ten predictors are

> `L4_SRC_PORT`, `L4_DST_PORT`, `PROTOCOL`, `TCP_FLAGS`, `L7_PROTO`, `IN_BYTES`, `OUT_BYTES`, `IN_PKTS`, `OUT_PKTS`, `FLOW_DURATION_MILLISECONDS`.

The five byte, packet and duration variables use $\log(1 + x)$ in student and conventional linear/kernel pipelines. Both ports are retained. This differs from the original dataset paper’s evaluation, which removed ports as well as addresses; we do not reproduce its classifier protocol. Source identity is used only for allocation and reporting. Addresses, other direct identifiers and attack family are excluded from model input.

The canonical panel has a mixed-support pool of 500 records and 768 balanced targets. Each of ten supports supplies fifty labels; its other 450 pool rows are donors. The separate evaluation uses five subsequently allocated balanced supports and 768 targets. Its target identities are selected without their outcomes. The later student is trained on the response-complete historical pool, including records that served as targets in the earlier panel; their outcome labels are not used for student fitting. Teacher responses do not exist for the later supports or targets. Appendix I.4 gives the exclusion and allocation record.

B STUDENT AND TASK-HEAD IMPLEMENTATION

Unless stated otherwise, student preprocessing, response scaling and student fitting use only permitted response donors; task-head scaling and fitting use only the current labeled support. Targets enter neither fitting stage.

B.1 SCALAR STUDENT PIPELINES

The scalar students use Ridge with $\alpha = 1$. Donor-fitted raw preprocessing uses median imputation and standardization for numerical variables, and modal imputation and one-hot encoding with unknown categories ignored for categorical variables. The teacher response is standardized on the same donors; Ridge fits in standardized response space and its prediction is mapped back to original units. The predicted scalar is standardized again on the current labeled support. Bank and Reading use this scalar alone; Credit-g appends it to the donor-preprocessed raw block. No row normalization is introduced.

The positive rescalings $P = 2D$ and MAG64$= 2|D|$ leave donor-standardized fitting targets and support-standardized head inputs unchanged. This equivalence is specific to the standardized pipeline.

Reading uses numerical standardization and dense one-hot encoding fitted on the 200 permitted donors. Imputation is unnecessary because its predictors are complete. Donor response standardization is inverted before support standardization and the scalar-only logistic head ($C = 1$, balanced weights, liblinear, at most 5,000 iterations). No Reading outcome enters response regression; each head uses only its current fifty support labels.

B.2 NETFLOW STUDENT PIPELINES

The canonical and later-evaluation students have the same dimensions but different training recipes (Table 2). The canonical 450 donors are split into 360 gradient-training and 90 response-validation rows. Raw and teacher scalers fit the gradient-training subset. The downstream raw block uses statistics from all 450 donors. The best response-validation epoch is retained without refitting.

Table 2: The two NetFlow MLP recipes. Both predict 64 standardized coordinates from ten raw measurements with one 64-unit ReLU hidden layer and Adam. No target metric chooses an epoch.

Setting	Canonical replay	Later evaluation
Response donors	450	1,268
Penalty `alpha`	0.001	0.01
Learning rate	0.001	0.001
Batch size	64	sklearn auto
Maximum epochs / iterations	600	500
Validation fraction	0.20	0.15
Stopping	MSE, patience 40	sklearn early stopping, 30
Improvement tolerance	10^{-5}	10^{-5}
Scaler fitting	Gradient-training rows	All historical donors
Seed	$28101 +$ support	28101

The later student uses sklearn's MLPRegressor with ReLU and Adam; all three fits reach the iteration cap. Both scalers fit all historical donors before the internal validation split. Evaluation supports and targets enter neither student training nor scaling. Downstream raw and response scalers fit only the current support.

B.3 TASK HEADS AND DECISION RULES

Logistic regression uses $C = 1$, balanced class weights, liblinear and at most 5,000 iterations. RBF-SVM uses $C = 1$, $\gamma = \text{scale}$, balanced weights and no probability calibration. Native fitted decisions are retained without target threshold selection. Each head sees only its current support outcomes. Linear-SVM and Extra-Trees references are specified in Appendix C.

C BASELINES AND MATCHED COMPARATORS

Reported historical recipes, historical support-CV recipes, the common support-only suite and fixed tabular foundation models are distinct comparisons. Target performance is not used to choose among incompatible recipes. Foundation models are same-support low-label references, not resource-matched response-donor systems or target-tuned best-case estimates.

C.1 BANK BASELINE ACCOUNTING

All Bank comparisons use the same five fifty-label supports, 384 targets and nineteen duration-excluded fields. The behavioral student averages bank metrics within each support, then supports.

C.1.1 REPORTED HISTORICAL RECIPES

Headline LR uses three stratified support folds to select $C \in \{0.03, 0.3, 1, 3, 30\}$ by mean-fold BA, with the first grid entry resolving ties. Mixed preprocessing fits inside each fold and is followed by row L2 normalization; balanced liblinear LR refits on all fifty support rows. Selected values are 3, 0.03, 30, 30 and 0.3.

Headline Extra Trees uses 300 estimators, Gini splits, square-root feature sampling, minimum leaf size 1, balanced weights, no bootstrap and no depth limit. It and the fixed 120-estimator CatBoost, XGBoost and LightGBM recipes fit raw preprocessing on the 200 permitted donors, then the classifier on the fifty supports. They receive raw donor features, not behavioral responses.

C.1.2 HISTORICAL SUPPORT-CV RECIPES

Five stratified folds select by pooled out-of-fold BA, with ties resolved by the recorded simplicity key and candidate order. There are five LR, six Extra Trees, sixteen CatBoost, eight XGBoost and eight LightGBM candidates (Prokhorenkova et al., 2018; Chen & Guestrin, 2016; Ke et al., 2017). Preprocessing fits inside each fold; CatBoost instead uses native categories, a fixed missing-category token and balanced weights. Final fits use all fifty labels.

LR uses $C \in \{0.01, 0.1, 1, 10, 100\}$. Extra Trees uses 400 estimators, feature fractions $\{\text{square root}, 0.5, 1.0\}$ and leaf sizes 1/2. CatBoost crosses 300/500 iterations, depths 4/6, learning rates 0.03/0.05 and penalties 3/10. The two eight-setting booster grids and fixed tie rules are retained in the protocol.

Table 3: Named Bank historical recipes, evaluated separately on identical supports and targets. Metrics are percentages; support SD describes the five fits. No target-score maximum combines recipes within a family.

Method / recipe	BA	Support SD	AUROC	Macro-F1
Named archived recipes				
Headline LR: row L_2, 3-fold CV	57.50	8.00	58.20	49.08
Headline Extra Trees: 300 trees	55.16	3.95	56.77	54.34
CatBoost: 120 iterations	52.24	3.14	55.72	51.53
XGBoost: 120 trees	50.62	3.18	54.64	47.67
LightGBM: 120 trees	51.98	5.59	53.73	50.92
Archived five-fold support-CV recipes				
Logistic Regression	52.03	4.12	57.37	39.47
Extra Trees	53.44	3.32	54.01	52.68
CatBoost	50.78	5.72	51.91	50.36
XGBoost	52.50	4.84	53.88	49.73
LightGBM	51.93	2.31	54.00	50.68

C.1.3 COMMON SUPPORT-ONLY PROTOCOL

All eight conventional families use the same three stratified folds within each support. Mean-fold BA selects settings; grid order resolves ties. Numerical median imputation and scaling, and categorical

modal imputation and one-hot encoding, fit inside each training fold. Selected models refit on all fifty support rows and retain native decisions. Targets enter neither selection, early stopping nor threshold choice.

LR uses $C \in \{0.01, 0.1, 1, 10, 100\}$; linear SVM uses $\{0.01, 0.1, 1, 10\}$; RBF-SVM crosses $C \in \{0.1, 1, 10\}$ with $\gamma \in \{\text{scale}, 0.1, 1\}$. All three use balanced weights. Extra Trees uses 500 estimators, balanced weights, no depth limit, feature fractions $\{\text{square root}, 0.5, 1.0\}$ and leaf sizes $\{1, 2, 4\}$. CatBoost receives the same one-hot inputs and balanced weights; LightGBM also uses balanced weights. XGBoost uses unit class weighting on the balanced supports. The ReLU MLP crosses hidden widths 16/32 with L2 penalties 0.01/1, using L-BFGS, at most 1,000 iterations and no early stopping. The six CatBoost and eight XGBoost/LightGBM candidates are given in Table 8.

Table 4: Common support-only Bank protocol. Hyperparameters use the shared three-fold support CV; no target labels enter selection. Metrics are percentages and support SD is descriptive. Some MLP L-BFGS fits reached the 1,000-iteration cap.

Method / recipe	BA	Support SD	AUROC	Macro-F1
Logistic Regression	51.56	3.08	59.59	38.44
Linear SVM	50.68	1.37	59.48	35.32
RBF-SVM	50.00	0.00	50.24	33.33
Extra Trees	51.35	4.62	53.20	48.24
CatBoost	52.45	1.45	54.77	50.87
XGBoost	53.59	3.89	54.79	51.93
LightGBM	54.06	4.31	55.35	52.63
MLP	53.28	4.86	57.88	44.84

A separate fixed capacity control uses a directly supervised Bank MLP with 64 hidden units, L-BFGS, penalty 0.01 and at most 300 iterations; it reaches 56.15% BA. A 64-unit Adam control is weaker. These models receive no response-regression targets and are not identical-capacity counterparts of the scalar Ridge student.

C.1.4 Fixed tabular foundation models

TabPFN-3.5 uses package 9.0.0 and checkpoint `tabpfn-v3.5-20260909.safetensors`; TabICLv2 uses package 2.2.0 and `tabicl-classifier-v2-20260212.ckpt`. Both use eight ensemble members and CPU inference through their official mixed-data interfaces. TabPFN uses `fit_preprocessors`, float32 and no tuning configuration. TabICL uses ensemble batch size 8 without AMP or GPU offloading. They receive no behavioral donor supervision.

Table 5: Bank student and fixed tabular foundation models. Metrics are percentages; support SD is descriptive. The student averages its three banks within each support and is not refitted.

Method / recipe	BA	Support SD	AUROC	Macro-F1
Behavioral student (saved predictions)	63.61	1.67	64.23	63.06
TabPFN-3.5 (9.0.0)	54.84	3.72	56.75	52.54
TabICLv2 (2.2.0)	54.38	3.11	56.07	51.48

The earlier TabPFNv2 evaluation (Hollmann et al., 2025) reached 55.00% BA using package 8.5.0, checkpoint `tabpfn-v2-classifier-finetuned-zk73skhh.ckpt` and eight members. Earlier Bank TabICLv2 reached 54.01% using package 2.1.1, 32 members and support-fitted ordinal categorical encoding. These remain distinct from Table 5.

Table 6: Bank student-minus-comparator BA differences. The 10,000-draw paired target bootstrap uses seed 28101 and the six historical epoch/class strata, with shared target resamples and all fitted models held fixed. Intervals are percentile intervals, unadjusted for multiplicity.

Comparator	Student advantage (pp)	95% paired interval
Archived headline LR	+6.11	[3.58, 8.54]
Archived headline Extra Trees	+8.45	[4.05, 12.69]
TabPFN-3.5	+8.77	[4.64, 12.71]
TabICLv2	+9.24	[5.07, 13.21]
Uniform LightGBM	+9.55	[4.91, 14.01]

C.2 Donor-matched semi-supervised controls

Self-training LR, self-training Extra Trees and kNN label spreading receive the same raw donor records as the behavioral student, without donor outcomes or LLM responses. Three-fold stratified support-only CV selects hyperparameters with validation rows excluded from fitting. Selected models refit on the full fifty-label support and unlabeled donors.

Table 7: Donor-matched semi-supervised controls, BA (%). Non-LLM controls receive the current 50-label support and the same raw donor records as the behavioral student, without donor outcomes or LLM responses. Behavioral columns reproduce the corresponding paper systems.

Panel	Self-training LR	Self-training Extra Trees	Label spreading	Behavioral LR	Behavioral RBF
Bank	54.06	53.91	50.73	63.61	–
Credit-g	64.07	64.86	58.86	65.43	–
NetFlow canonical	74.92	82.92	77.12	79.70	82.04
NetFlow evaluation	78.78	88.10	78.11	83.37	84.18

On Bank, the behavioral student exceeds self-training LR, self-training Extra Trees and label spreading by 9.55 [6.16, 12.78], 9.70 [5.14, 14.10] and 12.88 [8.96, 16.68] BA points, respectively (95% paired target-bootstrap intervals; 10,000 resamples within historical epoch/class strata, conditional on fixed supports and banks).

C.3 Reading same-panel baselines

All seven raw-input comparators use the same five fifty-label supports and 204 targets as the $K = 64$ student. The frozen student is replayed without refitting; LR and Extra Trees retain the original Reading recipes. No target outcome selects preprocessing, hyperparameters, thresholds, calibration or stopping.

LR uses numerical standardization and one-hot encoding followed by row L2 normalization, fitted within each training fold. Three stratified folds select $C \in \{0.03, 0.3, 1, 3, 30\}$ by mean-fold BA, with the first entry resolving ties; balanced liblinear LR refits on all fifty labels. Extra Trees uses 300 estimators, square-root feature sampling, minimum leaf size 1, balanced weights and no depth limit. Its raw encoder fits the 200 permitted donors, preserving the original recipe; donor outcomes and behavioral responses are not used.

XGBoost, LightGBM and CatBoost use the common support-CV grids (Table 8). Numerical standardization and dense one-hot encoding fit each training fold; selected models refit on all fifty support rows without row normalization. The same stratified three-fold selection and first-candidate tie rule apply. XGBoost uses the binary logistic objective, histogram trees, log-loss evaluation and balanced sample weights computed within the training fold. LightGBM uses balanced weights, deterministic column-wise fitting and the binary objective. CatBoost uses the same one-hot inputs, balanced weights and log-loss, without best-model selection. No validation or target set is used for early stopping. Native model decisions are retained.

Table 8: Reading support-only booster candidates, in fixed zero-based order. No target-based choice is made among grids or recipes.

Learner	ID	Candidate tuple (orders specified below)
XGBoost	0	100, 2, 0.03, 3, 0.8, 0.8, 5, 0.1
	1	100, 2, 0.08, 3, 1, 1, 5, 0.1
	2	100, 3, 0.03, 3, 0.8, 1, 5, 0.1
	3	100, 3, 0.08, 1, 1, 0.8, 1, 0
	4	250, 2, 0.03, 1, 1, 0.8, 1, 0
	5	250, 2, 0.08, 3, 0.8, 1, 5, 0.1
	6	250, 3, 0.03, 1, 0.8, 0.8, 1, 0.1
	7	250, 3, 0.08, 1, 1, 1, 1, 0
LightGBM	0	100, 5, 3, 0.03, 10, 5
	1	100, 5, 3, 0.08, 5, 1
	2	100, 7, 3, 0.03, 5, 5
	3	100, 15, 4, 0.08, 10, 1
	4	250, 5, 3, 0.03, 5, 1
	5	250, 7, 3, 0.08, 10, 5
	6	250, 15, 4, 0.03, 10, 5
	7	250, 15, 4, 0.08, 5, 1
CatBoost	0	200, 2, 0.03, 10
	1	200, 3, 0.05, 10
	2	300, 3, 0.03, 3
	3	300, 4, 0.05, 10
	4	400, 4, 0.03, 3
	5	400, 5, 0.05, 10

XGBoost tuples are (trees, depth, learning rate, minimum child weight, row fraction, feature fraction, L2, L1); LightGBM tuples are (trees, leaves, depth, learning rate, minimum leaf count, L2); CatBoost tuples are (iterations, depth, learning rate, L2). Other parameters use the recorded package defaults or fixed settings above.

Table 9: Reading support class counts and selected hyperparameters. Booster entries index Table 8; class 1 is Academic. The supports were allocated without outcomes and are not balanced.

Support	Class 0/1	LR C	XGBoost	LightGBM	CatBoost
0	21/29	0.3	0	0	3
1	19/31	3	0	3	5
2	10/40	30	1	1	0
3	21/29	0.3	1	0	2
4	19/31	3	3	1	1

TabPFN-3.5 uses package 9.0.0, repository `Prior-Labs/tabpfn_3_5`, checkpoint `tabpfn-v3.5-20260909.safetensors`, eight members, `fit_preprocessors`, float32 inference, one preprocessing job and no tuning configuration. The fourteen categorical column indices are supplied explicitly. TabICLv2 uses package 2.2.0, repository `jingang/TabICL`, checkpoint `tabicl-classifier-v2-20260212.ckpt`, eight members, ensemble batch size 8 and float32, without AMP, FlashAttention-3 or offloading. Both use raw mixed-type DataFrames with only current support labels; neither receives donor records, probe information, behavioral responses or student features. Both run on CPU through their official interfaces.

C.4 Credit-g and NetFlow reference configurations

Credit-g's Extra Trees comparator uses 400 trees and three-fold stratified support CV over feature fractions $\{\text{square root}, 0.5, 1.0\}$ and leaf sizes $\{1, 2\}$. Its five selected settings are (square root, 2), (square root, 1), (square root, 2), (square root, 2) and (0.5, 2). Its BA ties the separate 300-tree replay. The fixed directly supervised MLP32 uses Adam and reaches 65.71% using only current support

outcomes. TabICLv2 uses package 2.2.0, checkpoint `tabicl-classifier-v2-20260212.ckpt` and eight members.

Both Llama NetFlow panels use Extra Trees with 300 estimators, Gini splits, square-root feature sampling, minimum leaf size 1, balanced weights, no bootstrap and no depth limit. Canonical TabICLv2 uses package 2.2.0 and 32 members; the evaluation-panel baseline additionally uses ensemble batch size 8. It receives the ten untransformed float32 measurements and performs internal preprocessing. Raw comparators receive no response or student coordinate. Checkpoint, ensemble, threshold and seed are not selected using target performance; exact fit seeds and replay coverage are recorded in Appendix I and its implementation ledger.

D Full metrics and paired comparisons

D.1 Reading: deployment and teacher preservation

Deployment panel. Table 10 reports the 204-target comparison. Student metrics average the two banks within each support, then the five supports; predictions are not ensembled. Support SD uses divisor $5 - 1$ and is descriptive. AUROC and macro-F1 use the same aggregation. All targets remain outside student and head fitting.

Table 10: Reading on 204 held-out targets with 50 labels per head. Metrics are percentages. Support SD is the descriptive BA spread over five fixed supports; the behavioral student uses the prespecified Bank signed first moment, $K = 64$.

Method	BA	Support SD	AUROC	Macro-F1
Behavioral student	73.20	0.74	76.15	73.05
XGBoost	70.49	8.01	72.46	70.20
LightGBM	67.93	11.06	72.25	67.55
CatBoost	66.53	8.62	71.98	66.25
Logistic regression	66.09	6.58	73.01	65.82
TabPFN-3.5	65.87	9.13	75.89	64.62
Extra Trees	65.61	10.12	75.85	64.45
TabICLv2	65.35	6.60	73.17	65.11

The student has the highest BA point estimate. Its XGBoost interval includes zero; the other six intervals have positive lower bounds (Table 11). These are pointwise comparisons without multiplicity adjustment.

Table 11: Reading student-minus-comparator BA differences in percentage points on the same 204 targets. Intervals use 10,000 paired target resamples (seed 28101), conditional on the fixed supports, banks and fitted models.

Comparator	Student − comparator	95% paired interval
XGBoost	$+2.72$	$[-0.99, 6.31]$
LightGBM	$+5.27$	$[1.55, 9.01]$
CatBoost	$+6.67$	$[2.04, 11.32]$
Logistic regression	$+7.11$	$[1.95, 12.24]$
TabPFN-3.5	$+7.33$	$[2.50, 12.25]$
Extra Trees	$+7.59$	$[2.59, 12.55]$
TabICLv2	$+7.85$	$[2.97, 12.83]$

The bootstrap samples predictor groups with replacement within each target class, sharing each resample across methods and fits. The identical predictor pair is kept together. Each draw recomputes class denominators, averages class recalls and retains support/bank averaging. This is finite-target uncertainty conditional on fixed fits, not variation over new supports or probe banks. The five supports share donors and targets and are not independent datasets.

Teacher assay. Only 64 prespecified targets have true Qwen responses. Table 12 compares teacher, student and raw baselines on those same rows, a subset of the 204 deployment targets. The teacher head uses true support responses with the same fifty labels and scalar logistic settings. Across ten support/bank fits, reconstruction has mean $R^2 = 0.913$, Pearson $r = 0.957$ and MSE 1.2435 in original signed-response units. Target outcomes are not needed for reconstruction metrics.

Table 12: Reading teacher preservation on the 64-target assay only. Metrics are percentages; these are not teacher results on the full 204-target deployment panel.

Method	BA	Support SD	AUROC	Macro-F1
Logistic regression	63.03	9.13	72.23	62.23
Extra Trees	63.03	11.56	74.81	61.40
Behavioral student	71.14	1.63	73.17	70.30
True Qwen response	74.15	0.78	71.75	72.32

Table 13: Paired Reading contrasts restricted to the 64-target teacher assay. Differences are BA points. The same 10,000-draw class-stratified target-group procedure is applied within this subset.

Contrast	ΔBA	95% paired interval
Student − Logistic regression	$+8.11$	$[-1.22, 17.21]$
Student − Extra Trees	$+8.11$	$[-1.55, 17.55]$
True Qwen response − Logistic regression	$+11.11$	$[1.10, 20.65]$
True Qwen response − Extra Trees	$+11.11$	$[1.73, 19.88]$
Student − True Qwen response	-3.01	$[-10.54, 4.32]$

The student–teacher interval includes zero on the 64-target assay; the primary deployment estimand remains the 204-target panel.

D.2 Credit-g as a reconstruction-utility diagnostic

For Credit-g, the diagnostic response is the symmetric mean $\bar{Q}$. Ridge reconstructs it with $R^2 = 0.948$, Pearson $r = 0.974$ and Spearman $r_s = 0.973$, while BA rises from 64.57% for raw LR to 65.43%, a 0.86-point gain. The true response with raw features gives 65.50%, and direct MLP32 reaches 65.71%. Faithful reconstruction does not imply a large task gain. Appendix E gives matched functional comparisons.

D.3 Summary metrics and paired baseline comparisons

Table 14: Practical deployment comparisons for Credit-g and NetFlow evaluation, BA (%). The behavioral student uses the prescribed task-specific student and head; NetFlow evaluation uses RBF. Support and target identities match across methods. Response donors are an additional student-training resource. Bank comparisons are in Appendix C.1.

Panel	Behavioral student	Extra Trees	TabICLv2	Other direct raw reference
Credit-g	65.43	63.86	62.07	65.71 (direct MLP32)
NetFlow evaluation	84.18	88.44	87.76	—

Table 15: Student systems, with metrics in percent. Teacher BA uses the same readout and full target panel with true responses substituted for student responses. Reading has teacher responses only on its 64-target assay, not all 204 targets; the later NetFlow evaluation has no target teacher responses.

Panel	Student BA	AUROC	Macro-F1	Teacher BA
Bank	63.61	64.23	63.06	64.53
Reading (204 targets)	73.20	76.15	73.05	n/a
Credit-g	65.43	70.23	65.35	65.50
NetFlow, canonical	82.04	88.61	82.00	81.39
NetFlow, evaluation	84.18	91.31	83.91	n/a

Balanced accuracy is the mean of the two class recalls. A method's score averages BA over its specified support/bank fits. For a contrast, correctness is first averaged over fits for each target, and the paired target bootstrap samples the same IDs for both methods. Shared class weights make this equivalent to averaging the per-fit BA contrast. Raw repetitions across banks add no fitted model. Inference conditions on the supports, banks, students and heads.

Table 16: Paired student-minus-baseline comparisons for Credit-g and NetFlow. Differences and bounds are percentage points. These target intervals use 2,000 draws conditional on fixed fits. The principal NetFlow intervals use 10,000 draws and are reported separately in the text.

Panel	Comparator	Difference (pp)	95% interval
Credit-g	Extra Trees	+1.57	[-1.07, +4.29]
Credit-g	TabICLv2	+3.36	[+0.89, +5.96]
NetFlow, canonical	Extra Trees	-1.32	[-2.28, -0.35]
NetFlow, canonical	TabICLv2	-0.43	[-1.48, +0.60]
NetFlow, evaluation	Extra Trees	-4.26	[-5.56, -3.01]
NetFlow, evaluation	TabICLv2	-3.59	[-4.93, -2.37]

On the separately allocated NetFlow panel, the principal fixed-protocol gain is $+3.48$ points [1.64, 5.35] for LR and $+0.73$ [-0.74, 2.21] for RBF. Table 16 gives additional same-panel baseline comparisons. Appendix I records their provenance and replay coverage; cached-control intervals in Appendix E retain their separate resampling specifications.

E Response structure, controls, and mechanism

E.1 The response family

The semantic-aligned pair (m_0, m_1) has linear components $Q = (m_1 + m_0)/2$ and $D = (m_1 - m_0)/2$. Exchanging assignments preserves Q and reverses D. Historical $P = 2D$ gives the equivalent margins $m_0 = Q - D$, $m_1 = Q + D$ by an algebraic change of coordinates. The fixed family is

$$\bar{Q}, \qquad \bar{D}, \qquad \overline{|D|}, \qquad \mathrm{RMS}(D) = \left(K^{-1} \sum_j D_j^2 \right)^{1/2}, \qquad D, \qquad |D|. \tag{8}$$

The first four are scalars; the final two retain probe identity. Activity, RMS and magnitude profiles are even nonlinear summaries of the antisymmetric component, not additional linear symmetry components. Bank and Reading use $2\bar{D}$, diagnostic Credit-g uses $\bar{Q}$, and NetFlow MAG64 is $2|D|$.

If no probe index is privileged a priori, requiring a linear scalar $f(D) = w^\top D$ to satisfy $f(\Pi D) = f(D)$ for every index relabeling implies $\Pi^\top w = w$. Swapping any two indices forces their weights to be equal, giving $f \propto \bar{D}$. This restricts the summary, not the physical or semantic probes, and imposes no requirement on coordinate signs.

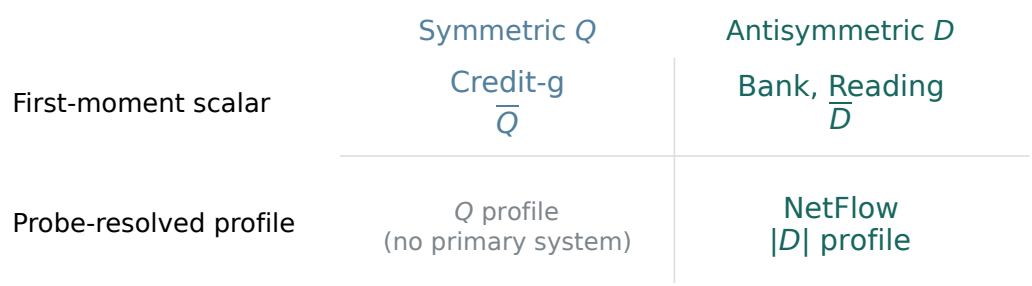


Figure 4: Positions of the response functionals used in the experiments within the paired response family; Reading uses the Bank scalar unchanged. The horizontal axis separates exchange-symmetric and exchange-antisymmetric components; the vertical axis separates first moments from probe-resolved profiles. NetFlow removes orientation from the antisymmetric profile.

E.2 Matched fitting and teacher assays

Comparisons preserve historical donors, supports, targets, feature exclusions and response banks. Students use donor responses without outcomes; each task head uses its own fifty-label support.

The response-only teacher assay standardizes each functional on the current support. Scalars use the fixed logistic head; profiles use fixed LR and RBF-SVM heads. These assays characterize the true response, not the deployed student. Separate teacher rows append the raw block for Credit-g scalars and NetFlow profiles, preserving the historical head input and preprocessing. Response-only teacher scores and raw-augmented student scores have different input blocks.

Bank and Credit-g scalar controls use exact Ridge recipes. Bank uses the predicted scalar alone, with one additional raw-plus-$\widehat{\widehat{D}}$ head. Credit-g retains raw-plus-scalar input for every control. NetFlow compares D and $|D|$ with the same MLP64 recipe, output dimension, donor split and LR/RBF heads; scalar collapses use Ridge. Only signed-versus-magnitude profiles match student capacity. The response-only teacher assay distinguishes scalar and profile retention before approximation.

The later NetFlow evaluation has responses for its historical donors but not its supports or targets. It contributes student-side structural and probe-count comparisons only, with no inferred teacher or target-reconstruction score.

Teacher results. BA, AUROC and macro-F1 are percentages, averaged over banks within support and then supports. Support SD is descriptive, in percentage points. Raw-block rows retain historical combined readouts; other rows are support-scaled response-only assays.

Bank.

Target	Readout	BA	AUROC	F1	Support SD
$\bar{Q}$	LR	47.47	47.60	47.05	14.80
$\bar{D}$	LR	64.53	64.32	64.11	0.41
$\overline{\lvert D\rvert}$	LR	50.02	46.44	36.41	0.35
$\mathrm{RMS}(D)$	LR	49.03	46.01	37.91	1.26
D	LR	50.90	49.74	49.24	8.31
D	RBF	50.69	55.23	37.11	1.07
$\lvert D\rvert$	LR	51.39	49.55	48.29	2.51
$\lvert D\rvert$	RBF	50.78	54.72	36.05	1.73

Credit-g.

Target	Readout	BA	AUROC	F1	Support SD
$\bar{Q}$	LR	60.54	63.85	60.48	0.52
$\bar{D}$	LR	50.68	52.08	50.41	3.86
$\overline{\lvert D\rvert}$	LR	50.64	52.21	50.24	4.06
$\mathrm{RMS}(D)$	LR	50.18	50.28	49.85	3.28
D	LR	58.07	60.84	57.87	2.55
D	RBF	60.50	64.86	59.90	1.52
$\lvert D\rvert$	LR	57.11	60.44	56.96	2.68
$\lvert D\rvert$	RBF	59.68	63.83	59.05	2.19
$\bar{Q}$	raw + LR	65.50	70.25	65.44	2.56
$\bar{D}$	raw + LR	64.39	69.43	64.31	2.14
$\overline{\lvert D\rvert}$	raw + LR	64.25	69.25	64.15	2.01
$\mathrm{RMS}(D)$	raw + LR	64.36	69.14	64.25	1.22

NetFlow canonical.

Target	Readout	BA	AUROC	F1	Support SD
$\bar{Q}$	LR	51.63	51.42	51.40	2.36
$\bar{D}$	LR	61.09	64.47	60.97	0.37
$\overline{\lvert D\rvert}$	LR	59.91	63.62	59.80	0.39
$\mathrm{RMS}(D)$	LR	55.53	58.07	55.29	1.79
D	LR	76.91	83.61	76.86	3.45
D	RBF	79.23	86.20	79.16	2.84
$\lvert D\rvert$	LR	77.23	84.05	77.17	3.13
$\lvert D\rvert$	RBF	79.37	86.24	79.29	2.78
D	raw + LR	79.27	85.55	79.23	3.24
D	raw + RBF	81.48	88.26	81.45	2.65
$\lvert D\rvert$	raw + LR	79.38	85.81	79.34	3.29
$\lvert D\rvert$	raw + RBF	81.39	88.19	81.35	2.55

Student results. Every row retains its stated student, head and raw block. Historical primary rows remain fixed even when a structural control scores higher. Aggregation is the same as for teacher results. NetFlow profiles use MLP64 and scalar collapses use Ridge.

Bank.

Target	Readout	BA	AUROC	F1	Support SD
$\bar{Q}$	LR (Ridge)	50.03	47.90	37.19	1.50
$\bar{D}$	LR (Ridge)	63.61	64.23	63.06	1.67
$\bar{D}$	raw + LR (Ridge)	55.47	61.91	46.69	4.36
$\overline{\lvert D\rvert}$	LR (Ridge)	50.38	53.45	36.10	1.22
$\mathrm{RMS}(D)$	LR (Ridge)	51.48	55.52	39.98	2.53

Credit-g.

Target	Readout	BA	AUROC	F1	Support SD
$\bar{Q}$	raw + LR (Ridge)	65.43	70.23	65.35	1.94
$\bar{D}$	raw + LR (Ridge)	64.57	69.87	64.49	2.42
$\overline{\lvert D\rvert}$	raw + LR (Ridge)	64.61	69.81	64.53	2.51
$\mathrm{RMS}(D)$	raw + LR (Ridge)	64.46	69.85	64.38	2.46

NetFlow canonical.

Target	Readout	BA	AUROC	F1	Support SD
$\lvert D\rvert$	raw + LR (MLP64)	79.70	85.15	79.67	2.39
$\lvert D\rvert$	raw + RBF (MLP64)	82.04	88.61	82.00	3.56
D	raw + LR (MLP64)	79.55	85.12	79.53	2.79
D	raw + RBF (MLP64)	81.93	88.59	81.90	2.94
$\bar{D}$	raw + LR (Ridge)	75.66	82.17	75.63	2.40
$\bar{D}$	raw + RBF (Ridge)	77.70	85.04	77.61	3.20
$\overline{\lvert D\rvert}$	raw + LR (Ridge)	75.73	82.18	75.70	2.37
$\overline{\lvert D\rvert}$	raw + RBF (Ridge)	77.69	85.06	77.59	3.30
$\mathrm{RMS}(D)$	raw + LR (Ridge)	75.53	82.16	75.50	2.31
$\mathrm{RMS}(D)$	raw + RBF (Ridge)	77.70	85.06	77.60	3.22

NetFlow evaluation.

Target	Readout	BA	AUROC	F1	Support SD
$\lvert D\rvert$	raw + LR (MLP64)	83.37	91.07	83.31	2.91
$\lvert D\rvert$	raw + RBF (MLP64)	84.18	91.31	83.91	2.90
D	raw + LR (MLP64)	83.80	91.30	83.76	2.39
D	raw + RBF (MLP64)	84.49	91.40	84.23	2.92
$\bar{D}$	raw + LR (Ridge)	80.06	88.07	79.39	3.98
$\bar{D}$	raw + RBF (Ridge)	83.84	90.99	83.37	2.72
$\overline{\lvert D\rvert}$	raw + LR (Ridge)	80.05	88.03	79.37	4.00
$\overline{\lvert D\rvert}$	raw + RBF (Ridge)	83.91	91.00	83.44	2.75
$\mathrm{RMS}(D)$	raw + LR (Ridge)	79.99	88.01	79.32	4.00
$\mathrm{RMS}(D)$	raw + RBF (Ridge)	83.64	90.95	83.15	2.96

Table 17: Bank head interaction at fixed response supervision. Both rows use the same Ridge prediction of the antisymmetric first moment and the same logistic settings. The second row appends donor-preprocessed raw attributes. Metrics are percentages; support SD describes the five observed supports.

Bank head input	BA	AUROC	F1	Support SD
Scalar only	63.61	64.23	63.06	1.67
Raw + scalar	55.47	61.91	46.69	4.36

Bank raw augmentation reaches 55.47% BA versus 63.61% for the scalar head: -8.14 points $[-11.51, -4.67]$. AUROC and macro-F1 are 61.91% and 46.69%, versus 64.23% and 63.06%; support SD rises from 1.67 to 4.36 points.

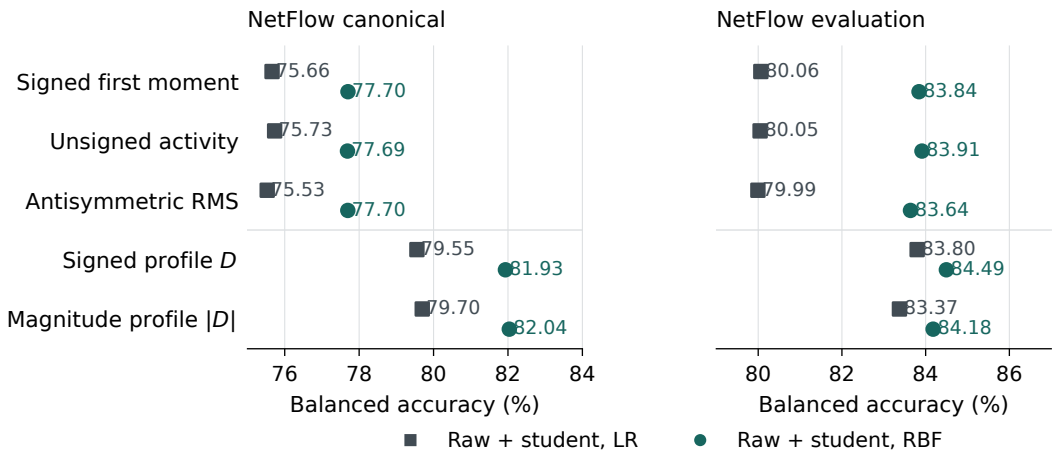


Figure 5: NetFlow structural controls on the canonical and separate evaluation panels. Heads retain their panel's raw block and fixed settings. Signed and magnitude profiles share the same MLP64 recipe within each panel; scalar collapses use Ridge. Thus the profile comparison matches student capacity, while the scalar comparison also changes it.

Paired comparisons. The response-functional comparison uses 10,000 target bootstrap draws with seed 28101. Each draw is shared across all supports, banks and comparisons within a panel. Bank uses epoch/class strata, Credit-g class strata, canonical NetFlow source/class strata and the separate evaluation source strata. Source-only draws recompute class denominators. Fits remain fixed; no multiplicity adjustment is made. Differences and bounds are percentage points. These structural intervals retain their own predictions and resampling specification; deployment and baseline comparisons are in Appendix D and Table 6.

Contrast	ΔBA	Lower	Upper
Bank			
$\bar{D}$/LR minus $\bar{Q}$/LR	+13.58	9.20	17.81
$\bar{D}$/LR minus $\overline{\lvert D\rvert}$/LR	+13.23	8.87	17.47
$\bar{D}$/LR minus $\mathrm{RMS}(D)$/LR	+12.14	8.02	16.13
raw + $\bar{D}$/LR minus $\bar{D}$/LR	−8.14	−11.51	−4.67
Credit-g			
raw + $\bar{Q}$/LR minus raw + $\bar{D}$/LR	+0.86	−0.25	1.96
raw + $\bar{Q}$/LR minus raw + $\overline{\lvert D\rvert}$/LR	+0.82	−0.32	1.96
raw + $\bar{Q}$/LR minus raw + $\mathrm{RMS}(D)$/LR	+0.96	−0.14	2.07
raw + $\bar{Q}$/LR minus Raw LR	+0.86	−0.25	1.96
raw + $\bar{Q}$/LR minus Direct MLP32	−0.29	−1.89	1.32
NetFlow canonical			
raw + $\lvert D\rvert$/LR minus raw + D/LR	+0.14	−0.10	0.39
raw + $\lvert D\rvert$/LR minus raw + $\bar{D}$/LR	+4.04	2.63	5.46
raw + $\lvert D\rvert$/LR minus raw + $\overline{\lvert D\rvert}$/LR	+3.97	2.56	5.38
raw + $\lvert D\rvert$/LR minus raw + $\mathrm{RMS}(D)$/LR	+4.17	2.77	5.59
raw + $\lvert D\rvert$/RBF minus raw + D/RBF	+0.10	−0.07	0.27
raw + $\lvert D\rvert$/RBF minus raw + $\bar{D}$/RBF	+4.34	2.81	5.90
raw + $\lvert D\rvert$/RBF minus raw + $\overline{\lvert D\rvert}$/RBF	+4.35	2.83	5.92
raw + $\lvert D\rvert$/RBF minus raw + $\mathrm{RMS}(D)$/RBF	+4.34	2.81	5.92
raw + $\lvert D\rvert$/RBF minus Raw RBF	+4.35	2.80	5.97
NetFlow evaluation			
raw + $\lvert D\rvert$/LR minus raw + D/LR	−0.42	−0.78	−0.07
raw + $\lvert D\rvert$/LR minus raw + $\bar{D}$/LR	+3.31	1.48	5.18
raw + $\lvert D\rvert$/LR minus raw + $\overline{\lvert D\rvert}$/LR	+3.33	1.49	5.21

Contrast	ΔBA	Lower	Upper
raw + $\lvert D\rvert$/LR minus raw + $\mathrm{RMS}(D)$/LR	$+3.38$	1.54	5.25
raw + $\lvert D\rvert$/RBF minus raw + D/RBF	-0.32	-0.55	-0.08
raw + $\lvert D\rvert$/RBF minus raw + $\bar{D}$/RBF	$+0.34$	-1.09	1.76
raw + $\lvert D\rvert$/RBF minus raw + $\overline{\lvert D\rvert}$/RBF	$+0.27$	-1.17	1.67
raw + $\lvert D\rvert$/RBF minus raw + $\mathrm{RMS}(D)$/RBF	$+0.54$	-0.90	1.98

On canonical NetFlow, magnitude minus signed profile is 0.14 points for LR $[-0.10, 0.39]$ and 0.10 for RBF $[-0.07, 0.27]$. The later evaluation slightly favors signed profiles: -0.42 $[-0.78, -0.07]$ and -0.32 $[-0.55, -0.08]$. Neither comparison changes the historical magnitude system. Intervals containing zero do not establish equivalence; scalar and profile students differ in capacity.

Reconstruction. Target responses are used without outcomes. Scalar R^2 uses target variance; profile R^2 averages nonconstant coordinates. Profile Pearson and Spearman flatten response matrices; cosine averages records. Standardized MSE divides errors by the response scales fitted on permitted training donors. MSE in D units is one quarter of MSE in historical P units; R^2, standardized MSE and standardized classification inputs are unchanged. Later NetFlow target responses are unavailable and excluded.

Bank.

Target	R^2	Pearson	Spearman	Cosine	MSE	Scaled MSE
$\bar{Q}$	-0.956	0.869	0.880	n/a	6.708	11.471
$\bar{D}$	0.695	0.879	0.898	n/a	0.687	4.189
$\overline{\lvert D\rvert}$	-0.861	-0.257	-0.294	n/a	0.973	19.647
$\mathrm{RMS}(D)$	-0.716	-0.216	-0.268	n/a	1.554	26.494

Credit-g.

Target	R^2	Pearson	Spearman	Cosine	MSE	Scaled MSE
$\bar{Q}$	0.948	0.974	0.973	n/a	1.060	0.057
$\bar{D}$	0.744	0.867	0.867	n/a	0.209	0.302
$\overline{\lvert D\rvert}$	0.734	0.861	0.858	n/a	0.180	0.314
$\mathrm{RMS}(D)$	0.664	0.820	0.806	n/a	0.298	0.402

NetFlow canonical.

Target	R^2	Pearson	Spearman	Cosine	MSE	Scaled MSE
$\lvert D\rvert$	0.669	0.881	0.878	0.964	0.094	0.338
D	0.676	0.883	0.881	0.963	0.097	0.331
$\bar{D}$	0.605	0.781	0.796	n/a	0.017	0.409
$\overline{\lvert D\rvert}$	0.612	0.785	0.802	n/a	0.015	0.397
$\mathrm{RMS}(D)$	0.575	0.761	0.762	n/a	0.024	0.406

E.3 Outcome-free donor response geometry

All geometry quantities use permitted donor responses without outcomes, separately for each support and bank before fitting heads. Signed coherence is $\lvert\bar{D}\rvert/(\mathrm{RMS}(D)+\epsilon)$. Q and D energies are mean squared margins in shared units, not independently standardized coordinates. Residual profile energy removes each record's probe mean: it measures what scalar averaging discards, not whether the discarded part predicts outcomes.

For effective rank, response matrices are centered across donors. If p_i are normalized squared singular values, entropy rank is $\exp(-\sum_i p_i \log p_i)$ and participation ratio is $(\sum_i p_i^2)^{-1}$. The table averages donor exclusions within each bank; residual columns give fractions of squared energy outside the per-record common mode, and rank is entropy rank.

Panel	Bank	Coherence	D residual	Q residual	Rank(D)	Rank(Q)
Bank	0	0.194	0.933	0.349	11.20	5.21
Bank	1	0.176	0.940	0.400	10.55	5.83
Bank	2	0.426	0.743	0.103	10.94	5.48
Credit-g	0	0.795	0.349	0.784	23.18	4.90
Credit-g	1	0.793	0.344	0.674	21.21	4.26
NetFlow canonical	0	0.824	0.319	0.468	12.01	4.28
NetFlow canonical	1	0.820	0.324	0.476	13.14	4.40
NetFlow canonical	2	0.824	0.339	0.430	9.63	4.46
NetFlow evaluation	0	0.824	0.318	0.458	12.43	4.41
NetFlow evaluation	1	0.819	0.324	0.469	13.35	4.52
NetFlow evaluation	2	0.822	0.336	0.422	9.97	4.61

The Bank signed mean is useful despite substantial residual profile energy. Its interpretation rests on the fixed scalar comparison and permutation symmetry, not a claim that probes share one sign.

E.4 Direct, one-sided, random, and shuffled Bank controls

Fixed comparison. The ablation uses the same Gemma-3-12B responses, five disjoint fifty-label supports, 384 targets and three 64-probe banks as the main Bank result. The other 200 pool records are donors. Each scalar uses donor-fitted raw preprocessing and response scaling, Ridge ($\alpha = 1$), and the fixed scalar-only logistic head in Appendix B. The teacher comparison uses that signal directly with the same support-fitted head. No outcome enters student fitting, and no ablation replaces the polarity predictor.

Controls are the shared level $\bar{Q}$, one-sided means $\bar{m}_0, \bar{m}_1$, the direct candidate margin, one fixed projection of P, and donor-shuffled polarity. AverageQ uses both assignments; it is a shared-level control, not cheaper one-sided acquisition. The direct margin averages the two semantic-aligned answer mappings for the candidate itself and is distilled and calibrated by the same Ridge/scalar head. Earlier direct Gemma baselines used different prediction or calibration procedures.

For bank b, one prespecified Rademacher vector $a_b \in \{-1, +1\}^{64}$ defines

$$r_b(x) = 64^{-1/2} \sum_{j=1}^{64} a_{bj} P_j(x). \tag{9}$$

The projection is not selected among alternatives. Five fixed shuffles independently permute donor polarity within each support/bank condition, leaving raw donors, targets and head recipes unchanged; each head refits on its support. Shuffling defines no teacher function on targets, so its teacher BA is undefined. Its reconstruction compares the resulting student with original unpermuted target polarity.

Table 18: Full Bank teacher-signal results. BA is percent. Static is a secondary vector control, not one of the scalar contrasts. The shuffled student averages all five fixed seeds; no artificial teacher score is assigned to its targets.

Teacher target	Teacher BA	Student BA
Signed first moment $\bar{P}$	64.53	63.61
AverageQ	47.47	50.03
Label-0 mean	50.87	49.88
Label-1 mean	47.29	49.83
Direct Gemma margin	50.89	50.68
Fixed random projection of P	54.24	55.31
Shuffled polarity	n/a	50.84
Frozen static representation	50.62	50.00

The static control uses the fixed 3,840-dimensional neutral final hidden state with row-L2 normalization. Multi-output Ridge predicts that map, and the head receives row-normalized predicted

coordinates. No layer, pooling rule, dimension or normalization is chosen from target accuracy. It remains a secondary vector control.

Table 19: Paired student minus each scalar control, in BA percentage points. The 10,000 draws resample target IDs within the six epoch/class strata, with the same draw for every support, bank and shuffle seed. These are descriptive intervals on the studied Bank panel, conditional on fixed fits and unadjusted for multiple comparisons.

Comparator student	Paired advantage	Lower 95%	Upper 95%
AverageQ	+13.58	+9.29	+17.86
Label-0 mean	+13.73	+9.43	+18.02
Label-1 mean	+13.78	+9.13	+18.44
Direct Gemma margin	+12.93	+8.98	+16.84
Fixed random projection of P	+8.30	+5.50	+11.11
Shuffled polarity	+12.77	+8.53	+17.02

All six lower bounds are positive; the smallest advantage is over the single random projection. This supports the signed first moment within the tested controls without requiring a common sign across probe coordinates.

Table 20: Target reconstruction in the Bank ablation. R^2, Pearson and Spearman are averaged over fits. Standardized MSE uses donor-fitted teacher scales. Static R^2 averages nonconstant coordinates and Pearson flattens the matrix; scalar Spearman is not applicable. Shuffled fidelity refers to original, unpermuted polarity.

Teacher target	R^2	Pearson	Spearman	MSE	Std. MSE
Signed first moment $\bar{P}$	0.695	0.879	0.898	2.750	4.189
AverageQ	-0.956	0.869	0.880	6.708	11.471
Label-0 mean	-6.114	0.600	0.651	5.093	8.736
Label-1 mean	0.095	0.901	0.916	9.699	10.594
Direct Gemma margin	0.059	0.799	0.802	7.641	14.497
Random projection	-0.937	0.692	0.744	36.717	12.996
Shuffled polarity	-0.771	-0.010	-0.011	16.066	24.620
Frozen static representation	-36.339	0.975	n/a	1.33×10^{-5}	23.879

Correlation alone does not imply accurate reconstruction: a shifted or rescaled prediction can have high correlation and negative R^2. The static vector's high flattened correlation coexists with poor coordinate reconstruction and chance-level student BA. Neither metric alone establishes utility.

E.5 Bank support-null mechanism and calibration

These diagnostics retain the five historical fifty-label supports, three banks, shared 384 targets and 200 response donors per condition, using Appendix B's recipe.

Table 21: Bank restriction diagnostics, BA (%), averaged over the five historical supports and three banks on the shared target panel. Erasure and reflection preserve support logits and retain the original scaler and head. The final row is a diagnostic decomposition of the fitted augmented score.

Historical system or diagnostic	Bank BA
Original scalar-only student	63.61
Support-null coefficient component erased	56.98
Support-null coefficient component reflected	55.38
Raw + scalar	55.47
Scalar-aligned component of augmented score	63.66

The affine student in original response units. Fix a support and bank. Let $T(x) \in \mathbb{R}^d$ be the donor-fitted encoding and μ_F, s_F the donor response mean and SD. Ridge fits $\widetilde{w}^\top T(x) + \widetilde{a}$ to $(F(x) - \mu_F)/s_F$, minimizing summed squared donor residuals plus $\|\widetilde{w}\|^2$ ($\alpha = 1$, unpenalized intercept). Inverting the scaler gives

$$S(x) = w^\top T(x) + a, \qquad w = s_F \widetilde{w}, \qquad a = s_F \widetilde{a} + \mu_F. \tag{10}$$

The diagnostics use $F = \bar{D}$; historical $\bar{P} = 2\bar{D}$ gives the same support-standardized coordinate and predictions. Interventions use inverse-scaled coefficients.

Projection and support preservation. Let X_S be the $50 \times d$ support encoding, $\bar{t}$ its row mean and $U = X_S - \mathbf{1}\bar{t}^\top$. The thin SVD $U = L\Sigma V^\top$ is computed in float64, retaining singular values $\sigma_j > \tau$, with $\tau = \max(50, d)\epsilon_{64}\sigma_1$ and $\epsilon_{64} = 2^{-52}$. For retained right singular vectors V_r,

$$\Pi = V_r V_r^\top = U^+ U, \qquad w_\parallel = \Pi w, \qquad w_\perp = (I - \Pi)w. \tag{11}$$

Here U^+ uses the same numerical rank. With $\bar{S} = 50^{-1} \sum_{i \in S} S(x_i)$, the fixed interventions are

$$S_\lambda(x) = \bar{S} + [T(x) - \bar{t}]^\top (w_\parallel + \lambda w_\perp), \qquad \lambda \in \{1, 0, -1\}. \tag{12}$$

The values denote original, erasure and reflection. An exact row-space projector gives $Uw_\perp = 0$, hence

$$S_\lambda(x_i) - S_1(x_i) = (\lambda - 1)[T(x_i) - \bar{t}]^\top w_\perp = 0 \qquad (i \in S).$$

The original support response scaler and logistic head are retained, preserving support logits, probabilities and likelihood. Across fifteen conditions, maximum absolute support-logit changes are 4.885×10^{-15} for erasure and 9.770×10^{-15} for reflection. Evaluation of Equation 12 verifies $\max_x |S(x) - S_1(x)| < 10^{-10}$ over available rows.

Table 22: Bank diagnostics by support, averaging BA percent over three banks on the shared targets. d/r gives encoded dimension and centered support rank. Erasure and reflection retain the scaler and head. "Aligned" is a diagnostic component of the fitted augmented logit (Equation 14).

Support	d/r	Original	Erased	Reflected	Raw + scalar	Aligned
0	54/32	64.76	50.00	50.00	58.33	64.84
1	53/33	60.85	57.73	52.34	58.33	60.76
2	54/32	63.19	50.09	49.83	50.00	63.45
3	54/34	64.58	63.28	61.89	51.48	64.76
4	54/34	64.67	63.80	62.85	59.20	64.50
Mean		63.61	56.98	55.38	55.47	63.66

The loss varies by support. The intervention identifies useful inherited coefficients that current-support outcome likelihood does not identify, not coefficients unknowable from raw data generally. Donor predictions also change, so response-fitting data distinguish these functions.

Regularization under raw augmentation. Let $\bar{S}, s_S$ be the support mean and SD used by the scalar head. Define

$$\xi(x) = \begin{bmatrix} T(x) \\ 1 \end{bmatrix}, \qquad v = \begin{bmatrix} w/s_S \\ (a - \bar{S})/s_S \end{bmatrix}, \qquad A_B(S(x)) = v^\top \xi(x).$$

The raw-plus-scalar logit is $\beta^\top \xi + \gamma v^\top \xi = \theta^\top \xi$, with $\theta = \beta + \gamma v$. The liblinear head uses $C = 1$ and default intercept scaling 1, penalizing the intercept as a constant-feature coefficient. Its objective sums logistic losses plus $(\|\beta\|^2 + \gamma^2)/(2C)$; balanced weights equal one on these 25/25 supports. Eliminating the redundant parameterization gives

$$\min_{\beta + \gamma v = \theta} (\|\beta\|^2 + \gamma^2) = \theta^\top (I + vv^\top)^{-1} \theta = \|\theta\|^2 - \frac{(v^\top \theta)^2}{1 + \|v\|^2}. \tag{13}$$

The minimum occurs at $\gamma = (v^\top \theta)/(1 + \|v\|^2)$ and $\beta = \theta - \gamma v$. Augmentation reduces the penalty along v but allows any raw coefficient vector; it does not impose the scalar-only restriction. This includes the intercept penalty. Fitted penalties agree with the right-hand side within 3.56×10^{-15}.

Decomposing the fitted augmented score. Let $\ell_{\mathrm{sc}}, \ell_{\mathrm{aug}}$ be the scalar and augmented logits. On the support only, form $H = [\mathbf{1}, \ell_{\mathrm{sc}}(X_S)]$ and obtain $c = H^{+}\ell_{\mathrm{aug}}(X_S)$ by least squares. Extend these two coefficients to targets:

$$g(x) = c_0 + c_1\ell_{\mathrm{sc}}(x), \qquad e(x) = \ell_{\mathrm{aug}}(x) - g(x), \qquad H^{\top}e(X_S) = 0. \tag{14}$$

Orthogonality concerns support score vectors, not raw coefficients. At threshold zero, g gives 63.66% BA versus 55.47% for the full augmented logit. Most loss is associated with the additional score component, rather than different scalar calibration. No target outcome fits c or selects the threshold.

The exactly matched raw head uses the same donor encoder and $C = 1$, reaching 54.58%; adding the scalar raises this to 55.47%. The stronger 57.50% headline comparator uses support-fitted preprocessing, row L2 normalization and support-only CV over C (Appendix C). The loss concerns relaxing the scalar restriction, not adding behavior to that exactly matched raw head.

Ranking, orientation and calibration. For a scalar f, the strictly monotone head $\sigma(\eta f + b)$ preserves AUROC when $\eta > 0$ and reverses it to $1 - \mathrm{AUROC}(f)$ when $\eta < 0$. The direct teacher margin has fixed semantic orientation and target AUROC 64.77%. Its slope is negative on supports 0 and 2 and positive on 1, 3 and 4. It does not depend on the probe bank: these are five orientations, not fifteen independent results. Mean calibrated AUROC is

$$0.6(0.6476508247) + 0.4(1 - 0.6476508247) = 0.5295301649.$$

Calibrated teacher BA is 50.89%; the matched direct-margin student's BA is 50.68%. The signed first moment has a negative slope on every support and bank for teacher and student, with more useful calibrated ranking and threshold transfer under this procedure.

Table 23: Bank ranking and transferred calibration. Metrics are percentages, averaged over the historical conditions. "Fixed AUC" uses the original scalar orientation, without target-based sign correction. $\Delta\mu/s$ is the mean target-minus-support location shift divided by the support scalar SD (population divisor). Calibrated metrics retain the original support-fitted heads.

Target	Representation	Fixed AUC	Calibrated AUC	BA	$\Delta\mu/s$
Direct margin	Teacher	64.77	52.95	50.89	3.865
	Student	64.76	52.68	50.68	5.897
Signed first moment	Teacher	35.68	64.32	64.53	0.989
	Student	35.77	64.23	63.61	0.355
Symmetric first moment	Teacher	38.02	47.60	47.47	0.304
	Student	36.66	47.90	50.03	−2.868

The location shifts describe calibration difficulty without establishing its cause. Low calibrated BA does not establish absence of target information.

E.6 Raw-feature and resource-matched auxiliary controls

These controls retain five historical supports, fifty labels per head, 200 donors per support, three banks and 384 targets. Duration remains excluded. Every auxiliary scalar uses donor-fitted mixed-data encoding, donor-only target standardization, Ridge ($\alpha = 1$), support standardization of the predicted scalar and balanced $C = 1$ liblinear LR. Only the regression target changes; the behavioral target remains the signed first moment.

PCA-1 uses the first component of the centered encoded donor matrix, the full deterministic solver and no outcome-based sign choice. Eight raw-random targets are $r_k^{\top}T(x)$, with independent Rademacher entries divided by $\sqrt{d}$. These project raw features, unlike Equation 9's response projection.

The autoencoder is $d \to 16 \to 1 \to 16 \to d$, with ReLU hidden layers, linear bottleneck and output, MSE and Adam. Learning rate is 10^{-3}, batch size 32, maximum 1,000 epochs and patience 50. A deterministic 160/40 donor split selects the best reconstruction-validation epoch. Its bottleneck supplies the regression target; deployment still uses Ridge and the scalar head. Neither fitting nor stopping uses outcomes.

Table 24: Non-LLM Bank auxiliary controls. BA, support SD, AUROC and macro-F1 are percentages; Δ is behavioral student minus row BA in points. R^2 compares the student with its own auxiliary target. The headline LR comparison is in Table 6. Intervals use a paired 10,000-draw target bootstrap, seed 28101, within the six historical epoch/class strata.

Auxiliary target	BA	SD	AUROC	F1	R^2	Δ	95% interval
Behavioral $\bar{D}$	63.61	1.67	64.23	63.06	0.695	0.00	$[0.00, 0.00]$
PCA-1	50.00	0.00	57.23	33.33	0.99995	13.61	$[9.29, 17.95]$
Raw random (eight)	50.15	1.82	51.05	37.61	0.357	13.46	$[9.29, 17.64]$
AE bottleneck	51.67	3.73	52.13	37.97	0.633	11.94	$[8.18, 15.75]$

PCA reconstruction is essentially exact but calibrated BA is 50.00%. Neither it nor the other tested raw targets reproduces the behavioral gain. No projection is selected: the eight support-averaged BAs have mean 50.15%, median 50.83%, SD 2.11 points and range [47.34, 52.76]%. These are separate readouts, not an ensemble.

Donor- and probe-matched representations. ProbeSim64 and DAE16 use the same 200 raw donors per support and their fitted encoding T. ProbeSim64 also uses each historical 64-record probe bank, with coordinates $\exp(-\|T(x) - T(u_j)\|^2/(2\sigma^2))$, where σ is the median nonzero donor-to-probe Euclidean distance. DAE16 reconstructs clean donor encodings through one 16-unit ReLU layer and a linear output, independently masking 20% of inputs to zero. A deterministic 160/40 donor split supplies reconstruction validation. Adam uses learning rate 10^{-3}, penalty 10^{-4}, batches of 32, at most 600 epochs, patience 40 and tolerance 10^{-5}. The validation mask is fixed and best reconstruction weights are restored.

Each representation is tested alone and with raw encoding. All four heads use balanced liblinear LR, $C \in \{0.03, 0.3, 1, 3, 30\}$, three stratified support folds, mean-fold BA selection and first-entry tie breaking. Coordinate standardization fits the training fold; final scaling and fitting use all fifty labels. There is no row normalization. Targets enter neither representation fitting nor selection.

Table 25: Resource-matched Bank representations on the same five supports and 384 targets. Metrics are percentages; Δ is behavioral student minus row BA in points. ProbeSim64 metrics average three banks within each support. The paired 10,000-draw target intervals use seed 28101 and the six historical epoch/class strata, with fits held fixed.

Representation	BA	AUROC	Macro-F1	Δ	95% interval
Behavioral student	63.61	64.23	63.06	0.00	$[0.00, 0.00]$
ProbeSim64	52.17	54.81	48.33	11.44	$[8.04, 14.81]$
Raw + ProbeSim64	52.86	57.38	49.22	10.75	$[7.50, 13.96]$
DAE16	53.33	57.80	41.28	10.28	$[6.81, 13.75]$
Raw + DAE16	52.97	52.03	43.79	10.64	$[7.22, 13.94]$

These controls use the same raw donor budget, with ProbeSim64 additionally using the same raw probe records. None approaches the behavioral student's 63.61% BA.

E.7 Matched support-null controls

Equation 12's decomposition and intercept adjustment apply to the behavioral, PCA and eight raw-random Ridge directions. No scaler or head is refitted. Table 26 averages the fixed fitting conditions. The largest support-logit change is 1.58×10^{-14}; the behavioral reflection change is 1.02×10^{-14}, consistent with the earlier 9.770×10^{-15} measurement up to floating-point precision.

Table 26: Matched support-null interventions. BA, AUROC and macro-F1 are percentages. BA loss is original minus intervened BA in points; negative loss is improvement. Intervals condition on the fixed fits and use the paired target bootstrap described above. The final column is the maximum absolute support-logit change.

Target	Intervention	BA	AUROC	F1	BA loss	95% interval	$\max \lvert\Delta\ell\rvert$
Behavioral $\bar{D}$	Original	63.61	64.23	63.06	0.00	$[0.00, 0.00]$	0
Behavioral $\bar{D}$	Erased	56.98	59.84	50.07	6.63	$[4.57, 8.65]$	4.88×10^{-15}
Behavioral $\bar{D}$	Reflected	55.38	51.16	48.29	8.23	$[5.78, 10.63]$	1.02×10^{-14}
PCA-1	Original	50.00	57.23	33.33	0.00	$[0.00, 0.00]$	0
PCA-1	Erased	50.00	56.78	33.33	0.00	$[0.00, 0.00]$	5.55×10^{-16}
PCA-1	Reflected	50.00	55.47	33.69	0.00	$[-0.16, 0.16]$	1.33×10^{-15}
Raw random (eight)	Original	50.15	51.05	37.61	0.00	$[0.00, 0.00]$	0
Raw random (eight)	Erased	50.40	52.76	36.62	-0.25	$[-0.53, 0.02]$	7.77×10^{-15}
Raw random (eight)	Reflected	51.20	53.04	40.60	-1.05	$[-1.46, -0.62]$	1.58×10^{-14}

Erasure and reflection leave PCA's mean BA at 50.00%. The raw-random mean increases from 50.15% to 50.40% and 51.20%, unlike the behavioral losses of 6.63 and 8.23 points. Among these auxiliary targets, support-null BA loss is specific to the behavioral direction.

A broader empirical null. The same auxiliary-target procedure covers 256 prespecified Rademacher raw directions and the first ten donor-only PCA directions. Random entries are divided by $\sqrt{d}$; PCA uses the full deterministic solver with no outcome-based sign orientation. Each scalar passes through the same donor-standardized Ridge ($\alpha = 1$) and support-standardized scalar head before erasure/reflection. Supports, donors and targets remain fixed; scalers and heads are not refitted after intervention.

Table 27: Empirical support-null distributions. Each random or PCA direction first averages its five support-specific metrics; the behavioral reference averages five supports and three banks. Original BA is percent, and losses are original minus intervened BA in points. Quantiles are across directions, not target intervals.

Direction family	Quantity	Median	5th	95th	Maximum
Random (256)	Original BA	51.15	48.01	54.43	55.99
	Erasure loss	0.00	-1.94	2.55	4.01
	Reflection loss	0.05	-3.35	3.18	5.52
PCA (10)	Original BA	49.95	47.05	59.86	62.55
	Erasure loss	-0.31	-3.24	3.71	5.21
	Reflection loss	0.08	-4.03	4.51	6.20
Behavioral	Original BA		63.61		
	Erasure loss		6.63		
	Reflection loss		8.23		

None of the 256 random directions or ten PCA directions reaches the behavioral original BA (63.61%), erasure loss (6.63 points) or reflection loss (8.23 points). The behavioral effect is unusual within this prespecified matched family. The maximum absolute support-logit change is 1.155×10^{-14}, below the fixed 10^{-10} tolerance.

E.8 Second-teacher Bank measurement and amortization

The existing Llama-3.1-8B-Instruct cache has complete paired responses for the same 250 pool records, 384 targets and three 64-probe banks. The signed functional, five supports, 200 donors per support, Ridge ($\alpha = 1$), scalar head and target panel remain fixed. No new responses, alternative functional or student setting are tried.

Table 28: Llama Bank by probe bank, averaging the five supports. T/S denotes the true-response teacher and its fixed Ridge student. BA, AUROC and positive-prediction fractions are percentages. Fidelity compares target predictions z with responses f; ρ is Pearson correlation, $\delta = (\mu_z - \mu_f)/s_f$, and s_z/s_f uses target population SDs. The final row averages all fifteen fits.

Bank	BA T/S	AUROC T/S	R^2/ρ	δ	s_z/s_f	Positive % T/S
0	61.56 / 56.67	66.56 / 63.62	$-0.923/0.890$	$+1.230$	1.024	54.2 / 18.1
1	63.28 / 53.23	66.35 / 58.01	$-0.061/0.820$	$+0.678$	0.960	46.9 / 31.8
2	64.01 / 61.04	67.68 / 64.28	$0.148/0.882$	$+0.601$	1.015	49.0 / 36.9
All	62.95 / 56.98	66.86 / 61.97	$-0.279/0.864$	$+0.837$	1.000	50.0 / 28.9

The Llama teacher reaches 62.95% BA, 5.45 points above raw LR [2.19, 8.70]. The student reaches 56.98%: -0.52 [-2.10, 1.06] versus raw LR and -5.97 [-9.15, -2.74] versus its teacher. Teacher/student macro-F1 is 62.86/53.11%; support SD is 0.43/3.46 points and bank SD 1.26/3.92. The teacher retains a useful signed response on the same panel, separating measurement utility from amortization fidelity.

Table 29: Llama Bank by historical support, averaging three banks. BA and Δ = student minus raw LR are percentages and percentage points, respectively. Reconstruction metrics refer to the Llama student. All five supports are retained.

Support	Teacher BA	Student BA	Raw LR	Δ	R^2	Pearson	Spearman
0	62.59	60.42	50.00	10.42	-0.362	0.912	0.913
1	62.76	59.64	65.89	-6.25	-0.173	0.687	0.707
2	63.37	53.65	65.89	-12.24	-1.318	0.905	0.908
3	63.45	58.25	55.73	2.52	-0.017	0.918	0.920
4	62.59	52.95	50.00	2.95	0.476	0.897	0.900

The student retains Pearson 0.864 and Spearman 0.870 despite $R^2 = -0.279$. Its exact target-response error decomposition is

$$\frac{\mathrm{MSE}(z,f)}{s_f^2} = \left(\frac{\mu_z - \mu_f}{s_f}\right)^2 + \left(\frac{s_z}{s_f} - 1\right)^2 + 2\frac{s_z}{s_f}(1-\rho).$$

Averaging fifteen fits gives 0.949 for location, 0.031 for scale mismatch and 0.299 for imperfect correlation: 74.2%, 2.4% and 23.4% of mean variance-normalized response error. These fractions do not apportion BA loss. Mean student response shifts upward by 0.837 teacher target SDs, while teacher/student decision thresholds stay close at 2.718/2.714 in $\bar{P} = 2\bar{D}$ units. With mostly negative slopes, the student predicts 28.9% positive targets versus 50.0% for the teacher.

Location mismatch is not the whole explanation. All fourteen same-sign student heads have lower AUROC than their teachers. On support 4, bank 1, the student slope is positive and teacher slope negative: student BA is 41.15% and AUROC 37.33%, despite response correlation 0.836. Location, orientation and ranking errors accompany the utility loss. No repair or recalibration is evaluated.

E.9 Automatic response-readout selection

The secondary selector exposes six deployment families from each cached $K = 64$ paired response: Raw, P_{mean}, Q_{mean}, P_{64}, Q_{64} and $\mathrm{MAG64} = |P|_{64}$. The means average the probe coordinates; profiles retain them. Raw uses no response supervision. Every response family uses Ridge ($\alpha = 1$), including profiles, rather than the historical NetFlow MLP. Raw preprocessing and coordinatewise response scaling fit only permitted donors, excluding the entire current support and all targets. The student predicts support and target responses, then inverts the donor response scaling. Response heads receive predicted responses alone, not appended raw features or true support responses. Historical primary systems remain unchanged.

Support-only choice. Each head is balanced logistic regression with `liblinear`, 5,000 maximum iterations, random seed 0 and zero decision-score threshold. Readout and $C \in \{0.03, 0.3, 1, 3, 30\}$

are chosen jointly. All thirty candidates share three stratified support folds, with shuffling and seed $27001 + s$ for zero-based support index s. Raw preprocessing fits the fold's training rows; response-head standardization also fits within the fold. Selection minimizes binary log loss pooled over one OOF probability per support row, assigning total weight $1/2$ to each class. Probabilities are clipped to $[10^{-6}, 1 - 10^{-6}]$ and logarithms are natural. It is not the mean of fold losses, and BA does not select candidates. Losses within 10^{-12} tie in the order Raw, P_{mean}, Q_{mean}, P_{64}, Q_{64}, MAG64, then the stated C order. The selected pipeline is refitted on all fifty support labels.

One preserved Diabetes support has 49 negatives and one positive. Its approved exception uses Raw at $C = 1$, with CV unavailable, retaining both banks. Fixed-family diagnostic heads also use $C = 1$ without CV for this support. No support is redrawn. All target predictions and selection records are saved and hashed before target scoring. All banks are retained; metrics average banks within support and then supports, never predictions.

Panels and comparator recipes. Bank uses five supports, 200 donors per condition, three banks and 384 targets; Credit-g uses five, 200, two and 280, respectively. Canonical NetFlow uses ten supports, 450 donors, three banks and 768 targets. Reading, Adult, Diabetes, Spirometry and Crimes each retain five supports, 200 donors and two banks; their target counts are 204, 16,281, 1,024, 204 and 204. These are the existing role ledgers and cached teacher backbones in Table 30; no new panels or responses are acquired.

The run's historical Raw LR comparator is distinct from its Raw candidate and from the main paper's stronger or matched comparator recipes. Bank replays fixed $C = 1$ with support preprocessing and row-L2 normalization. Credit-g retains its original five-fold $C \in \{0.01, 0.1, 1, 10, 100\}$ search and OOF threshold. NetFlow retains its source–class CV and $C \in \{0.03, 0.3, 3, 30\}$ grid; its historical raw branch uses unlogged float32 measurements, while response students retain the prescribed count/duration $\log(1 + x)$ transform. The five other datasets reuse verified original Raw LR predictions. Thus the 56.35/63.14/75.47% Bank/Credit-g/NetFlow raw values here do not replace the main historical results. All selector heads instead use the same stated grid and loss criterion.

Dataset	Teacher	Raw LR	Auto	Δ	95% interval
Bank	Gemma-3	56.35	59.95	$+3.59$	$[+2.08, +5.09]$
Reading	Qwen3	66.09	69.80	$+3.71$	$[+1.25, +6.24]$
Credit-g	Gemma-3	63.14	62.50	-0.64	$[-3.11, +1.82]$
NetFlow	Llama-3.1	75.47	74.80	-0.67	$[-1.28, -0.07]$
Adult	Qwen3	75.67	74.33	-1.34	$[-1.78, -0.88]$
Diabetes	Qwen3	50.65	53.19	$+2.54$	$[+0.71, +4.46]$
Spirometry	Qwen3	81.27	80.74	-0.53	$[-1.78, +0.74]$
Crimes	GPT-4o-mini	58.29	66.29	$+8.00$	$[+3.62, +12.14]$
Mean	—	65.87	67.70	$+1.83$	—

Table 30: Automatic support-only readout selection from cached paired responses by pooled class-balanced OOF log loss. BA (%); Δ is Auto minus the run's historical raw-LR comparator in points, with paired 95% target-bootstrap intervals conditional on the fitted supports, banks and models. These comparator recipes do not replace the main deployment systems. The mean weights datasets equally and has no pooled interval. No new LLM responses. This subsection gives comparator recipes and controls.

Raw-only same-CV control. The automatic selector searches both the response readout and logistic regularization. To isolate response selection from ordinary hyperparameter tuning, we repeat the identical support-only procedure while restricting the candidate family to Raw and selecting only C. The control uses the same frozen folds, pooled class-balanced OOF log loss, clipping and tie rule, including the Diabetes exception. It reuses already locked Raw-candidate predictions with zero refits; bank copies are verified identical. Equal-dataset mean BA is 65.87% for historical Raw LR, 65.66% for Raw-only and 67.70% for Auto: Auto minus Raw-only is $+2.04$ points. Dataset differences are four positive, one tied and three negative (Table 31).

Table 31: Raw-only same-CV control. Historical Raw LR retains the original recipe recovered by this run; Raw-only selects only C using the selector's frozen folds and pooled class-balanced OOF log loss. BA is in percent; Δ and intervals are Auto minus Raw-only in points. The mean weights datasets equally and has no pooled interval.

Dataset	Historical Raw	Raw-only	Auto	Δ	95% interval
Bank	56.35	56.25	59.95	$+3.70$	$[+2.17, +5.23]$
Reading	66.09	64.64	69.80	$+5.16$	$[+2.73, +7.62]$
Credit-g	63.14	62.50	62.50	$+0.00$	$[-2.29, +2.29]$
NetFlow	75.47	76.20	74.80	-1.40	$[-1.97, -0.84]$
Adult	75.67	74.49	74.33	-0.16	$[-0.59, +0.29]$
Diabetes	50.65	50.68	53.19	$+2.51$	$[+0.49, +4.61]$
Spirometry	81.27	82.47	80.74	-1.73	$[-2.61, -0.89]$
Crimes	58.29	58.06	66.29	$+8.24$	$[+4.24, +12.15]$
Mean	65.87	65.66	67.70	$+2.04$	—

Uncertainty. Both contrast tables use 10,000 paired target resamples with seed 28101 and fixed fitted support–bank conditions. Bank retains epoch–class strata; NetFlow source–class strata; Credit-g class-stratified rows; Diabetes patients within class. Adult resamples intact identical-predictor clusters without stratification. Reading, Spirometry and Crimes resample intact predictor clusters within pure-negative, pure-positive and mixed-label strata; mixed-label clusters are never split. Target resamples are shared across compared methods. These conditional percentile intervals are distinct from support SD and have no multiplicity adjustment or equal-dataset pooled interval. Displayed gains are rounded from the saved differences, not subtracted from rounded BA columns.

E.10 Controlled teacher substitution on Adult

We substitute DeepSeek V4 Flash 0731 for the existing historical Qwen3-14B cache, retaining Adult row identities, the 250-record response pool, five 50-record supports (200 donors each), two 64-probe banks, 16,281 targets and the fixed 64-target assay. Features, preprocessing, serialization, task wording, class semantics and both A/B mappings are unchanged. Both systems use $P_{\mathrm{mean}} = \mathrm{mean}_j(m_{1j} - m_{0j})$ at $K = 64$, donor-standardized Ridge ($\alpha = 1$), support-standardized scalar responses and a balanced `liblinear` logistic head with $C = 1$. Tokenizer and chat framing follow each teacher.

Table 32: Controlled Adult teacher substitution with the data roles, response functional, Ridge student, and downstream head fixed. Response R^2 is averaged over the fixed support-bank conditions. BA and AUROC are on the frozen full target panel.

Teacher	Student BA	AUROC	Response R^2
Qwen3-14B	63.30	69.11	0.522
DeepSeek V4 Flash 0731	69.77	78.05	0.659

Changing the teacher raises distilled BA from 63.30% to 69.77%: $+6.47$ points [5.93, 7.02]. The 95% paired interval uses 10,000 unstratified whole-predictor-cluster bootstrap draws with fixed supports and banks; metrics average banks within support and then supports. On the 64-record assay, true-response BA rises from 66.22% to 73.61%. The DeepSeek response is more useful on this assay and more accurately reconstructed by the same Ridge student (Table 32); these supporting measurements do not causally decompose the full-target gain. Teacher choice materially affects the supervision supplied by the paired response.

DeepSeek predictions were saved and hashed before target labels were parsed; no outcome enters response-student fitting. Both systems require zero LLM calls at deployment.

F Robustness and resource sensitivity

F.1 Support and predetermined probe-bank variation

For each support, metrics are computed separately by bank and then averaged; predictions are not ensembled. SD uses observed supports with divisor $n - 1$. No support bootstrap, binomial test or support-level p-value is reported. Supports share their panel's targets.

Table 33: Descriptive support variation. BA and its standard deviation are percent; gains are percentage points against the same-panel raw comparator. The sign columns count positive, tied and negative support gains. Canonical and evaluation NetFlow panels remain separate.

Dataset / student head	n	BA: mean $\pm$ SD	BA range	Mean gain	+/tie/−
Bank / polarity	5	63.61 $\pm$ 1.67	60.85–64.76	+6.11	3/0/2
Credit-g / LR	5	65.43 $\pm$ 1.94	64.11–68.75	+0.86	4/0/1
NetFlow canonical / LR	10	79.70 $\pm$ 2.39	75.22–84.20	+4.05	9/0/1
NetFlow canonical / RBF	10	82.04 $\pm$ 3.56	72.87–85.29	+4.35	10/0/0
NetFlow evaluation / LR	5	83.37 $\pm$ 2.91	79.16–86.32	+3.48	5/0/0
NetFlow evaluation / RBF	5	84.18 $\pm$ 2.90	80.58–87.60	+0.73	3/0/2

Bank student BA is 63.61% with support SD 1.67, versus raw LR's 57.50% and SD 8.00. Table 34 retains all five contrasts against the unchanged headline LR recipe.

Table 34: The five Bank supports. Student BA averages its three banks; raw LR is the historical headline recipe described in Appendix C. Support IDs use the original zero-based indexing. Values are BA percent and student-minus-raw percentage points.

Support ID	Polarity student BA	Raw LR BA	Gain
0	64.76	50.00	+14.76
1	60.85	65.89	-5.03
2	63.19	65.89	-2.69
3	64.58	55.73	+8.85
4	64.67	50.00	+14.67

Canonical NetFlow LR gains average 4.05 points, positive on nine of ten supports; RBF gains average 4.35, positive on all ten. Evaluation LR gains average 3.48, positive on all five supports; RBF gains average 0.73, positive on three of five. Credit-g's modest gain is positive on four of five supports. These counts describe fitted conditions, not independent replications.

Table 35: Credit-g: support-level BA (%) and student-minus-raw gain (percentage points). Support IDs are the original zero-based indices. Bank metrics are averaged within support.

Support ID	Raw LR	Student LR	Δ LR
0	68.21	68.75	+0.54
1	62.86	65.54	+2.68
2	63.21	64.11	+0.89
3	64.64	64.29	-0.36
4	63.93	64.46	+0.54

Table 36: NetFlow canonical: support-level BA (%) and student-minus-raw gain (percentage points). Support IDs are the original zero-based indices. Bank metrics are averaged within support.

Support ID	Raw LR	Student LR	Δ LR	Raw RBF	Student RBF	Δ RBF
0	71.09	75.22	+4.12	77.99	84.11	+6.12
1	75.39	78.69	+3.30	77.34	82.51	+5.16
2	74.09	77.69	+3.60	70.70	72.87	+2.17
3	79.04	78.91	-0.13	80.47	80.60	+0.13
4	73.70	81.55	+7.86	78.78	83.72	+4.95
5	75.39	80.64	+5.25	79.04	82.42	+3.39
6	77.34	80.69	+3.34	79.56	82.94	+3.39
7	77.99	84.20	+6.21	78.78	85.29	+6.51
8	76.69	79.77	+3.08	80.21	84.90	+4.69
9	75.78	79.60	+3.82	73.96	80.99	+7.03

Table 37: NetFlow evaluation: support-level BA (%) and student-minus-raw gain (percentage points). Support IDs are the original zero-based indices. Bank metrics are averaged within support.

Support ID	Raw LR	Student LR	Δ LR	Raw RBF	Student RBF	Δ RBF
0	74.85	79.16	+4.30	83.60	80.58	-3.01
1	76.21	81.70	+5.49	79.72	81.86	+2.14
2	84.05	84.50	+0.45	85.40	84.93	-0.47
3	82.24	86.32	+4.08	83.09	85.92	+2.83
4	82.13	85.19	+3.06	85.42	87.60	+2.17

Predetermined banks. Each bank is evaluated separately at the historical probe count. Bank means average the same supports; their SD and range describe the available banks, not arbitrary future probe compositions. No best bank is selected. Values are percentages and spreads are percentage points.

System	Bank 0	Bank 1	Bank 2	SD	Min	Max
Bank	64.01	63.85	62.97	0.56	62.97	64.01
Credit-g	65.29	65.57	n/a	0.20	65.29	65.57
NetFlow canonical LR	79.49	80.40	79.19	0.63	79.19	80.40
NetFlow canonical RBF	81.45	82.04	82.62	0.59	81.45	82.62
NetFlow evaluation LR	83.09	83.64	83.39	0.27	83.09	83.64
NetFlow evaluation RBF	84.74	84.44	83.35	0.73	83.35	84.74

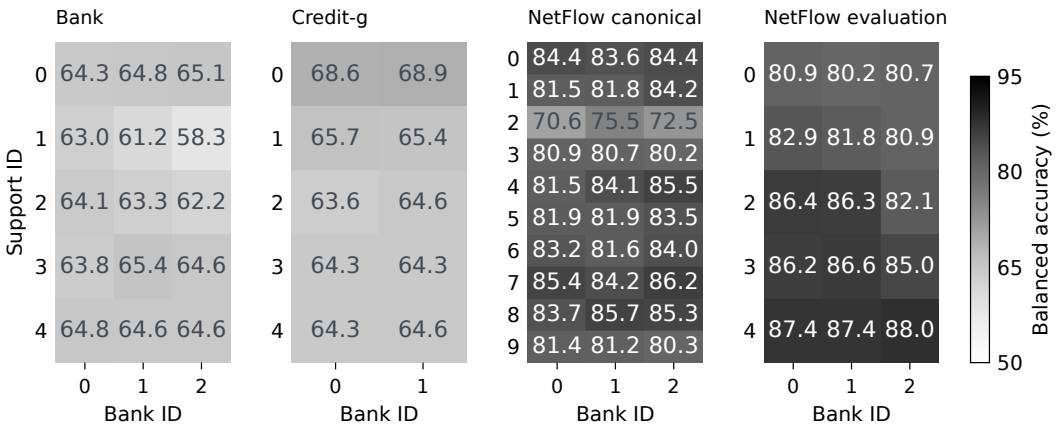


Figure 6: Historical student BA by support and predetermined probe bank. Cells are percentages; each support uses the same outcome-label budget. Bank identity is retained rather than selected. The spread describes the available banks and supports, not invariance to arbitrary probe composition.

F.2 Nested cached probe subsets

The fixed sensitivity grid is $K \in \{8, 16, 32, 64\}$. For each bank, eight predetermined permutations supply nested first-K subsets. Selected columns are restored to native cache order before fitting so coordinate permutation does not introduce an unrelated MLP initialization change.

Each subset refits the historical student for its prescribed functional: $\bar{D}$ on Bank, $\bar{Q}$ on Credit-g and magnitude profiles on both NetFlow panels. Ridge settings remain fixed. NetFlow retains hidden width, training recipe and both heads; only output width becomes K. Raw preprocessing does not change with the subset. No new response is acquired and no K is selected.

At $K = 64$, all permutations give the same native-ordered bank and reuse the primary fit. Subset SD is zero by construction, not zero support or bank uncertainty. At smaller K, SD is across eight subset means after averaging supports and banks. The tables separately give support SD; all-condition spread is retained in the detailed results. These are descriptive quantities for overlapping subsets.

Bank rises from 60.00% BA at $K = 8$ to 63.61% at $K = 64$, Credit-g changes little, and both NetFlow RBF systems are slightly higher at 32 than 64 (Figure 7).

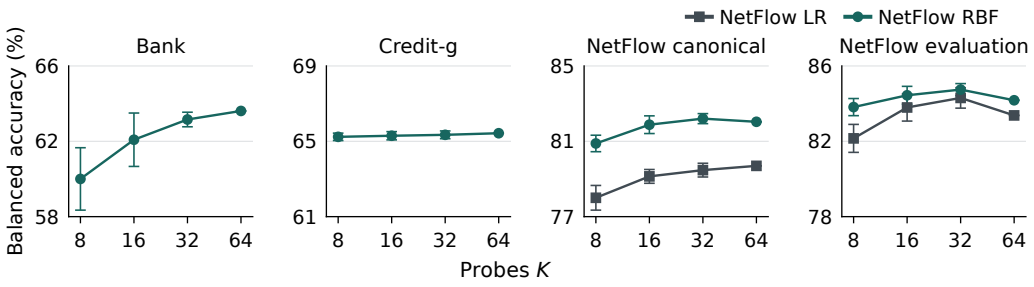


Figure 7: Probe-count sensitivity. All panels span eight BA points. Curves average fixed supports, banks, and eight outcome-independent nested subsets; bars show SD across subset means, not confidence intervals. At $K = 64$, the subsets coincide.

Bank (LR).

K	BA	Subset SD	Support SD	AUROC	F1	R^2
8	60.00	1.66	2.02	62.60	58.00	0.499
16	62.08	1.42	0.96	63.83	61.04	0.611
32	63.16	0.39	1.37	64.25	62.42	0.680
64	63.61	0.00	1.67	64.23	63.06	0.695

Credit-g (LR).

K	BA	Subset SD	Support SD	AUROC	F1	R^2
8	65.24	0.20	1.84	70.20	65.14	0.920
16	65.29	0.22	2.00	70.18	65.20	0.936
32	65.34	0.21	1.97	70.22	65.26	0.943
64	65.43	0.00	1.94	70.23	65.35	0.948

NetFlow canonical (LR).

K	BA	Subset SD	Support SD	AUROC	F1	R^2
8	78.00	0.66	2.36	83.93	77.98	0.669
16	79.14	0.37	2.49	84.85	79.11	0.672
32	79.47	0.36	2.70	85.06	79.45	0.669
64	79.70	0.00	2.39	85.15	79.67	0.669

NetFlow canonical (RBF).

K	BA	Subset SD	Support SD	AUROC	F1	R^2
8	80.89	0.44	2.70	87.63	80.84	0.669
16	81.88	0.47	2.95	88.29	81.84	0.672
32	82.20	0.27	3.07	88.45	82.17	0.669
64	82.04	0.00	3.56	88.61	82.00	0.669

NetFlow evaluation (LR).

K	BA	Subset SD	Support SD	AUROC	F1	R^2
8	82.15	0.74	3.44	89.86	81.72	n/a
16	83.80	0.72	3.02	91.07	83.54	n/a
32	84.30	0.54	2.45	91.28	84.19	n/a
64	83.37	0.00	2.91	91.07	83.31	n/a

NetFlow evaluation (RBF).

K	BA	Subset SD	Support SD	AUROC	F1	R^2
8	83.82	0.45	2.86	91.37	83.45	n/a
16	84.44	0.47	2.71	91.51	84.16	n/a
32	84.75	0.32	2.50	91.36	84.52	n/a
64	84.18	0.00	2.90	91.31	83.91	n/a

Reconstruction at each K uses the corresponding target response where available; the later NetFlow panel has no target fidelity value.

Reading cached prefixes. Reading uses one fixed prefix of each bank at $K \in \{8, 16, 32, 64\}$, rather than eight permutations. Each prefix refits the same Ridge/head recipe from cached signed responses. Table 38 keeps the nonmonotonic curve and separate assay. The primary $K = 64$ functional was prespecified; no smaller-K result replaces it.

Table 38: Reading cached-probe sensitivity. BA is percent; numbers in parentheses are target counts. Reconstruction uses only the 64-record teacher assay. No additional LLM responses are acquired.

K	Student (204)	Student (64)	Teacher (64)	R^2 (64)	r (64)
8	65.83	66.80	60.35	0.739	0.864
16	63.75	61.98	61.62	0.828	0.913
32	64.07	63.17	63.31	0.859	0.929
64	73.20	71.14	74.15	0.913	0.957

F.3 Bank response-donor sample size

Each historical Bank support uses eight predetermined permutations of its 200 donors with nested prefixes $N \in \{16, 32, 50, 100, 150, 200\}$. The same subset is used across three banks. Raw preprocessing and response scaling refit on selected donors; Ridge ($\alpha = 1$), support standardization and the scalar head remain fixed. At $N = 200$, all permutations reuse the historical fit. Targets select neither subsets nor settings.

Table 39: Bank donor-size ladder. Metrics average five supports, three banks and eight fixed permutations. BA, AUROC and macro-F1 are percentages; sMSE divides response error by donor response variance. Support, bank and subset SDs first average the other two axes and use divisor $n - 1$. Intervals are conditional paired target-bootstrap percentiles; Δ compares with the historical 57.50% raw LR. The $N = 200$ headline baseline interval is reported in Table 6.

N	R^2	Pearson	Spearman	sMSE	BA	95% interval	AUROC	F1
16	-0.886	0.288	0.295	31.880	50.81	$[50.26, 51.36]$	54.26	37.87
32	-0.239	0.452	0.471	18.435	52.22	$[51.06, 53.39]$	55.87	42.82
50	-0.076	0.559	0.587	15.975	53.96	$[52.21, 55.74]$	57.52	46.53
100	0.241	0.748	0.782	10.968	57.71	$[54.54, 60.95]$	61.49	54.50
150	0.479	0.837	0.868	7.523	60.69	$[56.76, 64.60]$	63.27	59.49
200	0.695	0.879	0.898	4.189	63.61	$[59.29, 67.95]$	64.23	63.06

N	Support SD	Bank SD	Subset SD	Δ versus raw LR	95% interval
16	0.54	0.25	0.65	-6.69	$[-8.70, -4.61]$
32	0.90	0.27	1.59	-5.28	$[-6.91, -3.62]$
50	1.11	0.32	1.33	-3.54	$[-5.03, -2.02]$
100	1.51	0.54	1.63	0.21	$[-1.74, 2.14]$
150	0.61	0.65	1.33	3.19	$[0.85, 5.51]$
200	1.67	0.56	0.00	6.11	Table 6

Reconstruction and downstream BA improve over the tested grid, from $R^2 = -0.886$ and 50.81% BA at 16 donors to $R^2 = 0.695$ and 63.61% at 200. The historical system remains $N = 200$; the ladder does not select a new operating point. Zero subset SD there follows from identical donor sets, not zero support or bank uncertainty.

F.4 Outcome-blind support resampling

We draw fifty support IDs uniformly without replacement from each historical response pool: 1,000 draws from the 250-record Bank pool, 1,000 from Credit-g's 250-record pool and 100 from canonical NetFlow's 500-record pool. IDs are fixed before reading outcomes; no support is balanced, rejected or replaced. Remaining 200, 200 and 450 rows are response donors. All draws contain both classes. Targets, the three/two/three banks, prescribed functionals, students and heads remain fixed. Bank raw LR uses Appendix C's row-normalized three-fold support-CV recipe; other raw comparators retain historical settings.

Draws overlap. Their distributions describe support selection conditional on fixed historical pools, targets and probe banks, not independent datasets or target-bootstrap intervals.

Table 40: Support-resampled student-minus-raw BA gain Δ in percentage points. Metrics are averaged across banks within each draw. SD uses divisor $n-1$; percentile columns and the positive-draw fraction describe the empirical support distribution. NetFlow denotes the canonical panel.

Panel / head	Draws	Mean	Median	SD	5th	25th	75th	95th	$\Pr(\Delta > 0)$
Bank / LR	1000	4.54	5.69	9.68	-14.67	-1.58	13.65	14.84	0.636
Credit-g / LR	1000	0.50	0.54	1.22	-1.43	-0.18	1.25	2.50	0.637
NetFlow / LR	100	4.14	4.06	3.05	-0.75	2.42	6.08	8.53	0.930
NetFlow / RBF	100	4.06	4.49	2.59	-0.27	2.55	5.75	7.47	0.880

True-response mean gains under the same heads are 5.17, 0.47, 3.32 and 2.22 points for Bank, Credit-g, NetFlow LR and NetFlow RBF. Bank teacher and student gains have Pearson correlation 0.693: both exceed raw LR on 582 draws, only teacher on 87, only student on 54 and neither on 277. Mean, median and positive-draw fraction describe distinct aspects; negative tails remain in Table 40.

F.5 Bank raw-label budget curve

Twenty predetermined class-wise permutations of the existing 125/125 pool form balanced nested supports with $B \in \{16, 32, 50, 100, 150, 200, 250\}$. Raw LR uses the same fixed support-CV procedure as above. This curve describes accuracy after labels are acquired; supports are balanced using those labels. At $B = 250$ all ladders share the full pool, although predetermined fold seeds can select different C values.

Table 41: Raw-LR label-budget curve on the fixed Bank targets. All twenty nested balanced ladders are retained. SD describes ladder variation; intervals are conditional target-bootstrap percentiles. Metrics are percentages.

Labels	BA	Ladder SD	95% interval	AUROC	F1
16	52.07	7.26	$[51.41, 52.72]$	55.95	40.42
32	53.07	5.66	$[52.07, 54.05]$	61.10	40.85
50	54.37	8.33	$[52.94, 55.78]$	58.53	45.76
100	55.49	5.41	$[53.50, 57.47]$	62.00	46.89
150	57.27	4.19	$[54.31, 60.21]$	62.36	52.24
200	58.06	3.60	$[54.73, 61.37]$	61.65	53.84
250	58.68	1.57	$[54.80, 62.57]$	61.79	55.44

The $B = 50$ row averages the twenty balanced label ladders and is not the historical five-support raw-LR estimate. Within the tested Bank pool and label grid, raw LR remains below the historical 63.61% behavioral-system mean through $B = 250$ (58.68%). This comparison concerns label count only, not acquisition cost.

F.6 Reuse as a fixed downstream prior

For each historical fifty-label anchor, the response student fits its permitted donors excluding the anchor, and the logistic prior head fits only those fifty labels. Encoder, student and deployed prior head remain fixed as downstream labels increase. *Raw* is the raw-feature booster; *Prior only* is the frozen fifty-label prior; *Feature* adds its logit as an input; *PriorInit* retains that logit and fits an additive raw-feature booster residual. The final logit is fixed prior logit plus residual.

PriorInit uses five-fold out-of-fold (OOF) logits for original anchor rows; added supports and targets use the frozen full fifty-label head. Feature augmentation uses the same convention. All decisions retain a zero-logit threshold.

Five anchors each have six balanced augmentation sequences, giving 30 nested ladders at $B \in \{50, 100, 250, 500, 1000\}$ where available. Each prefix has $B/2$ records per class, including the original 25/25 anchor; allocation uses eligible class membership. Additions exclude donors, probes and targets. Bank retains chronology and excludes duration. Each NetFlow source has five anchors, 450 donors per anchor and 192 same-source targets (96 per class). Its three historical banks supply

16 measured same-source magnitude coordinates each; MLP64 changes only its output width. Bank and Credit-g keep 200 donors per anchor, their prescribed scalars and three/two banks. All fits use cached responses and CPU computation.

Booster selection and uncertainty. Fixed recipes use 80 trees, depth 3 and learning rate 0.06; LightGBM uses seven leaves and CatBoost L2 penalty 5. At $B \in \{50, 250, 1000\}$, Raw and PriorInit are tuned independently over four candidates per family (Table 49) using mean support-only stratified three-fold BA. Temporary CV prior calibrators exclude validation-anchor labels and use inner OOF logits for training-anchor rows. Added labels never recalibrate a prior. Selected boosters refit on all B labels with the unchanged deployed prior. Targets select no settings, threshold or stopping rule.

Metrics average banks within ladder, then ladders. The 95% hierarchical paired percentile intervals use 10,000 draws (seed 28101), sampling historical anchors, augmentation sequences within anchors and targets within strata. Methods, budgets and fixed/tuned comparisons share resamples. Bank uses epoch/class strata; Credit-g and each NetFlow source use class strata. Overlapping ladders are not independent datasets; $B = 50$ augmentation repeats share five anchors.

Bank. Table 42 gives the fixed curve and Table 43 the independently tuned comparison. All nine tuned gain intervals are positive, and every tuned ladder has a positive gain. Prior only remains at 63.61%; no reported residual mean exceeds it. Reuse preserves much of the prior signal within boosters rather than improving on the prior-only predictor. For each fixed learner, Raw does not reach PriorInit at 50 labels by $B = 1000$, a descriptive comparison restricted to tested budgets.

Table 42: Bank fixed-recipe learning curves. BA (%); Δ = PriorInit − Raw in points. Feature appends the fixed prior logit. Intervals include anchor, augmentation and target variation.

Learner	B	Raw	Prior only	Feature	PriorInit	Δ	95% interval
XGBoost	50	51.56	63.61	50.49	58.00	+6.44	[2.36, 10.16]
	100	50.77	63.61	47.63	58.54	+7.77	[4.33, 11.34]
	250	51.62	63.61	49.39	60.40	+8.78	[5.09, 12.77]
	500	51.96	63.61	51.10	61.67	+9.71	[5.70, 13.95]
	1000	52.56	63.61	53.14	62.42	+9.86	[5.93, 13.95]
LightGBM	50	51.61	63.61	57.03	62.03	+10.42	[3.58, 16.84]
	100	50.74	63.61	49.93	60.09	+9.35	[5.04, 13.77]
	250	51.26	63.61	48.99	60.49	+9.23	[5.25, 13.37]
	500	50.46	63.61	48.24	61.41	+10.95	[6.83, 15.11]
	1000	50.96	63.61	51.27	62.72	+11.75	[7.49, 16.12]
CatBoost	50	51.98	63.61	50.99	61.75	+9.77	[4.86, 14.81]
	100	51.76	63.61	48.72	61.44	+9.67	[5.77, 13.73]
	250	51.94	63.61	50.09	61.98	+10.04	[5.96, 14.16]
	500	51.22	63.61	49.84	62.14	+10.91	[6.64, 15.30]
	1000	52.16	63.61	51.84	63.28	+11.12	[6.89, 15.49]

Table 43: Bank support-CV selection, separately for Raw and PriorInit. Gain summaries describe the 30 ladders; SD uses divisor $n - 1$ and p_+ is the positive-ladder fraction. Prior-only and feature-augmentation results retain their fixed recipes in Table 42.

Learner	B	Raw	PriorInit	Δ	95% interval	Median	SD	p_+
XGBoost	50	53.54	61.53	+7.99	[2.57, 13.63]	9.98	3.34	1.00
	250	51.64	60.13	+8.49	[4.82, 12.36]	8.64	3.18	1.00
	1000	53.03	62.49	+9.46	[5.56, 13.39]	10.24	2.72	1.00
LightGBM	50	51.04	62.52	+11.48	[5.07, 17.36]	13.72	4.80	1.00
	250	51.28	60.03	+8.75	[5.09, 12.68]	8.98	3.04	1.00
	1000	52.47	62.43	+9.96	[5.90, 14.09]	10.29	3.50	1.00
CatBoost	50	52.19	60.03	+7.85	[3.35, 12.38]	7.03	2.59	1.00
	250	51.79	60.92	+9.13	[5.45, 12.95]	10.11	3.38	1.00
	1000	52.91	62.71	+9.80	[5.56, 14.19]	10.46	3.50	1.00

Same-source NetFlow. These are separate same-source learning problems, not source-wise slices of the mixed-source predictors reported in Appendix G. Tables 44–47 give complete fixed and tuned results. Fixed LightGBM gains 10.14 points at $B = 50$ on BoT-IoT [5.80, 14.72] and 5.07 on ToN-IoT [0.45, 10.90]. Tuning changes these to -1.70 and $+1.15$; neither tuned interval is above zero. CSE-CIC-IDS2018 shows no consistent benefit; UNSW-NB15 is often neutral or negative. No source/learner has a positive tuned mean gain at $B = 250$ or 1000. Low-label fixed-setting gains largely disappear with more labels or stronger support-selected raw learners. In the two fixed LightGBM cases, Raw first reaches PriorInit at 50 labels at $B = 100$; there is no interpolation or extrapolation.

Table 44: NF-BoT-IoT same-source learning curves. BA (%); Δ = PriorInit − Raw, with hierarchical paired 95% intervals. Feature appends the fixed prior. Dashes denote methods not retuned in the CV comparison.

Learner	B	Raw	Prior only	Feature	PriorInit	Δ	95% interval
Fixed configurations							
XGBoost	50	88.54	87.29	87.81	87.81	-0.73	$[-2.67, 1.04]$
	100	91.49	87.29	90.71	90.41	-1.08	$[-2.78, 0.41]$
	250	94.39	87.29	93.44	93.33	-1.06	$[-2.19, -0.16]$
	500	94.79	87.29	94.33	94.36	-0.43	$[-1.50, 0.53]$
	1000	95.12	87.29	94.68	94.83	-0.30	$[-1.06, 0.35]$
LightGBM	50	77.92	87.29	86.39	88.06	$+10.14$	$[5.80, 14.72]$
	100	92.01	87.29	90.95	90.86	-1.16	$[-2.48, 0.08]$
	250	94.53	87.29	93.72	93.32	-1.21	$[-2.27, -0.39]$
	500	95.14	87.29	94.35	94.40	-0.73	$[-1.55, -0.08]$
	1000	95.17	87.29	94.72	94.81	-0.36	$[-1.08, 0.21]$
CatBoost	50	87.92	87.29	88.13	88.23	$+0.31$	$[-2.60, 3.33]$
	100	91.93	87.29	90.40	89.32	-2.60	$[-4.68, -0.94]$
	250	94.70	87.29	92.70	91.38	-3.33	$[-5.38, -1.69]$
	500	95.10	87.29	93.95	92.62	-2.48	$[-4.17, -1.15]$
	1000	95.31	87.29	94.65	93.52	-1.79	$[-3.26, -0.66]$
Independent support-CV selection							
XGBoost	50	88.54	—	—	87.01	-1.53	$[-4.10, 1.35]$
	250	94.24	—	—	93.26	-0.98	$[-2.04, -0.09]$
	1000	95.19	—	—	94.97	-0.22	$[-0.82, 0.34]$
LightGBM	50	90.62	—	—	88.92	-1.70	$[-3.61, -0.10]$
	250	94.46	—	—	93.27	-1.19	$[-2.36, -0.30]$
	1000	95.24	—	—	94.93	-0.31	$[-0.92, 0.22]$
CatBoost	50	89.90	—	—	88.75	-1.15	$[-3.26, 0.73]$
	250	95.03	—	—	93.28	-1.76	$[-3.14, -0.65]$
	1000	95.33	—	—	94.87	-0.46	$[-1.05, -0.05]$

Table 45: NF-CSE-CIC-IDS2018 same-source learning curves. BA (%); Δ = PriorInit − Raw, with hierarchical paired 95% intervals. Feature appends the fixed prior. Dashes denote methods not retuned in the CV comparison.

Learner	B	Raw	Prior only	Feature	PriorInit	Δ	95% interval
Fixed configurations							
XGBoost	50	84.69	81.46	85.00	84.37	-0.31	$[-3.65, 2.67]$
	100	88.58	81.46	88.41	86.50	-2.08	$[-4.02, -0.31]$
	250	91.35	81.46	91.20	89.87	-1.49	$[-2.97, -0.21]$
	500	91.94	81.46	91.86	90.93	-1.02	$[-2.25, 0.05]$
	1000	92.52	81.46	92.58	91.61	-0.90	$[-2.09, -0.03]$
LightGBM	50	85.10	81.46	84.72	82.19	-2.92	$[-6.28, 0.49]$
	100	89.03	81.46	88.63	86.54	-2.49	$[-4.12, -1.02]$
	250	91.30	81.46	90.76	89.75	-1.55	$[-2.78, -0.44]$
	500	91.55	81.46	91.39	90.87	-0.68	$[-1.78, 0.26]$
	1000	92.31	81.46	92.27	91.68	-0.63	$[-1.75, 0.16]$
CatBoost	50	86.67	81.46	86.35	83.40	-3.26	$[-6.15, -0.49]$
	100	89.58	81.46	89.59	87.06	-2.52	$[-4.13, -0.97]$
	250	90.78	81.46	90.55	89.10	-1.68	$[-3.26, -0.26]$
	500	91.13	81.46	90.92	89.88	-1.24	$[-2.93, 0.19]$
	1000	91.89	81.46	91.44	90.49	-1.41	$[-3.12, 0.12]$
Independent support-CV selection							
XGBoost	50	83.96	—	—	84.31	$+0.35$	$[-2.54, 3.54]$
	250	91.30	—	—	89.84	-1.46	$[-3.15, 0.02]$
	1000	92.55	—	—	91.73	-0.82	$[-1.79, -0.10]$
LightGBM	50	85.52	—	—	83.89	-1.63	$[-4.51, 1.32]$
	250	91.06	—	—	89.77	-1.29	$[-2.51, -0.20]$
	1000	92.45	—	—	91.75	-0.70	$[-1.65, 0.03]$
CatBoost	50	86.56	—	—	83.09	-3.47	$[-7.64, 0.17]$
	250	91.02	—	—	89.81	-1.21	$[-2.23, -0.29]$
	1000	91.74	—	—	91.05	-0.69	$[-1.56, -0.06]$

Table 46: NF-ToN-IoT same-source learning curves. BA (%); Δ = PriorInit − Raw, with hierarchical paired 95% intervals. Feature appends the fixed prior. Dashes denote methods not retuned in the CV comparison.

Learner	B	Raw	Prior only	Feature	PriorInit	Δ	95% interval
Fixed configurations							
XGBoost	50	90.62	90.59	90.24	90.97	$+0.35$	$[-2.95, 3.89]$
	100	94.46	90.59	93.65	94.74	$+0.28$	$[-1.50, 1.94]$
	250	97.27	90.59	96.97	97.18	-0.10	$[-0.83, 0.58]$
	500	98.49	90.59	98.23	98.35	-0.14	$[-0.99, 0.75]$
	1000	99.03	90.59	98.95	98.91	-0.12	$[-0.97, 0.73]$
LightGBM	50	86.04	90.59	90.52	91.11	$+5.07$	$[0.45, 10.90]$
	100	95.21	90.59	94.72	95.45	$+0.24$	$[-1.38, 1.86]$
	250	97.27	90.59	97.24	97.62	$+0.34$	$[-0.41, 1.18]$
	500	98.37	90.59	98.25	98.67	$+0.31$	$[-0.45, 1.17]$
	1000	98.96	90.59	98.78	99.13	$+0.17$	$[-0.54, 1.09]$
CatBoost	50	91.25	90.59	91.42	90.83	-0.42	$[-3.78, 3.19]$
	100	93.39	90.59	93.36	92.55	-0.84	$[-3.00, 1.52]$
	250	95.26	90.59	95.24	94.18	-1.08	$[-2.49, 0.14]$
	500	96.75	90.59	96.19	95.49	-1.27	$[-2.81, -0.01]$
	1000	97.47	90.59	97.20	96.59	-0.87	$[-2.18, 0.20]$
Independent support-CV selection							
XGBoost	50	89.69	—	—	90.52	$+0.83$	$[-3.12, 5.52]$
	250	97.20	—	—	97.19	-0.02	$[-1.00, 0.91]$
	1000	99.67	—	—	99.18	-0.49	$[-1.13, -0.05]$
LightGBM	50	90.73	—	—	91.88	$+1.15$	$[-1.53, 4.13]$
	250	97.80	—	—	97.45	-0.34	$[-1.08, 0.30]$
	1000	99.53	—	—	99.17	-0.36	$[-0.90, -0.03]$
CatBoost	50	91.88	—	—	91.91	$+0.03$	$[-4.34, 4.13]$
	250	97.71	—	—	97.36	-0.35	$[-1.06, 0.27]$
	1000	99.32	—	—	99.09	-0.23	$[-0.89, 0.21]$

Table 47: NF-UNSW-NB15 same-source learning curves. BA (%); Δ = PriorInit − Raw, with hierarchical paired 95% intervals. Feature appends the fixed prior. Dashes denote methods not retuned in the CV comparison.

Learner	B	Raw	Prior only	Feature	PriorInit	Δ	95% interval
Fixed configurations							
XGBoost	50	85.62	77.60	81.87	81.81	-3.82	$[-7.43, -0.17]$
	100	91.81	77.60	90.60	88.29	-3.52	$[-6.06, -1.06]$
	250	94.65	77.60	94.41	92.83	-1.82	$[-3.40, -0.29]$
	500	96.75	77.60	96.65	95.56	-1.20	$[-2.33, -0.13]$
	1000	97.59	77.60	97.54	96.89	-0.69	$[-1.68, 0.32]$
LightGBM	50	79.27	77.60	79.55	80.03	$+0.76$	$[-4.79, 6.46]$
	100	89.86	77.60	89.04	88.30	-1.56	$[-3.31, 0.19]$
	250	94.91	77.60	94.81	93.47	-1.44	$[-2.88, -0.10]$
	500	97.00	77.60	96.89	96.22	-0.78	$[-1.90, 0.35]$
	1000	97.90	77.60	97.71	97.44	-0.46	$[-1.49, 0.63]$
CatBoost	50	85.31	77.60	82.64	81.81	-3.51	$[-7.54, 0.87]$
	100	89.90	77.60	88.81	84.57	-5.32	$[-8.28, -2.45]$
	250	93.44	77.60	92.32	88.13	-5.31	$[-7.52, -3.10]$
	500	96.15	77.60	95.27	90.44	-5.71	$[-8.33, -3.26]$
	1000	97.05	77.60	96.61	92.23	-4.81	$[-7.18, -2.63]$
Independent support-CV selection							
XGBoost	50	83.02	—	—	80.38	-2.64	$[-8.30, 2.78]$
	250	94.67	—	—	93.29	-1.38	$[-2.86, 0.00]$
	1000	98.12	—	—	98.08	-0.04	$[-0.67, 0.76]$
LightGBM	50	83.85	—	—	82.85	-1.01	$[-6.08, 4.20]$
	250	95.33	—	—	93.94	-1.39	$[-3.02, 0.07]$
	1000	98.26	—	—	98.16	-0.10	$[-0.78, 0.76]$
CatBoost	50	86.15	—	—	84.48	-1.67	$[-5.10, 2.29]$
	250	96.06	—	—	94.99	-1.07	$[-2.42, 0.31]$
	1000	98.65	—	—	98.31	-0.34	$[-0.85, 0.25]$

Credit-g. Only $B = 50$ is available. After fixed role exclusions, the extra pool has 342 negatives and no positives, so larger balanced supports are impossible. Prior only is 60.39% BA; tuned gains are -1.04, -1.86 and $+1.93$ for XGBoost, LightGBM and CatBoost, with all intervals including zero.

Table 48: Credit-g at $B = 50$. BA (%), paired gain and hierarchical 95% interval. All larger requested budgets are unavailable under the fixed roles. Dashes denote methods not retuned.

Learner	Raw	Prior only	Feature	PriorInit	Δ	95% interval
Fixed configurations						
XGBoost	62.50	60.39	64.36	63.25	$+0.75$	$[-1.93, 3.39]$
LightGBM	57.79	60.39	59.86	59.96	$+2.18$	$[-0.82, 5.36]$
CatBoost	64.50	60.39	65.43	65.57	$+1.07$	$[-1.54, 3.64]$
Independent support-CV selection						
XGBoost	61.79	—	—	60.75	-1.04	$[-5.00, 2.21]$
LightGBM	63.07	—	—	61.21	-1.86	$[-7.32, 2.47]$
CatBoost	61.43	—	—	63.36	$+1.93$	$[-0.93, 5.25]$

Table 49: Four support-CV candidates per booster; candidate 0 is the fixed recipe. T is the number of trees, d depth and η learning rate. Min. denotes minimum child weight for XGBoost and minimum leaf count for LightGBM; λ is the library L2 penalty. Other settings remain at their recorded fixed values.

Learner	Candidate	T	d	η	Leaves	Min.	λ
XGBoost	0	80	3	0.06	—	1	1
	1	200	2	0.05	—	3	5
	2	200	3	0.05	—	1	1
	3	400	5	0.03	—	5	5
LightGBM	0	80	3	0.06	7	20	0
	1	200	2	0.05	3	10	5
	2	200	3	0.05	7	5	1
	3	400	5	0.03	15	10	5
CatBoost	0	80	3	0.06	—	—	5
	1	200	2	0.05	—	—	10
	2	200	3	0.05	—	—	3
	3	400	5	0.03	—	—	10

Candidate ties use fixed priority orders $(3, 1, 0, 2)$, $(1, 3, 2, 0)$ and $(1, 3, 0, 2)$ for XGBoost, LightGBM and CatBoost. Class balancing yields unit relative weights on full balanced supports; folds use their own counts. XGBoost retains base margins at prediction. LightGBM and CatBoost add the fixed target margin to their raw correction.

G NetFlow source-wise decomposition

The analyses below partition the fixed canonical and evaluation target panels by source. Models retain their original mixed-source fits and are not refit by source; the separate same-source learning experiments are reported in Appendix F.6.

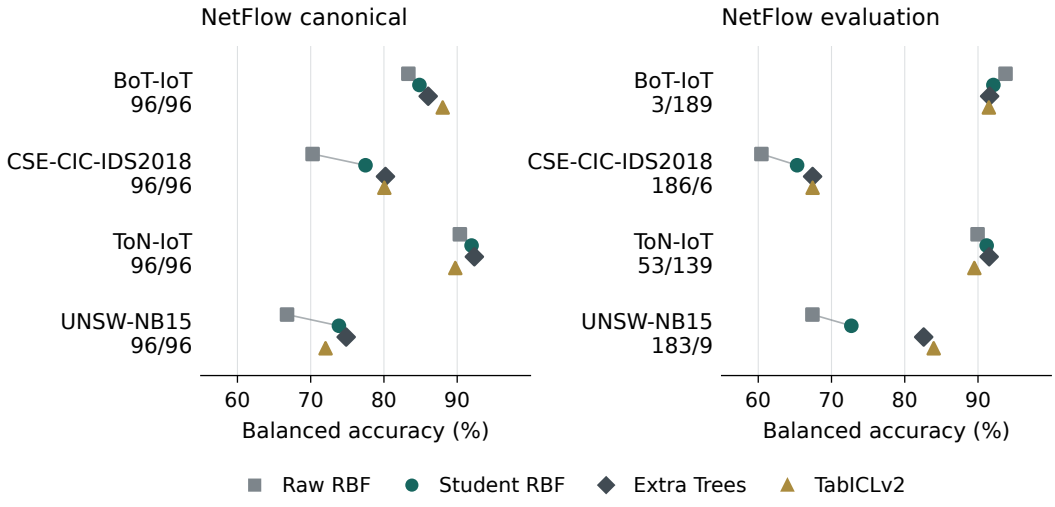


Figure 8: Source partitions of fixed canonical and evaluation NetFlow target panels. RBF systems retain their mixed-source fits, with no refitting by source. Labels give benign/attack counts; small minority cells make evaluation source scores descriptive.

Table 50: NetFlow source BA (%), averaged over fixed mixed-source supports and banks without source-specific refitting. +Student adds the historical MAG64 student within each readout; ET denotes Extra Trees. Names omit NF-. Counts are benign/attack. Small minority cells make evaluation source scores descriptive.

		Logistic		RBF			
Source	n_0/n_1	Raw	+Student	Raw	+Student	ET	TabICLv2
Balanced canonical panel							
BoT-IoT	96/96	77.81	82.60	83.33	84.84	86.04	88.02
CSE-CIC-IDS2018	96/96	74.27	77.29	70.26	77.48	80.21	80.05
ToN-IoT	96/96	86.30	89.06	90.36	91.96	92.34	89.74
UNSW-NB15	96/96	64.22	69.83	66.77	73.85	74.84	72.03
Evaluation panel							
BoT-IoT	3/189	88.73	83.69	93.76	92.10	91.59	91.48
CSE-CIC-IDS2018	186/6	56.13	66.22	60.43	65.30	67.42	67.42
ToN-IoT	53/139	92.67	88.73	89.94	91.18	91.52	89.50
UNSW-NB15	183/9	63.70	70.39	67.40	72.71	82.59	83.95

Canonical student RBF exceeds TabICLv2 on ToN-IoT (91.96% versus 89.74%) and UNSW-NB15 (73.85% versus 72.03%), and improves on raw RBF in every canonical source. In the separate evaluation it exceeds TabICLv2 on BoT-IoT (92.10% versus 91.48%) and ToN-IoT (91.18% versus 89.50%); on BoT-IoT it also exceeds Extra Trees (92.10% versus 91.59%). Several evaluation minority cells are small, so source scores are descriptive.

H AMORTIZATION FIDELITY AND RESOURCE ACCOUNTING

Amortization fidelity. Table 51 compares true-response and student readouts on matching panels and heads. Table 52 gives historical reconstruction metrics; reconstruction requires no outcome labels. Bank target R^2 uses target variance, while standardized MSE uses donor response scales. No target teacher response exists for the later NetFlow evaluation.

Table 51: Teacher-readout preservation and reconstruction fidelity on the same panel and with the same outcome head. Reading is the 64-target assay only; NetFlow is canonical. The later NetFlow evaluation has no target teacher score. BA is percent; R^2 measures target response reconstruction, averaging coordinates for NetFlow.

Dataset	Readout	Teacher BA	Student BA	R^2
Bank	Scalar LR	64.53	63.61	0.695
Reading (64-target assay)	Scalar LR	74.15	71.14	0.913
Credit-g	Raw + scalar LR	65.50	65.43	0.948
NetFlow	Raw + MAG64 RBF	81.39	82.04	0.669

Table 52: Historical target reconstruction, averaged over permitted support/bank student fits. NetFlow R^2 averages coordinates; its Pearson coefficient flattens the response matrix and cosine averages rows. Scalars use ordinary R^2, Pearson and Spearman. Standardized MSE uses donor response scales.

Dataset	Student	R^2	Pearson	Spearman	Cosine	MSE	Std. MSE
Bank	Ridge	0.695	0.879	0.898	n/a	2.750	4.189
Credit-g	Ridge	0.948	0.974	0.973	n/a	1.060	0.057
NetFlow	MLP64	0.669	0.881	n/a	0.964	0.376	0.338

Unstandardized MSE in Table 52 uses historical response units. For Bank and NetFlow, $P = 2D$ and MAG64$= 2|D|$, so the $D/|D|$ MSE values in Appendix E are one quarter of these historical-unit

values. R^2, correlations and standardized MSE are unchanged by positive rescaling. Reading assay fidelity is reported in Appendix D.1 and Table 51.

Recorded student footprint. Table 53 includes preprocessing and excludes the final task head; it is not end-to-end deployment latency. Sizes are per bank-specific student.

Table 53: Recorded student feature-generator footprint and CPU inference timings. Preprocessing is included and the final task head is excluded. Serialized size is per bank-specific student, not a language-model checkpoint.

Student	Parameters	Bytes	Single row (ms)	Batched ms/row
Bank Ridge	54–55	6,682–6,722	8.23	0.031
Credit-g Ridge	70–72	4,581–4,621	7.98	0.041
Canonical NetFlow MLP64	4,864	161,831–163,413	5.41	0.009
Later NetFlow MLP64	4,864	140,549	n/a	n/a

The NetFlow MLP has 4,864 parameters; three bank-specific students are retained separately. Recorded CPU times depend on host, software and batching and are not a matched cross-hardware LLM speed comparison. No Reading deployment latency is inferred.

Acquisition accounting. For K probes, paired acquisition requires $4K$ logical contexts per candidate and probe bank: K probes $\times$ two hypothetical labels $\times$ two A/B answer mappings.

Reading acquisition and deployment. Acquisition covers 250 pool candidates and 64 teacher-assay candidates with two 64-probe banks: 160,768 contexts, including 128,000 pool and 32,768 assay contexts. Candidate-prefix keys/values are reused across 128 probe queries. Unpadded input tokens total 62,291,440 without reuse and 27,950,132 with reuse; mean full context length is 387 tokens and maximum 403. Query batch size is 64. BF16, SDPA and non-thinking next-token inference are used on an NVIDIA A100-SXM4-80GB.

The 40 acquisition blocks total 4,165.42 s (69.42 min), including recorded block overhead. The 69.42 min measurement covers response acquisition; the 79.1 min session total additionally includes setup, pilot and downstream work.

True responses on 64 targets are an offline assay only and enter neither student loss nor task-head fitting. All 204 student target predictions use raw attributes and the ten fitted Ridge/head objects, without LLM inference. No teacher BA is inferred for the remaining 140 targets, and end-to-end Reading deployment latency was not measured.

Bank offline acquisition. The complete cache contains 634 candidates measured against three 64-probe banks, with 256 logical contexts per candidate/bank. It records 486,912 contexts, 161,924,424 unpadded input tokens and 3,804 forward calls at effective batch size 128 on an NVIDIA A100-SXM4-80GB. Tokens here are measured inputs, not generated outputs or FLOPs. Accelerator capacity is not a peak-memory measurement.

Table 54: Bank paired-context acquisition scopes. The rows overlap and must not be summed. The 250-row donor union is shared across five support exclusions. Evaluation responses belong to the complete historical cache, not donor-only student-training cost.

Scope	Candidates	Banks	Logical contexts
One support, one bank	200	1	51,200
One support, three banks	200	3	153,600
Donor union	250	3	192,000
Evaluation targets	384	3	294,912
Complete archive	634	3	486,912

Complete-archive measurement	Recorded value
Probes per bank	64
Contexts per candidate per bank	256
Recorded input tokens	161,924,424
Batched forward calls	3,804
Effective batch size	128
Response-loop time (s)	21302.08
Response-loop time (h)	5.92
Hardware	NVIDIA A100-SXM4-80GB

Logical-context counts describe acquisition scope, not additive marginal wall-clock costs. The response loop totals 21,302 seconds (5.92 hours) across three sessions, excluding model loading and separate direct/static passes. No donor-only wall time or single-record teacher latency is inferred.

These measurements support offline acquisition accounting, not a matched wall-clock break-even estimate. Deployment on a new record uses only preprocessing, the student and the fitted task head, with no LLM inference.

I Experimental provenance and replay coverage

This appendix records analysis chronology, target-access status, prediction locks, artifact identities, and the scope and limitations of deterministic replay.

I.1 Analysis chronology and target access

Bank, Credit-g and canonical NetFlow functionals were developed historically with access to panel outcomes. They support diagnostic, mechanistic and retrospective analyses, not prospective functional selection. The separate NetFlow evaluation was allocated from unused records under a fixed protocol: target identities and LR/RBF procedures were fixed before scoring, and predictions preceded target-outcome access.

Its fixed-protocol intervals remain LR $+3.48$ [1.64, 5.35] and RBF $+0.73$ [-0.74, 2.21]. NetFlow TabICLv2 was added after those outcomes were available, using the same supports and targets without selecting settings from target performance.

I.2 Prospective Reading transfer and prediction locks

Reading applies the Bank signed first moment unchanged with Qwen3-14B. The functional and $K = 64$ were specified before target scoring. Outcome-blind roles, prompts, acquisition design and seeds were fixed before acquisition; target predictions were locked before outcomes were loaded. This is prospective transfer of a fixed functional, not a general functional-selection procedure or a claim that dataset/model choice was randomized.

The original five supports, two banks and target panel remain unchanged. The later comparator study replays the frozen student and LR/Extra Trees predictions, fixes support-only grids and foundation-

model settings, and locks added predictions before scoring. It is a subsequent comparison, separate from the original prospective acquisition.

Original deployment lock (SHA256):
`ed607a9022f695f94fbbb191252f02e15ec3acf4fdf44dea527a6c80d77b7b29`
Expanded comparator lock (SHA256):
`ae9fddc479dfa34bd8c7d0114b9be0cb06d8bf3b130d6440e035f1a0f5b836a9`

The artifact record includes the primitive-response cache, deployed objects, row ledger, workbook, executed acquisition/comparison notebooks, selected hyperparameters and unrounded metrics. It preserves zero-based and Excel row numbers, support/bank membership, probe order, teacher-assay membership, checksums and exact model/tokenizer revisions. Original target scores from all ten $K = 64$ deployed objects are reproduced by preprocessing, Ridge and head alone; LR/Extra Trees scores and decisions also agree with their original predictions.

The Qwen model and tokenizer revision is:
`40c069824f4251a91eefaf281ebe4c544efd3e18`. The official template has an empty non-thinking block with `enable_thinking=False`; A/B token IDs are 32/33. Every serialized row and tokenized context is retained alongside mapping-specific semantic margins and raw answer logits. Allocation uses seed 28101 and assay selection 28102; the duplicate predictor pair comprises zero-based rows 433/434. Full paths, package stacks and seed inventories are centralized in the accompanying implementation ledger and protocol records.

I.3 Post-hoc controls and fixed-analysis status

Bank baselines. Baseline accounting loads saved student predictions without refitting. All requested baseline predictions precede target-truth loading; ten historical replays match native predictions and recorded CV choices. Historical, common-protocol and foundation-model families stay distinct, without a target-score maximum over incompatible recipes. An older Credit-g tree aggregate could not be matched sufficiently to its role/configuration records; it remains flagged as unmatched and is not used for the main result.

Bank diagnostics. Raw-feature controls, support-null interventions, donor ladders, outcome-blind support draws, the label-budget curve and cached Llama comparison preserve the prescribed signed response and student/head settings unless a diagnostic explicitly changes the auxiliary target. IDs, seeds, subsets and comparison rules precede scoring. The label-budget curve is retrospective accounting within the existing labeled pool. Cached-response diagnostics acquire no new LLM responses. The Llama cache is finite, its stored signed differences equal $m_1 - m_0$, and candidate/probe identities and raw-input hashes match the historical panel; its revision is `0e9e39f249a16976918f6564b8830bc894c89659`.

Structural controls. The Q/D comparisons, raw augmentation, signed-versus-magnitude profiles, bank summaries and K sensitivity preserve historical panels and model recipes. Comparison sets precede scoring; these are controls on existing systems, not prospective task selection. Outcome-free cross-bank diagnostics compare scalar Pearson/Spearman correlations on common donor IDs and centered Gram matrices, also after row normalization. Coordinates across banks are not treated as matched probes. Sign diagnostics retain positive/negative/zero fractions and coordinate stability; detailed results remain machine-readable.

Automatic readout selection. This is a secondary cached-response analysis on eight task panels already evaluated historically. Within the selector run, support-only decisions and all candidate target scores are written and hashed before target outcomes are loaded. The computational lock prevents target-driven selection within that run; it does not turn the historical panels into new prospective experiments. The Raw-only control subsequently reuses the locked support-CV tables and candidate predictions, with no refits. No additional LLM responses are acquired.

Prior reuse. Learning curves are post-hoc analyses on existing task panels. The student and fifty-label prior freeze before adding downstream labels; boosters use support-only CV and target predictions are saved before scoring. NetFlow uses separate same-source fits. The original $B = 50$ replay precedes learning curves; each library's prior-plus-correction arithmetic is checked.

Mechanism diagnostics. Appendix E.5 operates on the fitted historical Bank system without replacing its functional, student, head or threshold. Its direction statements condition on the prescribed functional. The affine-score audit reproduces aggregates without refitting, using the fixed numerical rank and least-squares conventions. No repair or recalibration is evaluated.

Archived later NetFlow Qwen. This aggregate-only teacher-response system uses a different bank and representation from the Llama panels. Its later comparison aggregates two supports and is not used to estimate a backbone effect; scores and configurations remain in the supplement.

I.4 Training-resource and allocation provenance

Fifty labels means outcomes used by each task head. Response-bearing donors are additional training data and are not interchangeable with outcome labels. Historical donor pools and later NetFlow supports retain their recorded allocation, including label-based balancing where stated. Bank's normalized visible-record identities enforce role exclusions.

Before later NetFlow allocation, all 2,768 historical candidate identities were excluded. Only 1,268 had complete finite MAG64 in all three banks; 1,500 incomplete rows were neither imputed nor zero-filled. The complete intersection was chosen before allocating the later panel. Its historical pool includes earlier targets but not their labels in response regression.

The allocation record reports 8,676,290 non-target labels screened to form five balanced supports, supported by the saved protocol, allocation code and pool arithmetic. No new full-dataset scan was used to reverify this count. Older Bank cache variants using Ridge $\alpha = 0.1$ are historical artifacts; all reported main Bank results use Ridge $\alpha = 1$.

I.5 Replay coverage and known limitations

For the original-task compact caches, equivalent m_0, m_1 margins are reconstructed algebraically from stored Q, P; the primitive next-token logits and complete chat transcripts are not part of the compact replay. Reading retains primitive A/B logits, mapping margins and serialized/tokenized contexts. The Bank mechanism cache also retains semantic-aligned mapping margins; its loader verifies their aggregation into m_0, m_1, P. Mean stored Bank P coordinates reproduce every checked historical scalar teacher value.

CPU replay uses stored responses without reloading a teacher, regenerating next-token margins or extracting responses. It preserves coordinate order and preprocessing/scaling conventions. The $P = 2D$ standardized-pipeline equivalence and full-probe outputs are checked against stored/native predictions where supported. Canonical NetFlow's support-indexed split identities are replayed.

Later NetFlow refitting requires exact raw replay inputs or the original CSV with its allocation manifest and donor binding. Historical raw values preserve JSON precision; the later panel uses canonical CSV precision. Native student and structural-control predictions are supplied.

Known replay limits. Teacher responses are unavailable for later NetFlow supports and targets. Credit-g and NetFlow TabICLv2 comparisons in the compact response-control package, and the archived later NetFlow Qwen case, are aggregate-only and cannot yield target-level replay from that package. The missing original historical NetFlow bootstrap receipt limits interval verification to printed/preserved precision. Cached-response controls have separate predictions and resampling specifications: Appendix D's later RBF interval $[-0.74, 2.21]$ is not replaced by the cached-control bound $[-0.76, 2.21]$. The artifact README maps inputs, predictions and scripts and records probe-identity and calibration-object coverage limits.

J Comparison with related supervision mechanisms

Table 55 compares selected related methods by the training-time signal, the object retained or reused afterward, where reference outcomes enter, and whether an LLM remains in the prediction path for new examples. The rows are intended to distinguish supervision structure rather than rank methods or equate their experimental settings.

Table 55: Selected related supervision mechanisms. The outcome column identifies where reference outcomes enter the listed transfer or curation stage, not the total labeling budget. The LLM column refers to the final prediction path. CE means cross-entropy; Hyper-ICL also reports alignment-only training. Our separate task head fits 50 outcome labels.

Method	Training-time signal	Retained / reused object	Where outcomes enter	LLM on new rows?
FeatLLM (Han et al., 2024)	Generated executable class rules	Rule features + fitted task head	Rule prompt / head loss	No
LUPI / generalized distillation (Vapnik & Izmailov, 2015; Lopez-Paz et al., 2016)	Privileged/teacher responses or features	Ordinary-input transferred predictor/features	Privileged/teacher construction and downstream supervision; formulation-dependent	N/A
Chang & Jia (2023)	Correct-class development margins	Demonstration-identity / position surrogate; curated prompts	Outcome-defined margin target	Yes
NUGGETS (Li et al., 2024b)	Improved anchor-answer likelihood score	Curated instruction–answer data	Anchor scoring / fine-tuning	Yes
Hyper-ICL (Alipour Talemi et al., 2026)	Demonstration-conditioned query hidden states	Adapted multimodal LM/query-token mechanism	CE in main supervised objective	Yes
Ours	Paired hypothetical-label responses on unlabeled probes	Raw-input response student + separate task head	Not in response-student loss; 50-label task head	No