SHAPLEY DATA AUDITING CERTIFICATES: BOUNDING DOWNSTREAM FAIRNESS WITHOUT MODEL ACCESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Alignment auditing is organizationally separated from model development: those who assemble and curate data, those who train models, and third parties who assess the system for fairness and privacy compliance, rarely hold the same information. We take this role separation as our design constraint and ask what an auditor can certify using only the information its role grants. We formalize a staged pipeline in which each block consumes the output of the previous one, and attach to a data block a *stage-local certificate*: a Shapley attribution of downstream accuracy and group-fairness violation to that block's input features, computed from dependence measures over data variables alone, with no access to the model or its predictions. A tight one-sided guarantee makes the certificate meaningful under restricted access: the mutual-information and conditional-mutual-information certificates upper-bound the sensitive-attribute leakage—hence the statistical-parity and equalized-odds (EO) gaps—of *every* predictor built downstream, with equality exactly when the downstream model realizes all of that leakage. By benchmarking 26 measures on seven datasets, with robustness across seven downstream predictors, we quantify the *pre-hoc attribution gap*—the fidelity a data auditor forfeits by not holding model access—and trace it to whether a measure's scale moves with coalition size. A final study prices the audit itself: permutation sampling brings a 20-feature certificate whose exact evaluation costs 71–150 hours down to sub-1% error in 30–70 minutes, while declared causal constraints move fidelity non-monotonically, so the constraint set must be reported. When a biased upstream annotator emits a stage's supervision, a two-stage study on 26 U.S. states shows that label-dependent certificates register the upstream block, the statistical-parity certificate is blind to it, and an EO certificate on inherited labels does not decisively reveal whether downstream harm rises or falls. Our results yield construction guidelines indexed by access regime—rather than by objective, bounding the price of the information asymmetry that alignment governance is built on.

1 INTRODUCTION

Proposals for governing advanced AI systems converge on a division of labor. Internal audit functions review systems before release (Raji et al., 2020); third-party auditors assess artifacts they did not produce (Raji et al., 2022; Brundage et al., 2026); dataset and model documentation is written by different teams at different stages (Gebru et al., 2021); and risk-management and conformity regimes provide for independent assessment (NIST, 2023). But this creates an information asymmetry that most technical auditing methods quietly assume away: they require query access to a trained model, and often to its internals (Casper et al., 2024; Marks et al., 2025).

The asymmetry is sharper at the data stage. The party best positioned to detect a fairness or privacy problem at its source—the team assembling, curating, and encoding the data—is often far from the deployed model, separated by organizational boundaries, minimization and siloing requirements, and legal limits on who may hold sensitive attributes. Yet upstream failures are the expensive ones: data cascades compound downstream and are systematically undervalued (Sambasivan et al., 2021); per-stage fairness guarantees do not compose (Dwork & Ilvento, 2019); and model-side procedures such as RLHF cannot repair a defective supervision signal (Casper et al., 2023; Gao et al., 2023). As supervision is increasingly emitted by earlier learned blocks (preference models, constitutional self-labeling, synthetic annotation), the biases a system inherits are pipeline, not model, artifacts

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

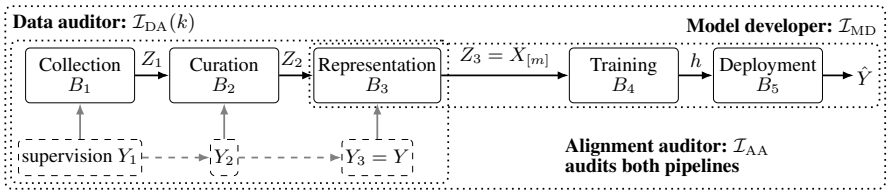

Figure 1: A staged pipeline and audit roles. Block B_k maps Z_{k-1} to Z_k; the data (B_1–B_3) and model (B_4–B_5) pipelines meet at the representation boundary $Z_3 = X_{[m]}$. Supervision Y_k for auditing stage k may be emitted by an earlier block, carrying its bias. Dotted regions are the *information-access sets* of Definition 1: the data auditor sees A, stage inputs and appropriate supervision, but no model or predictions; the model developer sees $X_{[m]}$ and h; the alignment auditor spans both.

(Bai et al., 2022; Shumailov et al., 2024). This paper asks *what each role in that division can certify, and at what cost.* We treat role and pipeline separation as primitives rather than deployment details.

Pipelines and roles, not models. We model a system as a chain of blocks, split into a *data pipeline* and a *model pipeline* (Figure 1). Auditing is stage-indexed: its object is a block together with the information available at its boundary, and since the supervision used to audit a block may be the output of an earlier block, bias amplification is localized to the stage that introduced it. We separate the *data auditor*, *model developer*, and *alignment auditor* who reviews both, each with an explicit information-access set; a method is admissible for a role only if it is computable from that set. This rules out post-hoc explanation for the data auditor and motivates *stage-local certificates*: Shapley attributions of downstream accuracy and group-fairness violation to a stage's input features, computed from dependence measures over data variables alone and attached to the representation boundary, the last point at which a property can be certified without reference to a particular predictor.

Tight guarantees, then measurement. A certificate without model access is useful only if it says something about models that do not yet exist. We give a tight one-sided guarantee: MI and CMI certificates upper-bound sensitive leakage of *any* predictor restricted to audited features. We then measure the residual across 26 measures, 7 datasets, and up to 7 downstream predictors, quantifying the *pre-hoc attribution gap*: the fidelity a data auditor gives up relative to an alignment auditor.

Contributions. (i) A staged-pipeline, role-indexed *formalization* of data auditing in which the admissibility of an auditing method is fixed by an information-access set, together with an explicit statement of which of its cases our experiments instantiate (Section 3); (ii) stage-local certificates with a tight one-sided guarantee valid for every downstream predictor, and an explicit statement of what such certificates cannot establish (Section 4); (iii) a benchmark organized by the question each role asks, with the main fidelity benchmark run against seven heterogeneous predictors rather than a single architecture (Section 5); (iv) a measurement of the price of role separation, the cost of issuing a certificate at governance cadence, and five resulting guidelines for measure construction (Sections 5.3–6); (v) a two-stage study on 26 states of what a stage can certify when its supervision is inherited (Section 5.5). All proofs, our full benchmark tables, and the estimator-sensitivity, known-generative-structure, and cadence studies are in the appendices, cross-referenced throughout.

2 Related Work

Auditing under restricted access. Internal and third-party auditing frameworks establish the organizational separation we formalize, now extended to frontier developers (Raji et al., 2020; 2022; Brundage et al., 2026). Casper et al. (2024) argue that black-box access is insufficient for rigorous audits and catalog the access levels an auditor may hold; Chan et al. (2024) formalize visibility into deployed systems. Our access sets are in this spirit but push *below* black-box access: the data auditor holds no model at all. Model-side auditing, such as recovering hidden objectives (Marks et al., 2025), is complementary and inadmissible for the data auditor. Two observations motivate not routing all assurance through the model: post-hoc explanations can be gamed by the audited party (Slack et al., 2020), and models can behave differently when they infer they are being evaluated (Greenblatt et al., 2024). Evidence obtained upstream escapes both.

Data-centric alignment. Data valuation via Shapley values and influence functions attributes model performance to training *examples* (Ghorbani & Zou, 2019; Koh & Liang, 2017). We instead attribute

accuracy and fairness violation to *features* of a stage, without fitting the downstream model, motivated by the compounding of data-stage defects (Sambasivan et al., 2021; Shumailov et al., 2024) and by the non-composition of stagewise fairness guarantees (Dwork & Ilvento, 2019).

Fairness criteria as measured constructs. We audit statistical parity (Calders et al., 2009) and equalized odds (Hardt et al., 2016), mutually incompatible in general (Barocas et al., 2023). Treating them as measurements requires attention to construct validity (Jacobs & Wallach, 2021): a certificate is a measurement instrument, so we report where its readings disagree with realized downstream behavior. Proxy features carrying sensitive information without being sensitive attributes (Datta et al., 2017; Plecko & Bareinboim, 2024) are what feature-level attribution is meant to surface, unlike representation-level interventions (Zemel et al., 2013) that act after the fact.

Shapley attribution. We use the Shapley value (Shapley, 1953) as the aggregation rule, following the removal-based view of explanation (Covert et al., 2020; 2021); known caveats around feature correlation and the on- versus off-manifold restriction (Kumar et al., 2020; Chen et al., 2020) lead us to adopt the conditional restriction. Pre-hoc measures over data variables were introduced by Khodadadian et al. (2021) and Pelegrina et al. (2024); the restricted-predictor semantics we rely on are analyzed by Nafea & Aldarmini (2026). Exact aggregation is exponential in the number of features, so we use the permutation-sampling estimator of Castro et al. (2009) and the causally constrained variant of Frye et al. (2020). A companion study characterizes how its cost and fidelity scale with dimension and quantization (Anonymous Authors, 2026); we take that profile as given.

3 Staged Pipelines and Audit Roles

Notation. For $S \subseteq [m] \triangleq \{1, \dots, m\}$, $X_S \triangleq \{X_i : i \in S\}$ and $S^c \triangleq [m] \setminus S$; $A^{(n)}$ is n samples of A, p_A, $p_{A|B}$ distributions, and $I(\cdot;\cdot)$, $I(\cdot;\cdot \mid \cdot)$ (conditional) mutual information (MI, CMI).

Staged pipelines. A pipeline is a chain of blocks $B_1, \dots, B_K$ in which B_k consumes Z_{k-1} and emits Z_k. Blocks $B_1, \dots, B_{K_d}$ form the *data pipeline* and the rest the *model pipeline*; the two meet at the representation boundary $Z_{K_d} = X_{[m]}$. Alongside $X_{[m]}$ we have a sensitive attribute $A \in \{0, 1\}$ and a per-block supervision signal Y_k.

The role of Y_k is what makes the framework stage-local rather than merely pre-hoc: Y_k is any well-defined supervision signal for stage k, and may be produced by an earlier statistical or learned block—an annotator model, a preference model, a synthetic labeler—carrying that block's inherited biases (Bai et al., 2022; Shumailov et al., 2024). When supervision emitted by block j is the reference for auditing block $k > j$, the certificate computed at k is taken relative to that inherited signal; its difference from the certificate computed against a reference supervision isolates the dependence introduced *between* j and k, localizing a change to a stage rather than to the system as a whole (Section 5.5). We write $Y \triangleq Y_{K_d}$ and drop the stage index when unambiguous. The benchmark of Sections 5.1–5.4 instantiates the single-stage case $k = K_d$ with ground-truth supervision $Y_k = Y$; Section 5.5 instantiates a two-stage pipeline in which Y_k is emitted by an upstream block.

Roles and access sets. Three roles are each defined by the information they hold (Figure 1):

Definition 1 (Access sets)**.** *The* data auditor *at stage k holds $\mathcal{I}_{\mathrm{DA}}(k) = \{(Z_{k-1}, A, Y_k)^{(n)}\}$: the stage's input, the sensitive attribute, and a stage-appropriate supervision signal. The* model developer *holds $\mathcal{I}_{\mathrm{MD}} = \{(X_{[m]}, A, Y)^{(n)}, h\}$ with h the trained predictor, and the* alignment auditor *holds $\mathcal{I}_{\mathrm{AA}} \supseteq \{(X_{[m]}, A, Y)^{(n)}, \hat{Y}\}$, spanning both pipelines. A method is* admissible *for a role if its output is a function of that role's access set.*

Admissibility is restrictive in a useful way. Gradient-, activation-, and prediction-based explanations are inadmissible for the data auditor, since $h \notin \mathcal{I}_{\mathrm{DA}}(k)$; so is any procedure that trains a reference model on the stage's data, because at stage $k < K_d$ the downstream feature space does not yet exist and the auditor is often barred from holding the data to fit one. Conversely, when only $\hat{Y}$ is exposed and the data withheld—the standard third-party position—data-side certificates are inadmissible and prediction-aware ones are not. Neither role dominates; Section 5.3 measures the gap.

The audit questions. Each role asks a different question of the same pipeline; we organize our benchmark around them. **Q1 (data auditor)**: which features of my stage's output drive downstream accuracy and group-fairness violation, without knowing which model will be trained? **Q2 (model**

developer): is that report robust to a model choice not yet made? **Q3 (alignment auditor)**: how much fidelity is forfeited by auditing data without model access? **Q4 (all three)**: can a certificate be issued at governance cadence, and what discretion does the estimator leave its issuer? **Q5 (downstream data auditor)**: what can a stage certify from possibly biased supervision emitted upstream?

4 STAGE-LOCAL AUDIT CERTIFICATES

Shapley aggregation. Let $v : 2^{[m]} \to \mathbb{R}$ be a set measure evaluated on feature coalitions of the audited stage. The Shapley value assigns feature i the average marginal contribution

$$\phi_i(v) = \sum\nolimits_{\pi \in \Pi} \tfrac{1}{m!}\, \Delta_i(\pi), \quad \Delta_i(\pi) \triangleq v(S_i^\pi \cup \{i\}) - v(S_i^\pi), \tag{1}$$

with Π the permutations of $[m]$ and S_i^π the predecessors of i in π; it is the unique linear, symmetric, and efficient rule. A *stage-local certificate* is $\{\phi_i(v)\}_{i\in[m]}$ for any v computable from $\mathcal{I}_{\mathrm{DA}}(k)$.

Approximation and constraints. Equation (1) enumerates $m!$ orderings, so beyond a handful of features a certificate is issued through the unbiased permutation-sampling estimator $\hat{\phi}_i(v) = \frac{1}{T}\sum_{t=1}^{T} \Delta_i(\pi^t)$, π^t uniform (Castro et al., 2009). Uniform sampling assumes all orderings are equally admissible, which a data auditor holding causal knowledge may not want; restricting the support to orderings consistent with a known ancestor–descendant relation gives the asymmetric Shapley value (Frye et al., 2020). The budget T and the constraint set are therefore governance-relevant choices of the issuing party that change the reported numbers (Section 5.4).

Prediction-agnostic and prediction-aware measures. A measure is *prediction-agnostic* (model-agnostic, pre-hoc) if $v(X_S) = f(X_{[m]}, X_S, A, Y)$, and *prediction-aware* if it additionally depends on a restricted model output $\hat{Y}_{|S}$. We say prediction-agnostic rather than prediction-free since the supervision Y is an input to the audit; only the audited *model's* outputs are withheld. Following Frye et al. (2021), the restriction of a probabilistic predictor $h(x_{[m]}) \approx p_{Y|X_{[m]}=x_{[m]}}(1)$ to a coalition is the conditional expectation over the omitted features,

$$h|_{X_S}(x_{[m]}) \triangleq \mathbb{E}_{x_{S^c} \sim p_{X_{S^c}|X_S=x_S}}[h(x_S, x_{S^c})], \tag{2}$$

with output $\hat{Y}_{|S}$. Any prediction-agnostic measure v has a prediction-aware counterpart $v^*(X_S) \triangleq v(\hat{Y}_{|S})$ obtained by substituting $\hat{Y}_{|S}$ for X_S wherever v remains well defined. The pair (v, v^*) is exactly what the data auditor and the alignment auditor can each compute, so we benchmark both.

We use the conditional restriction (2) rather than mean-value or marginal imputation because only it keeps the evaluation on the data manifold and yields consistent removal semantics (Chen et al., 2020; Nafea & Aldarmini, 2026); off-manifold alternatives can inflate attribution error by up to $1.84\times$ on real data (interventional marginalization; Appendix C). Whereas Frye et al. (2021) fit models for this expectation, we use the empirical conditional distribution induced by the measures' equal-frequency quantization: for each test point, h is averaged over training samples whose retained quantized coordinates match x_S, so no auxiliary generative model is introduced (Appendix C.1).

Certificate catalog. Accuracy certificates quantify dependence between Y and a coalition X_S; statistical-parity (SP) certificates dependence between A and X_S; equalized-odds (EO) certificates that dependence conditioned on Y. The information-theoretic certificates are built from MI, CMI, and the partial information decomposition of Bertschinger et al. (2014) into unique and shared information (SI): $v_{A,MI} = I(Y; X_S)$, $v_{A,CMI} = I(Y; X_S \mid A, X_{S^c})$, $v_{SP,MI} = I(A; X_S)$, $v_{SP,SI:1} = SI(A; X_S, Y)$, $v_{SP,SI:2} = SI(X_S; A, Y)$, $v_{EO,CMI} = I(A; X_S \mid Y)$, and the product certificate $v_{D,Prod} = I(A; X_S)\, I(A; X_S \mid Y)\, SI(Y; X_S, A)$ (Covert et al., 2020; Khodadadian et al., 2021; Dutta et al., 2022; Aldarmini & Nafea, 2025). Each is extended along two axes: a Gaussian closed-form surrogate (Barrett, 2015) (subscript g), and four dependence metrics—distance covariance (DC) (Székely et al., 2007), maximum mean discrepancy (MMD) (Gretton et al., 2006), the Hilbert–Schmidt independence criterion (HSIC) (Gretton et al., 2005), and the normalized cross-covariance operator (NOCCO) (Fukumizu et al., 2007)—with their Y-conditional versions, e.g. $\mathrm{CHSIC}_Y(A, B) \triangleq \mathbb{E}_{y\sim p_Y}[\mathrm{HSIC}(p_{A|Y=y}, p_{B|Y=y})]$. With the counterparts v^* this yields 26 measures and 52 certificate variants. Distinct rows can nevertheless coincide after aggregation (Appendix D), which is why the catalog is benchmarked rather than assumed to be diverse.

4.1 A TIGHT ONE-SIDED GUARANTEE

A certificate computed without model access is meaningful only if it constrains models not yet built. Two of the information-theoretic certificates do.

Proposition 1 (Certificates dominate downstream leakage)**.** *Let $\hat{Y}_{|S}$ denote* any *predictor whose inputs are restricted to X_S. Then $v_{SP,MI}(X_S) \geq I(A;\hat{Y}_{|S})$ and $v_{EO,CMI}(X_S) \geq I(A;\hat{Y}_{|S} \mid Y)$, with equality in the first inequality if and only if the Markov chain $A - \hat{Y}_{|S} - X_S$ holds, and in the second if and only if $A - \{\hat{Y}_{|S}, Y\} - X_S$ holds.*

The proof applies the data-processing inequality to the restriction map, with equality following from the stated Markov conditions (Appendix A.1). The bound holds for *every* predictor restricted to X_S—including the conditional restriction (2)—and assumes no model class, no training procedure, and no access to h, so it is computable from $\mathcal{I}_{\mathrm{DA}}(k)$ alone while quantifying over the whole model pipeline. That is the sense in which a data auditor certifies a stage without holding a model. Three further consequences make Proposition 1 the formal content of "stage-local certificate".

Remark 1 (Chaining to fairness gaps)**.** *For binary A with $\pi = P(A=1)$, there is a strictly increasing convex $f : [0,1] \to \mathbb{R}_{\geq 0}$ with $v_{SP,MI}(X_S) \geq I(A;\hat{Y}_{|S}) \geq f\Big(\Delta_{\mathrm{SP}}(\hat{Y}_{|S})\Big)$, and analogously $v_{EO,CMI}(X_S) \geq f(\Delta_{\mathrm{EO}}(\hat{Y}_{|S}))$, where Δ_{SP}, Δ_{EO} are the SP and EO gaps. Admissible choices include $f(\delta) = 8\pi^2(1-\pi)^2\delta^2$ (Hamman & Dutta, 2024, Lemma 1) and two Pinsker-type chains (Cui et al., 2023; Gupta et al., 2021); Appendix A.2 derives all three and states when each is tightest. A small stage-local certificate is therefore a* governance *statement: it caps the realized fairness gap of every downstream predictor built on that coalition.*

Remark 2 (The guarantee is one-sided, and that is the point)**.** *Proposition 1 bounds leakage from above, so a data auditor can certify the* absence *of a fairness channel at a stage but not that a downstream harm will occur: the bound is slack exactly when the model suppresses part of the dependence, and tight only when the model realizes all of it. Stage-local certificates are consequently screening instruments for the data auditor and evidence—not verdicts—for the alignment auditor; the empirical sections quantify how slack the bound is in practice.*

Remark 3 (Inherited supervision)**.** *Write $v(\cdot \mid Y_k)$ for a certificate computed with supervision Y_k emitted by an earlier block. Proposition 1 holds with Y_k in place of Y: $v_{EO,CMI}(X_S \mid Y_k) \geq I(A;\hat{Y}_{|S} \mid Y_k)$. The certificate therefore bounds equalized odds measured against the inherited label, not against the true label Y, and is silent about the latter when Y_k itself depends on A. The SP certificate $v_{SP,MI} = I(A;X_S)$ does not involve the supervision and is invariant to it.*

5 BENCHMARK BY AUDIT QUESTION

Reference attributions. We evaluate certificate *fidelity* against realized downstream behavior. For a downstream predictor with accuracy b_t^A (zero–one loss) and fairness gaps b_t^{SP}, b_t^{EO} on held-out data, we remove each feature i by restricting the predictor to $X_{[m]\setminus\{i\}}$ via (2) and record $d_i^\bullet \triangleq b_t^\bullet - b_i^\bullet$. Both $d_i^\bullet$ and the certificate values are normalized by their maximum absolute value, and the error is the mean absolute deviation across features, $e(v_A) \triangleq \frac{1}{2m}\sum_i |\bar{\phi}_i(v_A) - \bar{d}_i^A| \in [0,1]$, analogously for SP and EO, averaged over 3 seeds.

This reference is a leave-one-out (LOO) effect on the deployed predictor, not a model-based Shapley value: a *reference proxy* rather than a gold standard. It is what a model owner can compute and report, and what an auditor's certificate is asked to anticipate, but it does not satisfy the Shapley axioms and inherits the correlation sensitivity of removal-based attribution (Covert et al., 2021; Kumar et al., 2020). Errors should be read as agreement with this proxy, not distance from truth; Appendix F.4 repeats the comparison where the generative structure is known by construction.

Datasets and implementation. We use seven datasets, as (target, sensitive attribute, samples): three U.S. Census ACS tasks (Ding et al., 2021)—**ACS Income** (income, race, 1.66M), **ACS Employment** (employment, disability, 3.24M), **ACS Public Coverage** (coverage, race, 1.14M)—plus **Adult** (income, gender, 48,842) (Becker & Kohavi, 1996), **Census Income KDD** (income, sex, 399,285) (UCI, 2000), **COMPAS** (recidivism, race, 5,334) (Larson et al., 2016), and **Heritage Health** (health

Table 1: **Q1.** Fidelity of stage-local certificates computable from $\mathcal{I}_{\mathrm{DA}}$: attribution error (%, mean $\pm$ s.d. over 3 seeds) against the NN reference, in row blocks by objective (accuracy, SP, EO). Lowest per-dataset error in **bold**; mean runtime in parentheses. Lower is better.

Certificate (runtime)	ACS Income	ACS Pub. Cov.	ACS Emp.	COMPAS	Income (KDD)	Adult	Health
$v_{A,MI}$ (1.5s)	**4.9 ± 0.3**	**10.4 ± 0.5**	1.2 ± 0.1	**6.3 ± 1.1**	**9.3 ± 1.7**	**8.4 ± 0.6**	**10.5 ± 3.3**
$v_{A,CMI}$ (17.0s)	5.2 ± 0.3	11.3 ± 0.5	**1.0 ± 0.1**	8.0 ± 1.1	9.6 ± 1.9	8.5 ± 0.6	11.5 ± 3.2
$v_{SP,MI}$ (1.5s)	12.4 ± 1.0	9.7 ± 3.0	4.1 ± 1.4	14.4 ± 3.4	23.2 ± 1.2	6.1 ± 2.2	17.0 ± 8.5
$v_{SP,SI:1}$ (1.3m)	12.8 ± 1.0	**8.9 ± 3.0**	4.2 ± 1.4	16.4 ± 3.8	20.0 ± 1.0	9.5 ± 2.0	17.6 ± 6.5
$v_{SP,SI:2}$ (1.4h)	**5.9 ± 1.4**	22.7 ± 2.6	**3.3 ± 1.6**	**7.1 ± 1.8**	**18.7 ± 0.8**	6.7 ± 1.6	**14.1 ± 7.7**
$v_{D,Prod}$ (1.4m)	15.8 ± 1.9	10.0 ± 2.3	3.9 ± 1.2	16.8 ± 3.8	23.5 ± 1.2	**5.9 ± 2.3**	21.1 ± 9.4
$v_{EO,CMI}$ (3.5s)	24.3 ± 3.1	12.7 ± 2.4	7.0 ± 2.5	**17.1 ± 4.6**	**30.6 ± 2.1**	6.2 ± 3.5	**19.6 ± 6.1**
$v_{D,Prod}$ (1.4m)	**21.2 ± 2.8**	**11.6 ± 2.4**	**5.5 ± 2.8**	17.2 ± 4.6	31.8 ± 2.3	**6.1 ± 3.6**	23.1 ± 5.9

index, age, 60,000) (Goldbloom & Hamner, 2011), with five to six features each. Continuous features are quantized into 10 equal-frequency bins; MI and DC use `dit` (James et al., 2018) and SI uses `BROJA_2PID` (Makkeh et al., 2018) with MOSEK (MOSEK ApS, 2019) and ECOS (Domahidi et al., 2013) as backup. Kernel metrics use an RBF kernel with $\sigma = 0.5$; MMD and HSIC are subsampled so the estimated measure variance stays below 2% of its value, and NOCCO uses batches of 512. The neural network (NN) is a fully connected classifier over the quantized features, tuned on a validation split of the 67% training partition over 3 seeds; the six other predictors (CatBoost, LGBM, XGB, Random Forest, Logistic Regression, HistGradientBoosting) use the same partition. Splits, seeds, solvers, and hardware are in Appendix J.1.

5.1 Q1: What the Data Auditor Can Report

Table 1 gives fidelity and cost for the information-theoretic certificates on all seven datasets. For accuracy, $v_{A,MI}$ is best on 6/7 datasets and cheapest, more than $10\times$ faster than $v_{A,CMI}$, which attains the single best value ($1.0 \pm 0.1\%$ on ACS Employment). For SP the picture inverts: $v_{SP,SI:2}$ is best on 5/7 datasets and reaches $3.3 \pm 1.6\%$, but its unique-information program is by far the most expensive (1.4 hours against under two minutes for every other certificate), since the high-dimensional coalition X_S is the reference variable. For EO, $v_{D,Prod}$ is best on 4/7 datasets ($5.5 \pm 2.8\%$) at 1.4 minutes, against 3.5 seconds for $v_{EO,CMI}$, which wins elsewhere.

Two extensions matter for the data auditor's choice. MMD/CMMD gives the most consistent gains among alternative metrics—8 of 21 errors reduced, as low as $1.6 \pm 0.2\%$ (accuracy) and $2.6 \pm 1.6\%$ (SP) on ACS Employment—while NOCCO is the most stable (16 of its 19 relative increases below 100%, no severe degradation). DC and HSIC are unusable pre-hoc: relative increases reach 3331% and 2602%, and every DC certificate exceeds 28% absolute error. The Gaussian surrogates are dataset-dependent but never catastrophic, holding error below 25% for 45 of 56 configurations at seconds per certificate. Tables 5–7 report all three families, and Appendix F.4 repeats the comparison on a synthetic pipeline with known generative structure.

Why the instruments separate. The ordering across metrics is not arbitrary, and one property explains it. A Shapley certificate differences a set measure across coalitions of *different sizes*, so any part of a measure's value that varies with $|S|$ rather than with the dependence it is meant to capture enters the attribution as bias. DC and HSIC are unnormalized statistics whose scale grows with the dimension of the compared vector, so their marginal contributions confound added dependence with added dimension—which is why they fail pre-hoc by an order of magnitude rather than marginally. NOCCO normalizes by the covariance operators, the natural reading of its being the most stable metric we test; MMD/CMMD keeps a bounded kernel-mean form while measuring exactly the group-conditional discrepancy the SP and EO gaps are defined by, matching its advantage on those objectives. The same mechanism predicts the inversion in Section 5.3: prediction-aware restriction replaces X_S by the scalar $\hat{Y}_{|S}$, fixing the compared dimension at one for every coalition, after which DC becomes the most faithful SP metric we observe (Appendix F.7 gives a closed form for a Gaussian instance). This mechanism fits the measurements but is not separately tested.

5.2 Q2: Does the Report Survive the Model Developer's Choice?

A certificate is issued before the downstream model exists, so its value depends on whether its fidelity is stable across model classes. We therefore make the multi-predictor evaluation the *main*

Table 2: **Q2.** Main-benchmark fidelity across seven heterogeneous downstream predictors on the three large-scale ACS datasets: attribution error (%) against the NN, and the mean ± s.d. and min–max range across six tree/linear predictors (CatBoost, LGBM, XGB, Random Forest, Logistic Regression, HistGradientBoosting). A certificate is useful to a data auditor to the extent that the range is narrow relative to the mean.

	ACS Income			ACS Public Coverage			ACS Employment		
Certificate	NN	Mean (6)	Range	NN	Mean (6)	Range	NN	Mean (6)	Range
$v_{A,MI}$	4.9	6.6 ± 1.4	4.8–8.4	10.4	13.4 ± 2.8	10.1–17.4	1.2	1.3 ± 0.6	0.9–2.6
$v_{A,CMI}$	5.2	6.9 ± 1.4	5.1–8.7	11.3	14.2 ± 2.8	11.0–18.2	1.0	1.0 ± 0.6	0.6–2.3
$v_{SP,MI}$	12.4	16.7 ± 2.9	12.8–20.1	9.7	14.5 ± 5.4	9.1–22.6	4.1	4.0 ± 1.1	2.7–5.5
$v_{SP,SI:1}$	12.8	17.2 ± 2.7	13.6–21.2	8.9	14.1 ± 5.0	8.5–21.8	4.2	3.9 ± 1.0	2.7–5.4
$v_{SP,SI:2}$	5.9	9.5 ± 2.8	6.0–13.3	22.7	20.7 ± 3.0	15.3–25.0	3.3	2.9 ± 0.8	2.0–4.2
$v_{D,Prod}$	15.8	18.7 ± 2.9	14.6–21.9	10.0	14.9 ± 4.6	9.9–22.2	3.9	3.9 ± 1.1	2.6–5.5
$v_{EO,CMI}$	24.3	29.6 ± 5.8	23.6–41.7	12.7	16.8 ± 3.5	13.0–23.2	7.0	6.5 ± 2.7	3.0–10.1
$v_{D,Prod}$	21.2	27.9 ± 7.5	20.1–43.5	11.6	15.4 ± 3.3	11.9–21.6	5.5	5.1 ± 2.7	2.1–8.8

fidelity benchmark rather than a robustness appendix: Table 2 reports errors on the three large-scale ACS datasets against seven predictors spanning neural, boosted-tree, bagged-tree, and linear families, with the neural network reported alongside the other six so the two are directly comparable.

The certificates transfer, with clear structure. On ACS Income and ACS Public Coverage, accuracy certificates are the most stable across predictors (standard deviation at most 21.2% of mean error), SP certificates the least (up to 29.5% and 37.2%), with EO in between. On ACS Employment those ratios rise to 60%, but only as an artifact of the denominator: absolute errors there are 0.6–2.6%. The relevant reading is the range column: across six unseen model families the standard deviation is 0.5–2.9× the seed standard deviation of one family for the fairness certificates (larger only for accuracy, whose seed spread is near zero), and the best certificate within an objective is preserved except for SP on ACS Public Coverage under Random Forest and Logistic Regression, and for EO on ACS Income under Logistic Regression, the consistent outlier there (41.7% and 43.5% against 20–29% for the tree ensembles), in the expected direction: a linear model realizes a different subset of the available dependence, and by Remark 2 the certificate bounds rather than predicts what it realizes. Per-predictor errors are in Tables 11–13. Operationally, a certificate can be issued without naming the downstream model, provided it reports a ranking with an error band rather than a point estimate, and a later change of model family does not invalidate it.

5.3 Q3: The Price of Role Separation

For a certificate v with prediction-aware counterpart v^*, define the *pre-hoc attribution gap*

$$\mathrm{gap}(v) = \big(e(v) - e(v^*)\big)/e(v^*). \tag{3}$$

A positive gap is the fidelity forfeited by the data auditor relative to the alignment auditor; a negative gap means the model-agnostic certificate is the more faithful of the two.

The gap is real and heterogeneous (Figure 3). Over datasets and objectives, the model-agnostic certificate is more faithful than its prediction-aware counterpart in 7/56 configurations for accuracy, 15/84 for SP, and 13/56 for EO: model access helps in 82% of cases, and for accuracy almost always. The magnitude is metric-dependent in a way that matters for governance. Prediction-aware restriction rescues certificates that fail outright without model access: $v^*_{A,DC}$ attains the lowest post-hoc accuracy error we observe ($0.5 \pm 0.3\%$ on ACS Employment) with a pre-hoc gap of 7625%, and $v^*_{SP,DC}$ is best on 4/7 datasets at $3.2 \pm 1.8\%$, though their pre-hoc counterparts exceed 40% error. Conversely, HSIC certificates are *worse* with model access on at least 5 of 7 datasets (gaps as low as -65.6%). No prediction-aware certificate runs in under 10 seconds—a cost the data auditor avoids.

The alignment-relevant conclusion is asymmetric. For the certificates a data auditor should actually use—MI, CMI, PID, MMD, NOCCO—the gap stays below 800%, and above 300% only where post-hoc error is under 5%; every gap above 1000% is DC, a poor pre-hoc metric (Tables 8–10, Appendix F.6). The price of role separation is thus small if the data auditor selects the certificate for the access regime it operates in, and catastrophic otherwise—the practical content of Definition 1.

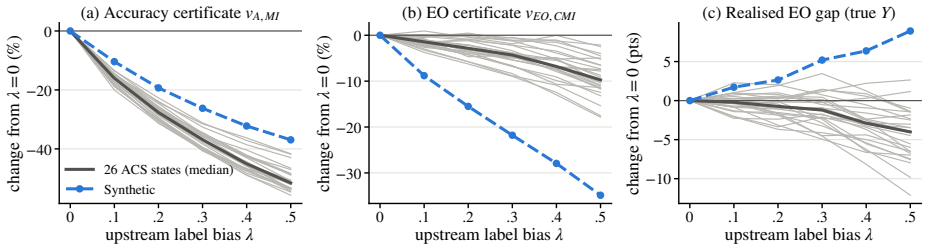

Figure 2: Certificates computed on inherited supervision Y_2 as the upstream label bias λ grows; gray: each of 26 ACS states, dark: their median, dashed blue: synthetic pipeline. (a, b) Both label-dependent certificates fall everywhere; SP (not shown) is unchanged. (c) The realised EO gap against the true label (mean of NN, HGB, LR) falls in most states and rises in the synthetic pipeline.

5.4 Q4: Issuing a Certificate at Audit Cadence

The studies above hold the feature set small enough for exact aggregation; governance does not, since a certificate must be re-issued whenever a stage changes and the width of a representation boundary is set by the pipeline, not the auditor. We therefore audit ACS Public Coverage at 20 features—the 18 of Ding et al. (2021) plus `GCL` and `COW`—with the Gaussian certificates v_{A,MI_g}, v_{SP,MI_g}, v_{EO,CMI_g}, taking exact Shapley values as the reference for the sampled ones.

Exact aggregation at this width costs 71 hours for the accuracy and SP certificates and 150 for EO—an annual exercise, not a release gate. Permutation sampling drives approximation error below 1% within 30–70 minutes, a speedup of over $60\times$, so the binding cost of certification is the estimator, not the access constraint: a changed stage can be re-certified overnight without holding a model (Appendices H.1, F.9–F.10; the full cost–fidelity scaling is in Anonymous Authors, 2026).

The declared constraints matter more than the budget. From the ancestor–descendant pairs C1: (AGEP, SCHL) through C6: (MIG, PINCP) we form nested constraint sets CK1$=\{$C1$\}$ through CK5$=\{$C1–C6$\}$ and re-run sampling under each. At a fixed 6000-second budget, error against the downstream reference is lowest under CK3 for v_{A,MI_g} and v_{SP,MI_g} and under CK1–CK2 for v_{EO,CMI_g}, while the fully constrained CK5 is the *worst* configuration for the accuracy and SP certificates. Causal knowledge is therefore not monotonically helpful, and an issuing party that reports only a number is reporting a choice as if it were a measurement (Appendix H.2).

The choice is nevertheless auditable; Appendix H.2 shows how a third party holding the declared constraint set can check it, which makes the discretion in (P5) reviewable.

5.5 Q5: Certifying a Downstream Stage Under Inherited Supervision

Design. An upstream annotation block B_1 emits the supervision Y_2 from the ground truth Y with two-sided label bias of strength $\lambda \in \{0, 0.1, \ldots, 0.5\}$: negatives with $A = 1$ are relabelled positive and positives with $A = 0$ negative, each with probability λ. The data auditor at B_2 holds $(X_{[m]}, A, Y_2)$, never Y. For each λ we compute the certificates on Y_2 and, as a reference only an alignment auditor could hold, on Y (exact aggregation), train a neural network, HistGradient-Boosting and logistic regression on $(X_{[m]}, Y_2)$, and measure their EO gap against the true Y on held-out data. We run ACS Income (features `COW`, `SEX`, `WKHP`, `SCHL`; A = race (White vs. Black); $Y = \mathbf{1}[\text{income} > \$50\text{k}]$; 10,000 samples, 5 seeds) in each of the 26 states with the most minority-group positives, where an EO gap is estimable, and a synthetic pipeline with balanced groups and a strong proxy for A (10 seeds). Hypotheses were fixed before the state runs (Appendix I).

The downstream stage detects that its supervision changed. Both label-dependent certificates fall as the upstream block injects bias (Figure 2a,b). The accuracy certificate falls strictly monotonically in all 26 states, by 42–56% at $\lambda = 0.5$. The EO certificate falls in all 26 (sign test $p = 3 \times 10^{-8}$; paired t-test $p < 0.05$ in 23), by 2–18% with median 9.7%, and by 35% in the synthetic pipeline. The difference between the certificate on Y_2 and on Y is thus a stage-attributable footprint of B_1—the localization of Section 3—but it requires the reference Y, which the downstream auditor lacks. The SP certificate is identical under Y_2 and Y in every state and seed (Remark 3): an SP audit cannot see label bias introduced upstream. Accuracy attribution, usually a step toward the fairness certificates, is useful on its own: it is the most sensitive detector of corrupted supervision we observe.

The certificate does not reveal the direction of downstream harm. The harm the EO certificate is meant to anticipate does not follow it (Figure 2c). In the synthetic pipeline the realised EO gap against the true label rises by 8.9 points while the certificate on Y_2 falls by 35%: an auditor holding only the inherited label would certify the stage as increasingly fair as it becomes increasingly harmful. In the states the gap falls in 24 of 26, as downstream accuracy against Y drops from about 0.74 to 0.46–0.71. Neither proxy strength $I(A; X_{[m]})$ (pre-specified; Spearman $\rho = +0.05$ across states, $p = 0.82$) nor the loss of downstream accuracy (exploratory; $\rho = -0.03$, $p = 0.90$) predicts which way the gap moves. Remark 3 gives the reason: on inherited supervision the certificate bounds EO relative to Y_2, so its decrease says nothing about EO relative to Y. A stage-local certificate is interpretable only together with the provenance of its supervision—which block emitted Y_k, and whether a reference exists—and that record belongs in the certificate (P5).

6 Discussion

Guidelines for measure construction. The benchmark supports no single ranking of certificates; it supports five construction guidelines, each indexed by the access regime the measure is used in. They are guidelines rather than principles: each is an inference from the benchmark reported here, under its stated evaluation assumptions, not a claim established in general. **(P1) Construct for the access set, not for the objective.** The ranking of dependence metrics inverts under prediction-aware restriction—DC moves from worst to best for SP—so a measure validated for an alignment auditor is not a measure for a data auditor. **(P2) Prefer a certificate that carries a one-sided guarantee when the downstream model is unknown.** MI and CMI certificates bound the leakage of every restricted predictor (Proposition 1); PID and kernel certificates buy fidelity without that quantifier, and a certificate that drops a PID term is wrong in direction rather than merely loose—a Bayes-optimal predictor can violate EO (Appendix E). **(P3) Build the fairness objective into the measure by conditioning, not by correction.** Conditioning on Y yields EO certificates directly, whereas representation-level correction acts after the stage it was meant to audit; SP and EO are incompatible already at the data stage (Appendix B), so the objective must be chosen when the measure is built. Conditioning is only as good as the label conditioned on: on inherited supervision the EO certificate moves independently of downstream harm (Section 5.5). **(P4) Match estimator cost to reporting cadence, and report a band rather than a point**, since fidelity typically varies more across model families than across seeds. **(P5) Report the estimator's discretion as part of the certificate.** Beyond a handful of features the certificate is a sampled quantity under declared causal constraints that move fidelity non-monotonically, so budget, constraint set, and the provenance of the supervision belong in the certificate, where a third party can check them (Appendices F.2–F.3, H.2). Appendix G.2 instantiates the guidelines as a certificate choice for each role.

Why privacy and safety push auditing upstream. Data-minimization, federation, and siloing frequently place the party holding A apart from the party holding h, so stage-local certificates are compatible with regimes in which no single party can assemble the joint distribution post-hoc auditing presumes; Appendix G.3 develops this, and claims compatibility only.

Limitations. We state a few limitations of our work. First, the reference attributions are LOO effects on a trained predictor, so reported errors measure agreement with a removal-based proxy, not distance from a ground-truth attribution; a model-based Shapley reference would be stronger and costs exactly the factor that motivates pre-hoc auditing (Appendix F.4). Second, our main datasets are tabular with five to six features and the cadence study reaches only 20; coalition enumeration is exponential, and we do not claim the numbers extrapolate to high-dimensional learned representations; extending these certificates to the encoder outputs of modern pipelines is the natural next step. Third, the benchmark of Sections 5.1–5.4 reports means over 3 seeds without significance testing (only the two-stage study tests), and several choices—$\sigma = 0.5$, NOCCO batch size—are fixed rather than swept, so differences of a few points should not be over-read (Appendices F.2, F.8). Fourth, and most consequentially, the guarantee is one-sided (Remark 2): a clean certificate is evidence that a stage does not supply a fairness channel, not that the deployed system is fair—a lower bound on what an alignment auditor should demand, not a replacement for auditing the model. Fifth, the staged formulation is exercised only in its simplest multi-stage form: one upstream block, a known label-bias mechanism, and tabular features (Section 5.5).

AI USE STATEMENT

We used generative AI tools for editing assistance, including reformatting our written manuscript and copy-editing prose for grammar and concision, and for code assistance. We did not use them for research ideation, for deriving or checking the theoretical results, or for analyzing or interpreting experimental results; all measures, experiments, and numbers reported here were designed, implemented, and run by the authors. Retrieval and citation assistance was verified manually: every reference was checked against the published record. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only public benchmark datasets (ACS/Folktables, UCI Adult, Census Income KDD, ProPublica COMPAS, Heritage Health) under their published terms, with no attempt at re-identification. Two considerations are specific to our contribution. First, the guarantee of Proposition 1 is one-sided: a small certificate bounds the fairness gap a downstream model can realize, but cannot establish that a deployed system is fair, and must not be used as a compliance verdict or a substitute for a model-level audit. Second, the certificates require the auditing party to hold the sensitive attribute A; the framework is meant for parties that already hold it lawfully for fairness assessment, and neither requires nor justifies collecting it otherwise. COMPAS is used as a measurement benchmark, not an endorsement of recidivism prediction.

REPRODUCIBILITY STATEMENT

Assumptions and complete proofs are in Section 4 and Appendix A; the empirical protocol is in Section 5 and Appendix J.1, and the two-stage study, including its data checks and pre-specified hypotheses, in Appendix I. Anonymized code for the two-stage study is provided as supplementary material.

REFERENCES

Sokrat Aldarmini and Mohamed S. Nafea. Information-theoretic quantification of inherent discrimination bias in training data for supervised learning. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2025.

Anonymous Authors. Shapley data auditing for explainable feature attribution: Scalability analysis, 2026. Concurrent anonymous submission; included as supplementary material.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

Adam B Barrett. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Physical Review E*, 91(5):052802, 2015.

Barry Becker and Ronny Kohavi. Adult. *UCI Machine Learning Repository*, 1996.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

Miles Brundage, Noemi Dreksler, Aidan Homewood, et al. Frontier AI auditing: Toward rigorous third-party assessment of safety and security practices at leading AI companies. *arXiv preprint arXiv:2601.11699*, 2026.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18. IEEE, 2009.

Stephen Casper, Xander Davies, Claudia Shi, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023.

Stephen Casper, Carson Ezell, Charlotte Siegmann, et al. Black-box access is insufficient for rigorous AI audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024. doi: 10.1145/3630106.3659037.

Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

Alan Chan, Carson Ezell, Max Kaufmann, et al. Visibility into AI agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.

Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*, 2020.

Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17212–17223, 2020.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

Yue Cui, Chen Ma, Kai Zheng, Lei Chen, and Xiaofang Zhou. Controllable universal fair representation learning. In *Proceedings of the ACM Web Conference 2023 (WWW)*, pp. 949–959, 2023. doi: 10.1145/3543507.3583307.

Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. Proxy non-discrimination in data-driven systems. *arXiv preprint arXiv:1707.08120*, 2017.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Alexander Domahidi, Eric Chu, and Stephen Boyd. ECOS: An SOCP solver for embedded systems. In *Proceedings of the 2013 European Control Conference*, pp. 3071–3076. IEEE, 2013.

Sanghamitra Dutta, Praveen Venkatesh, and Pulkit Grover. Quantifying feature contributions to overall disparity using information theory. *arXiv preprint arXiv:2206.08454*, 2022.

Cynthia Dwork and Christina Ilvento. Fairness under composition. In *Proceedings of the 10th Innovations in Theoretical Computer Science Conference (ITCS)*, 2019. doi: 10.4230/LIPIcs.ITCS.2019.33.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. doi: 10.1561/0400000042.

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric Shapley values: Incorporating causal knowledge into model-agnostic explainability. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, pp. 10835–10866. PMLR, 2023.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021. doi: 10.1145/3458723.

Amirata Ghorbani and James Zou. Data Shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning (ICML)*, pp. 2242–2251. PMLR, 2019.

Anthony Goldbloom and Ben Hamner. Heritage Health Prize. *Kaggle*, 2011.

Ryan Greenblatt, Carson Denison, Benjamin Wright, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert–Schmidt norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory (ALT)*, volume 3734 of *Lecture Notes in Computer Science*, pp. 63–77, 2005. doi: 10.1007/11564089_7.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample problem. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

Umang Gupta, Aaron M Ferber, Bistra Dilkina, and Greg Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):7610–7619, 2021.

Faisal Hamman and Sanghamitra Dutta. Demystifying local and global fairness trade-offs in federated learning using partial information decomposition. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021. doi: 10.1145/3442188.3445901.

Ryan G James, Christopher J Ellison, and James P Crutchfield. "dit": A Python package for discrete information theory. *Journal of Open Source Software*, 3(25):738, 2018.

Sajad Khodadadian, Mohamed Nafea, AmirEmad Ghassami, and Negar Kiyavash. Information theoretic measures for fairness-aware feature selection. *arXiv preprint arXiv:2106.00772*, 2021.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, pp. 1885–1894. PMLR, 2017.

I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with Shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning (ICML)*, pp. 5491–5500. PMLR, 2020.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. ProPublica COMPAS analysis—Data and analysis for machine bias. *ProPublica GitHub repository*, 2016.

Abdullah Makkeh, Dirk Oliver Theis, and Raul Vicente. BROJA-2PID: A robust estimator for bivariate partial information decomposition. *Entropy*, 20(4):271, 2018.

Samuel Marks, Johannes Treutlein, Trenton Bricken, et al. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965*, 2025.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

MOSEK ApS. *MOSEK Optimization Suite*, 2019.

Mohamed Nafea and Sokrat Aldarmini. On the role of restricted predictors in Shapley-based fairness auditing. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, 2026.

NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical Report NIST AI 100-1, NIST, 2023.

Guilherme Dean Pelegrina, Miguel Couceiro, and Leonardo Tomazeli Duarte. A preprocessing Shapley value-based approach to detect relevant and disparity prone features in machine learning. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 279–289, 2024. doi: 10.1145/3630106.3658905.

Drago Plecko and Elias Bareinboim. Causal fairness analysis: A causal toolkit for fair machine learning. *Foundations and Trends in Machine Learning*, 17(3):304–589, 2024. doi: 10.1561/2200000106.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, et al. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, 2020. doi: 10.1145/3351095.3372873.

Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. Outsider oversight: Designing a third party audit ecosystem for AI governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2022. doi: 10.1145/3514094.3534181.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M. Aroyo. "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. doi: 10.1145/3411764.3445518.

Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2:307–317, 1953.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024. doi: 10.1038/s41586-024-07566-y.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 180–186, 2020. doi: 10.1145/3375627.3375830.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

Census-Income (KDD). UCI Machine Learning Repository, 2000. doi: 10.24432/C5N30T.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning (ICML)*, pp. 325–333. PMLR, 2013.

APPENDIX

The appendices below describe the proofs, the full benchmark tables, and the sensitivity, known-generative-structure, and cadence studies referenced from the main text.

A PROOFS OF THE MAIN-PAPER RESULTS

This section proves Proposition 1 and derives the chains from the certificates to the SP and EO gaps stated in Remark 1. Information is measured in nats throughout.

A.1 Proof of Proposition 1

Proposition 2 (Certificates dominate downstream leakage; restatement of Proposition 1)**.** *Let $\hat{Y}_{|S}$ denote any predictor whose inputs are restricted to X_S. Then*

$$v_{SP,MI}(X_S) \geq I(A; \hat{Y}_{|S}),$$
$$v_{EO,CMI}(X_S) \geq I(A; \hat{Y}_{|S} \mid Y),$$

with equality in the first inequality if and only if the Markov chain $A - \hat{Y}_{|S} - X_S$ holds, and in the second if and only if $A - \{\hat{Y}_{|S}, Y\} - X_S$ holds.

Proof. Because $\hat{Y}_{|S}$ is generated from X_S alone, with any internal randomization independent of $(A, Y, X_{[m]})$, it is conditionally independent of (A, Y) given X_S. This gives both $A - X_S - \hat{Y}_{|S}$ and $A - \{X_S, Y\} - \hat{Y}_{|S}$ by construction, with no assumption on the causal position of A (in particular, A need not be a root variable; cf. Kusner et al., 2017). The data-processing inequality applied to the first chain gives

$$v_{SP,MI}(X_S) = I(A; X_S) \geq I(A; \hat{Y}_{|S}),$$

with equality if and only if the reversed chain $A - \hat{Y}_{|S} - X_S$ also holds, that is, when $\hat{Y}_{|S}$ is a sufficient statistic of X_S for A. Applying the data-processing inequality to the second chain, conditionally on each value of Y, gives

$$v_{EO,CMI}(X_S) = I(A; X_S \mid Y) \geq I(A; \hat{Y}_{|S} \mid Y),$$

with equality if and only if $A - \{\hat{Y}_{|S}, Y\} - X_S$ holds. □

A.2 Chaining to SP and EO gaps

Let $\hat{Y}$ be the binary decision derived from $\hat{Y}_{|S}$. It is again a function of X_S, so Proposition 1 gives $v_{SP,MI}(X_S) \geq I(A; \hat{Y})$, and chaining reduces to lower-bounding $I(A; \hat{Y})$ by an increasing function of the SP gap $\Delta_{\mathrm{SP}} = \max_{a,a'} |P(\hat{Y} = 1 \mid A = a) - P(\hat{Y} = 1 \mid A = a')|$. Three such bounds are available.

Multi-class A. For $A \in \{1, \ldots, n\}$ with $\pi_i = P(A = i)$, the Pinsker-based result of Cui et al. (2023, Lemma 9) gives

$$I(A; \hat{Y}) \geq \min_i \pi_i^2 \, \Delta_{\mathrm{SP}}^2, \qquad \text{equivalently} \qquad \Delta_{\mathrm{SP}} \leq \frac{\sqrt{I(A; \hat{Y})}}{\min_i \pi_i},$$

so $f(\delta) = \min_i \pi_i^2 \delta^2$ is admissible.

Binary A, refined Pinsker. For $A \in \{0, 1\}$ with $\pi = P(A = 1)$, Gupta et al. (2021, Theorem 2) write $I(A; \hat{Y}) = (1 - \pi)\,\mathrm{KL}(p_{\hat{Y}|A=0} \| p_{\hat{Y}}) + \pi\,\mathrm{KL}(p_{\hat{Y}|A=1} \| p_{\hat{Y}})$, note that the two L_1 distances to the mixture are πV and $(1-\pi)V$ with $V = \|p_{\hat{Y}|A=1} - p_{\hat{Y}|A=0}\|_1$, and lower-bound each divergence by

$$f_{\mathrm{KL}}(V) = \max\left\{ \log \frac{2+V}{2-V} - \frac{2V}{2+V},\ \frac{V^2}{2} + \frac{V^4}{36} + \frac{V^6}{288} \right\},$$

which gives $I(A; \hat{Y}) \geq g(\pi, V)$ with $g(\pi, V) = (1-\pi) f_{\mathrm{KL}}(\pi V) + \pi f_{\mathrm{KL}}((1-\pi)V)$. For a binary decision $V = 2\Delta_{\mathrm{SP}}$, and g is increasing in V, so $f(\delta) = g(\pi, 2\delta)$ is admissible. (The statement in Gupta et al. (2021) is written in terms of the parity gap; substituting $V = 2\Delta_{\mathrm{SP}}$ makes the dependence on the gap explicit.)

Binary A, closed form. Hamman & Dutta (2024, Lemma 1) give

$$\Delta_{\mathrm{SP}} \leq \frac{\sqrt{0.5\, I(A; \hat{Y})}}{2\pi(1-\pi)}, \qquad \text{i.e.} \qquad f(\delta) = 8\pi^2(1-\pi)^2\delta^2.$$

Which bound is tightest. For binary A, the second term of f_{KL} gives $f_{\mathrm{KL}}(V) \geq V^2/2$, so

$$g(\pi, 2\delta) \;\geq\; 2\pi(1-\pi)\,\delta^2 \;\geq\; 8\pi^2(1-\pi)^2\delta^2 \;\geq\; \min(\pi, 1-\pi)^2\,\delta^2,$$

where the middle inequality is $4\pi(1-\pi) \leq 1$ and the last holds because $8\max(\pi, 1-\pi)^2 \geq 2 \geq 1$. The refined-Pinsker route is therefore the tightest for binary A, the bound of Hamman & Dutta (2024) is the simplest closed form, and the bound of Cui et al. (2023) is the only one of the three that covers multi-class A. All three f are strictly increasing and convex on $[0, 1]$.

EO. Write $q_y = P(Y = y)$, $\pi_y = P(A = 1 \mid Y = y)$, and $\Delta_y = |P(\hat{Y} = 1 \mid A = 1, Y = y) - P(\hat{Y} = 1 \mid A = 0, Y = y)|$, so that $\Delta_{\mathrm{EO}} = \max_y \Delta_y$. Since $I(A; \hat{Y} \mid Y) = \sum_y q_y\, I(A; \hat{Y} \mid Y = y)$ and each term obeys the bounds above with π replaced by π_y,

$$v_{EO,CMI}(X_S) \;\geq\; I(A; \hat{Y} \mid Y) \;\geq\; q_{y^\star}\, f_{\pi_{y^\star}}(\Delta_{\mathrm{EO}}), \qquad y^\star = \arg\max_y \Delta_y,$$

where f_π is any of the binary-A bounds evaluated at π. With the closed form, $v_{EO,CMI}(X_S) \geq \min_y 8 q_y \pi_y^2 (1-\pi_y)^2\, \Delta_{\mathrm{EO}}^2$, which is the analogous strictly increasing convex f of Remark 1.

B Why Conditioning, Not Correction, Builds the Certificate

This section shows that SP and EO independence cannot in general hold together at the data stage, which is the basis of guideline P3.

Lemma 1 (Data-side incompatibility)**.** *Let X_S be a feature set, and let the label Y and the sensitive attribute A be binary, with all conditioning events of positive probability. Then the joint condition $\{X_S \perp A$ and $X_S \perp A \mid Y\}$ implies either (i) $X_S \perp Y$, or (ii) $A \perp Y$.*

Proof. For $a \in \{0, 1\}$ and $x \in \mathcal{X}_S$, the law of total probability and $X_S \perp A \mid Y$ give

$$\begin{aligned} p_{X_S|A=a}(x) = {} & p_{X_S|Y=1}(x) p_{Y|A=a}(1) \\ & + p_{X_S|Y=0}(x) p_{Y|A=a}(0). \end{aligned}$$

Independence $X_S \perp A$ equates this expression for $a = 0$ and $a = 1$. Substituting $p_{Y|A=a}(0) = 1 - p_{Y|A=a}(1)$ and rearranging yields, for every x,

$$\begin{aligned} & \big(p_{X_S|Y=1}(x) - p_{X_S|Y=0}(x)\big) \\ & \quad \cdot \big(p_{Y|A=0}(1) - p_{Y|A=1}(1)\big) = 0. \end{aligned}$$

If the second factor vanishes, $A \perp Y$ because Y is binary. Otherwise the first factor vanishes for every x, so $p_{X_S|Y=1} = p_{X_S|Y=0}$ and $X_S \perp Y$. □

Both alternatives are testable before training: the first says the audited features are uninformative for the target, the second that the groups have equal base rates. Outside these cases no data-side correction can deliver both independences, so the fairness objective has to be fixed when the measure is built; conditioning on Y builds EO into $v_{EO,CMI}$ directly.

C Why the Restricted Predictor Must Be Conditional

This section motivates the conditional restriction of Eq. (2) and describes its implementation.

Three of the feature-removal strategies catalogued by Covert et al. (2021) are relevant here: mean-value imputation, interventional marginalization, and conditional expectation. Only the conditional restriction obeys subset-function consistency with the data distribution (Covert et al., 2021); mean-value and interventional constructions can evaluate the predictor off the data manifold. Such off-manifold evaluations can conceal sensitive dependence: predictors that agree on the data manifold can receive different off-manifold attributions (Frye et al., 2021). Under both SP and EO, the conditional restriction has the blocking property that it returns zero measured bias when the observed features are independent of the sensitive attribute; the alternatives do not satisfy this property in general (Nafea & Aldarmini, 2026).

Nafea & Aldarmini (2026) also compare the three restrictions empirically, on 320 synthetic DAGs and seven real datasets. Among their DAGs, 175 are distinct SP–EO DAGs under the conditional restriction, with lowest SP and EO errors of 20.7% and 35.3%; on the best 50% of DAGs, the corresponding values are 11.3% and 19.3%. Under mean-value and interventional restrictions, the number of distinct SP–EO DAGs falls from 175 to 120; the best SP error changes from 20.7% to 22.8% and 21.7%, respectively, while the best EO error changes from 35.3% to 34.4% under interventional restriction. On ACS Income, the errors of $v_{SP,SI:2}$ under the conditional restriction are 6.9% for SP and 11.3% for EO; mean-value restriction gives 7.9% and 17.4%, whereas interventional restriction gives 6.9% and 19.2%. On ACS Public Coverage, the EO error of $v_{SP,SI:2}$ is 21.1% conditionally and rises to 36.8% and 38.8% under mean-value and interventional restrictions, respectively. Across their real-data comparison, replacing the conditional restriction with mean-value imputation inflates attribution error by up to $1.75\times$ (Nafea & Aldarmini, 2026); computed from their reported errors, interventional restriction inflates it by up to $1.84\times$ (the ACS Public Coverage case above), although the effect varies in sign across measures and datasets.

C.1 Empirical-conditional implementation

The restriction used here is

$$h|_{X_S}(x_{[m]}) = \mathbb{E}_{X_{S^c} \sim p_{X_{S^c}|X_S = x_S}}[h(x_S, X_{S^c})],$$

with output $\hat{Y}_{|S}$. It is implemented by equal-frequency quantization and quantized-match averaging: for each test point, the predictor is averaged over training samples whose retained, quantized coordinates match x_S. This empirical conditional construction uses no auxiliary generative model: whereas Frye et al. (2021) learn models to estimate the expectation, it uses the plug-in estimate of $p_{X_{S^c}|X_S}$ at the quantization resolution, so its accuracy depends on how well the quantized joint distribution is estimated.

D When Distinct Certificates Coincide

This section shows when distinct measures yield identical Shapley attributions.

Let $v(S) = v(S^c)$ for all coalitions. For $i \in [m]$, write

$$\phi_i(v) = \sum_{S \subseteq [m] \setminus \{i\}} \omega(S)\big[v(S \cup \{i\}) - v(S)\big],$$
$$\omega(S) = \frac{(m - 1 - |S|)!\,|S|!}{m!}.$$

Pair every coalition $S \subseteq [m] \setminus \{i\}$ with $T = S^c \setminus \{i\}$, so that $|T| = m - 1 - |S|$ and $\omega(T) = \omega(S)$. Since $T \cup \{i\} = S^c$ and $T = (S \cup \{i\})^c$, bipartition symmetry gives $v(T \cup \{i\}) - v(T) = v(S) - v(S \cup \{i\})$: the paired marginal contributions are negatives of each other, so $\phi_i(v) = -\phi_i(v)$ and hence $\phi_i(v) = 0$. By linearity of the Shapley value, two measures that differ only by a bipartition-symmetric term, such as the shared or synergistic information that X_S and X_{S^c} carry about A, therefore induce identical attributions, although the set functions themselves differ.

E A Faulty EO Certificate as a Negative Control

This section gives the counterexample behind the faulty EO measure $CI(A; X_S, Y)$, the synergistic information of the partial information decomposition with A as the target and (X_S, Y) as sources. It is used as a negative control in Appendix F.4.

Consider $A \to X \to Y$ with one feature $X = A$ and $A \sim \mathrm{Bern}(0.5)$. Let $Y \sim \mathrm{Bern}(p_0)$ when $X = 0$ and $Y \sim \mathrm{Bern}(p_1)$ when $X = 1$, where $0 < p_0 < p_1 < 1$. Because $A \perp Y \mid X$,

$$0 = I(A; Y \mid X) = UI(A; Y \setminus X) + CI(A; X, Y),$$

and both terms are non-negative, so $CI(A; X, Y) = 0$: the measure reports no EO-relevant dependence. Now let $\hat{Y} \sim \mathrm{Bern}(p_X)$ be drawn independently given X, so that $p_{\hat{Y}|X=x}(1) = p_{Y|X=x}(1)$ for every x. Then

$$p_{\hat{Y}|A=0,Y=1}(1) = p_0 \neq p_1 = p_{\hat{Y}|A=1,Y=1}(1),$$

so a predictor that reproduces the Bayes posterior violates EO; the deterministic Bayes classifier $\mathbf{1}[p_X > 1/2]$ violates it as well when $p_0 < 1/2 < p_1$, with EO gap 1. The measure is therefore wrong in direction rather than merely loose, which is why it is a negative control rather than an EO certificate.

F Estimator Details and Sensitivity

This section records the estimator choices behind the benchmark, their sensitivity, and their cost.

F.1 PID computation

Three PID solvers were assessed. The `dit` and `ComputeUI` approaches become impractical as the variable count or support grows; `BROJA_2PID`, formulated as an exponential-cone program, takes tens of seconds per evaluation of a PID-based unfairness measure rather than more than an hour. MOSEK is paired with ECOS as backup through CVXPY. The ratio of absolute disagreement between the two solvers' PID components to total MI rarely exceeds 5%; no joint failure occurred on the real-world datasets.

F.2 Quantization and subsampling

The benchmark's quantization sweep uses 2 to 20 bins in steps of 2; beyond 8 bins, errors decrease only marginally, and the companion sweep over $\{2, 4, \ldots, 16\}$ bins (Appendix F.9) likewise finds non-PID errors stable above 6 bins. The subsampling sweep varies 20% to 100% in 10% increments and leaves most error curves stable (Appendix F.8). The dependence of kernel measures on the dimension of the compared vector is treated in Appendix F.7.

F.3 Computational cost and scalability

Runtime and scaling. The following runtimes come from the companion scalability study (Anonymous Authors, 2026), whose feature sets and bin counts differ from those of Table 1. With 6 bins, $v_{A,MI}$ scales to 15 features on ACS Employment at approximately 11 hours, while $v_{A,CMI}$ is feasible only up to 12 features. Gaussian measures scale to 16 features and run under 2 hours there with errors below 30% beyond 6 features. PID measures are approximately 7.8 hours at 6 bins for $v_{SP,SI:2}$ and become infeasible at higher bin counts. In the benchmark, distribution-based MI, SP-MI, and Gaussian measures use an empirical equal-frequency quantization; the Gaussian RBF bandwidth is 0.5, MMD and HSIC are subsampled below 2% estimated-measure variance, and NOCCO batches contain 512 samples.

F.4 Synthetic data with known generative structure

This subsection repeats the comparison on synthetic data whose generative structure is known, so that DAGs in which the SP and EO mechanisms differ can be identified. The reference is still the downstream NN.

Table 3: Synthetic known-SCM study: normalized attribution error (%, mean ± s.d.) against the reference NN; objective is SP and EO, and lower is better.

Measure	Distinct SP–EO DAGs		Other DAGs		Best 50% DAGs	
	SP	EO	SP	EO	SP	EO
$v_{SP,MI}$	18.0 ± 3.3	46.4 ± 5.1	29.4 ± 3.7	34.9 ± 4.8	9.7 ± 1.8	28.9 ± 5.3
$v_{SP,MMD}$	19.1 ± 3.2	44.9 ± 4.9	29.3 ± 3.6	34.2 ± 4.7	11.0 ± 1.9	28.1 ± 4.8
$v_{SP,SI:1}$	20.1 ± 3.5	50.6 ± 5.3	34.6 ± 3.9	40.8 ± 5.1	10.8 ± 1.8	32.6 ± 5.4
$v_{SP,SI:2}$	$\mathbf{17.6 \pm 3.2}$	45.0 ± 5.0	27.7 ± 3.6	32.8 ± 4.7	$\mathbf{9.2 \pm 1.8}$	27.4 ± 4.9
$v_{EO,CMI}$	19.8 ± 3.2	45.4 ± 5.0	28.9 ± 3.5	33.8 ± 4.6	11.4 ± 2.0	28.1 ± 5.3
$v_{EO,CMMD}$	19.7 ± 3.1	44.0 ± 4.9	28.2 ± 3.5	32.7 ± 4.5	11.4 ± 2.0	27.1 ± 5.1
$CI(A; X_S, Y)$	36.9 ± 2.8	43.7 ± 3.9	40.2 ± 3.0	41.7 ± 3.6	28.5 ± 2.6	32.9 ± 3.8
$CI(X_S; A, Y)$	26.4 ± 2.9	$\mathbf{35.7 \pm 4.5}$	$\mathbf{23.9 \pm 3.2}$	$\mathbf{26.4 \pm 4.1}$	13.7 ± 2.6	$\mathbf{21.5 \pm 4.3}$
$v_{D,Prod}$	18.5 ± 3.2	44.7 ± 4.9	28.2 ± 3.5	33.2 ± 4.6	10.6 ± 2.0	27.8 ± 5.2
$I(A; X_S)I(A; X_S \mid Y)$	18.6 ± 3.2	44.4 ± 4.9	28.2 ± 3.5	33.1 ± 4.6	10.7 ± 2.0	27.7 ± 5.2

The synthetic SCM has $A \sim \mathrm{Bern}(p_a)$, roots $Z_1, Z_2 \sim \mathrm{Uniform}([0, 2])$, four features, and a binary sink Y. Its structural equations are

$$X_i = \alpha_{a,i} w_{a,i} A + \sum_{k \in [4] \setminus \{i\}} \alpha_{k,i} w_{k,i} X_k + \sum_{j \in [2]} \alpha_{z_j,i} w_{z_j,i} Z_j + \varepsilon_i, \quad i \in [4], \tag{4}$$

$$\widetilde{Y} = \sqrt{\eta} \frac{Y_{\mathrm{sig}} - \mathbb{E}[Y_{\mathrm{sig}}]}{\sqrt{\mathrm{Var}[Y_{\mathrm{sig}}]}} + \sqrt{1 - \eta} \varepsilon_y,$$

$$Y_{\mathrm{sig}} = \alpha_{a,y} w_{a,y} A + \sum_{k \in [4]} \alpha_{k,y} w_{k,y} X_k + \sum_{j \in [2]} \alpha_{z_j,y} w_{z_j,y} Z_j, \qquad Y = \mathbf{1}(\widetilde{Y} > th_y). \tag{5}$$

Here $\alpha \in \{0, 1\}$ indicates whether an edge is present (with $\alpha_{k,i} = 0$ unless X_k precedes X_i in the sampled DAG), $w \in [1, 2]$ is its weight, $\eta \in [0, 1]$ sets the share of label variance explained by the signal, th_y is the label threshold, and $\varepsilon_i, \varepsilon_y \sim N(0, 1)$. The experiment samples 7,200 DAGs by varying the graph density $p_d^{a,z,x} \in \{0.2, 0.4, 0.6, 0.8\}$, the feature-to-label density $p_d^{x,y} \in \{0.5, 0.7, 0.9\}$, the direct-effect probability $p_d^{a,y} \in \{0.2, 0.8\}$, the number of confounders $N_c \in \{0, 1, 2\}$, the demographic ratio $p_a \in \{0.1, 0.6\}$, and $th_y \in \{0.2, 0.3, 0.4, 0.5, 0.6\}$. For each of 720 parameter combinations it samples 10 DAGs, with $w \sim \mathrm{Uniform}[1, 2]$ ($w_{a,y} = 2$), $\eta \sim \mathrm{Uniform}([0.5, 1])$, and ten 40k-example datasets per DAG. The reference is a 3-unit single-hidden-layer NN trained on a 67/33 split. The study complements the real-data benchmark rather than standing alone.

On distinct SP–EO DAGs, $v_{SP,SI:2}$ has the lowest SP error, 17.6%, while $CI(X_S; A, Y)$ has the lowest EO error, 35.7%; when the two mechanisms differ, different certificates lead on SP and on EO. $CI(\cdot; \cdot, \cdot)$ is the synergistic information of the same decomposition, with the first argument as the target; $CI(A; X_S, Y)$ is the faulty EO measure of Appendix E, included as a negative control.

F.5 Graph hyperparameters

This subsection reports how the synthetic errors vary with the DAG-generation parameters. Downstream accuracy ranges from 86.28% to 98.97% across 95% of DAGs. Increasing graph density $p_d^{a,z,x}$ raises nearly all SP and EO errors except the faulty EO negative control and $CI(X_S; A, Y)$; varying N_c has only minor effects except at the largest density. As $p_d^{x,y}$ increases, SP errors decrease slightly except for the faulty EO negative control; EO errors show no noticeable change. For SP, let R_A and R_X be the sums over coalitions of $UI(A; X_S \backslash Y)/I(A; X_S)$ and $UI(X_S; A \backslash Y)/I(A; X_S)$, the shares of $I(A; X_S)$ held by the two unique-information terms. The correlation of R_A with the SP error of $v_{SP,MI}$ (which contains $UI(A; X_S \setminus Y)$) and of $v_{SP,SI:1}$ (which excludes it) is 0.117 and 0.130; that of R_X with the SP error of $v_{SP,MI}$ and of $v_{SP,SI:2}$ is -0.087 and -0.111. The

Table 4: Synthetic prediction-aware study: attribution error (%, mean ± s.d.) against the reference NN for SP and EO; lower is better.

Measure	Distinct SP–EO DAGs		Other DAGs		Best 50% DAGs	
	SP	EO	SP	EO	SP	EO
$v^*_{SP,MI}$	**17.0 ± 3.1**	43.4 ± 4.9	26.7 ± 3.5	31.6 ± 4.6	8.9 ± 1.7	26.2 ± 4.8
$v^*_{SP,MMD}$	18.0 ± 3.3	47.1 ± 5.1	30.0 ± 3.7	36.0 ± 4.9	9.1 ± 1.7	29.5 ± 5.1
$v^*_{SP,SI:1}$	17.4 ± 3.2	45.6 ± 5.0	30.0 ± 3.7	35.7 ± 4.8	**8.7 ± 1.7**	28.4 ± 5.0
$v^*_{SP,SI:2}$	17.2 ± 3.2	44.6 ± 4.9	27.5 ± 3.6	32.9 ± 4.7	8.9 ± 1.8	27.5 ± 4.9
$v^*_{EO,CMI}$	24.1 ± 2.8	32.6 ± 4.2	24.2 ± 3.1	25.6 ± 3.9	14.0 ± 2.4	20.0 ± 3.8
$v^*_{EO,CMMD}$	22.2 ± 2.9	35.4 ± 4.5	24.0 ± 3.2	26.4 ± 4.2	12.3 ± 2.2	21.0 ± 4.2
$CI^*(A; X_S, Y)$	27.2 ± 2.9	35.1 ± 4.1	30.5 ± 3.1	32.4 ± 3.9	16.7 ± 2.4	23.6 ± 3.7
$CI^*(X_S; A, Y)$	24.5 ± 2.9	35.9 ± 4.5	24.5 ± 3.2	27.2 ± 4.1	12.9 ± 2.5	21.4 ± 4.2
$v^*_{D,Prod}$	23.2 ± 2.8	34.8 ± 4.1	26.8 ± 3.1	28.2 ± 3.8	13.8 ± 2.2	22.4 ± 3.5
$I^*(A; X_S)I^*(A; X_S \mid Y)$	22.4 ± 2.8	**32.5 ± 4.2**	**23.5 ± 3.1**	**24.5 ± 3.8**	13.3 ± 2.4	**19.8 ± 3.7**

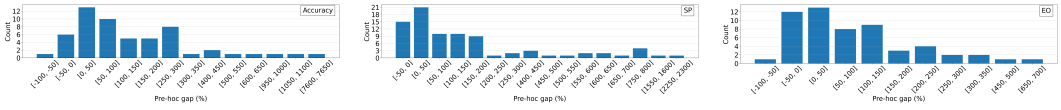

Figure 3: Distribution of the pre-hoc attribution gap (main paper, Eq. 3) across certificates and datasets, for accuracy (left), SP (centre), and EO (right). Positive values are the fidelity the data auditor forfeits by not holding model access; negative values are configurations where the model-agnostic certificate is more faithful. Per-certificate values are in Tables 8–10; the extreme positive tail comes from DC certificates.

EO counterparts, with $I(A; X_S \mid Y)$ in the denominator, are -0.0194, 0.0843, 0.122, and 0.104. Rankings are therefore stable across regimes, while absolute errors are not.

F.6 Measure error with access to predictions

This subsection repeats the synthetic comparison with X_S replaced by $\hat{Y}_{|S}$, the prediction-aware counterparts of Section 4. All SP-aligned measures improve by up to 19.4% relative to Table 3, except $v_{SP,MMD}$ on other DAGs (+2.4%); all EO-aligned measures improve by up to 28.8%, except $CI(X_S; A, Y)$ on distinct SP–EO and other DAGs (+0.6% and +3.0%); the two product measures improve on EO in every group but worsen on SP on distinct SP–EO and best 50% DAGs. Per-measure values are given below, with prediction-aware counterparts starred as in the main paper.

F.7 MMD dependence on dimension

This subsection gives a closed-form instance of the dimension effect discussed in Section 5.1, for a kernel measure where it can be computed exactly. For Gaussian $P = \mathcal{N}(\mu_1, \Sigma)$ and $Q = \mathcal{N}(\mu_2, \Sigma)$ on $\mathbb{R}^d$ and the kernel $k(x, x') = \exp(-\|x - x'\|^2/(2\gamma^2))$,

$$\mathrm{MMD}^2(P, Q) = 2\left(\frac{\gamma^2}{2}\right)^{d/2} \frac{1}{|\Sigma + \gamma^2 I/2|^{1/2}} \times \left[1 - \exp\!\left(-\Delta^\top (\Sigma + \gamma^2 I/2)^{-1} \Delta/4\right)\right],$$

where $\Delta = \mu_1 - \mu_2$. For $\Sigma = I$, $\mu_1 = [1, 0, \ldots, 0]$, and $\mu_2 = [0, \ldots, 0]$, this reduces to $\mathrm{MMD}^2(P, Q) = 2(\gamma^2/(\gamma^2 + 2))^{d/2}(1 - \exp(-1/(2(\gamma^2 + 2))))$, which decreases with d although the added coordinates carry no signal. Changes in an MMD attribution across coalitions of different sizes can therefore reflect $|S|$ rather than added dependence.

F.8 Effect of subsampling

Sweeping subsampling from 20% to 100% in 10% increments leaves most measure-error curves stable.

Table 5: Prediction-agnostic MI/PID certificates: attribution error (%, mean ± s.d. across 3 seeds) against the NN reference predictor, with average runtime in parentheses; lower is better. Objective is accuracy (top), SP (middle), and EO (bottom).

Measure (Avg. runtime)	ACS Income	ACS Pub. Cov.	ACS Emp.	COMPAS	Income (KDD)	Adult	Health
$v_{A,MI}$ (1.5s)	**4.9 ± 0.3**	**10.4 ± 0.5**	1.2 ± 0.1	**6.3 ± 1.1**	**9.3 ± 1.7**	**8.4 ± 0.6**	**10.5 ± 3.3**
$v_{A,CMI}$ (17.0s)	5.2 ± 0.3	11.3 ± 0.5	**1.0 ± 0.1**	8.0 ± 1.1	9.6 ± 1.9	8.5 ± 0.6	11.5 ± 3.2
$v_{SP,MI}$ (1.5s)	12.4 ± 1.0	9.7 ± 3.0	4.1 ± 1.4	14.4 ± 3.4	23.2 ± 1.2	6.1 ± 2.2	17.0 ± 8.5
$v_{SP,SI:1}$ (1.3m)	12.8 ± 1.0	**8.9 ± 3.0**	4.2 ± 1.4	16.4 ± 3.8	20.0 ± 1.0	9.5 ± 2.0	17.6 ± 6.5
$v_{SP,SI:2}$ (1.4h)	**5.9 ± 1.4**	22.7 ± 2.6	**3.3 ± 1.6**	**7.1 ± 1.8**	**18.7 ± 0.8**	6.7 ± 1.6	**14.1 ± 7.7**
$v_{D,Prod}$ (1.4m)	15.8 ± 1.9	10.0 ± 2.3	3.9 ± 1.2	16.8 ± 3.8	23.5 ± 1.2	**5.9 ± 2.3**	21.1 ± 9.4
$v_{EO,CMI}$ (3.5s)	24.3 ± 3.1	12.7 ± 2.4	7.0 ± 2.5	**17.1 ± 4.6**	**30.6 ± 2.1**	6.2 ± 3.5	**19.6 ± 6.1**
$v_{D,Prod}$ (1.4m)	**21.2 ± 2.8**	**11.6 ± 2.4**	**5.5 ± 2.8**	17.2 ± 4.6	31.8 ± 2.3	**6.1 ± 3.6**	23.1 ± 5.9

Table 6: Gaussian-surrogate certificates: attribution error (%, mean ± s.d. across 3 seeds) against the NN reference predictor, followed by relative error change (%) in parentheses and average runtime; lower error is better. Objective is accuracy (top), SP (middle), and EO (bottom).

Measure (Avg. runtime)	ACS Income	ACS Pub. Cov.	ACS Emp.	COMPAS	Income (KDD)	Adult	Health
v_{A,MI_g} (840ms)	**7.2 ± 0.1** (46)	**11.8 ± 0.4** (14)	**1.8 ± 0.2** (46)	3.9 ± 1.3 (**-39**)	16.2 ± 1.4 (74)	15.8 ± 0.7 (88)	**13.4 ± 5.6** (27)
v_{A,CMI_g} (2.5s)	7.4 ± 0.1 (41)	12.3 ± 0.3 (9)	2.3 ± 0.2 (132)	**3.6 ± 1.3** (**-55**)	**15.9 ± 1.4** (66)	**13.9 ± 0.7** (64)	13.5 ± 5.5 (17)
v_{SP,MI_g} (877ms)	**13.1 ± 2.3** (6)	21.5 ± 1.2 (121)	2.4 ± 1.5 (**-43**)	8.9 ± 3.0 (**-38**)	29.3 ± 0.4 (26)	**6.2 ± 1.5** (2)	30.7 ± 6.2 (81)
$v_{SP,SI_g:1}$ (1.8s)	**13.1 ± 2.3** (3)	21.5 ± 1.2 (140)	**2.3 ± 1.5** (**-45**)	17.0 ± 3.8 (4)	29.3 ± 0.4 (47)	11.9 ± 1.6 (25)	30.7 ± 6.2 (75)
$v_{SP,SI_g:2}$ (1.6s)	**13.1 ± 2.3** (121)	**18.9 ± 0.8** (**-17**)	2.4 ± 1.5 (**-29**)	**8.6 ± 3.0** (21)	29.3 ± 0.4 (57)	8.9 ± 2.9 (32)	**28.1 ± 6.3** (99)
$v_{D,Prod_g}$ (4.8s)	16.4 ± 2.3 (4)	20.9 ± 2.1 (110)	3.7 ± 1.5 (**-6**)	**8.6 ± 2.9** (**-49**)	**22.9 ± 0.4** (**-2**)	6.4 ± 1.7 (10)	30.0 ± 6.7 (42)
v_{EO,CMI_g} (1.9s)	19.7 ± 3.8 (**-19**)	**18.4 ± 3.2** (45)	**4.5 ± 2.7** (**-36**)	10.5 ± 4.1 (**-39**)	**28.5 ± 1.9** (**-7**)	**6.1 ± 3.2** (**-1**)	35.8 ± 8.9 (83)
$v_{D,Prod_g}$ (4.8s)	**15.8 ± 4.0** (**-25**)	20.5 ± 3.1 (77)	5.5 ± 2.8 (1)	**9.4 ± 4.2** (**-45**)	31.5 ± 1.8 (**-1**)	6.4 ± 3.1 (5)	**30.1 ± 8.5** (30)

F.9 Scalability with bin counts

The underlying cost–fidelity sweep is reported in full by the companion submission (Anonymous Authors, 2026). Across bins $\{2, 4, \ldots, 16\}$ under a 12-hour cap, non-PID errors change only slightly above 6 bins; at 16 bins $v_{A,MI}$ and $v_{SP,MI}$ stay below 200 seconds, $v_{A,CMI}$ is over $10\times$ slower, $v_{EO,CMI}$ is $2\times$ slower, and $v_{SP,SI:2}$ is approximately 7.8 hours at 6 bins.

F.10 Scalability with feature dimension

Drawing on the companion submission (Anonymous Authors, 2026): with 6 bins, $v_{A,MI}$ reaches 15 features at approximately 11 hours, $v_{A,CMI}$ only 12, $v_{SP,SI:2}$ only 6, and $v_{EO,CMI}$ 14. Gaussian measures reach 16 features under 2 hours with errors below 30% beyond 6 features; kernel measures are feasible only to 11 features even without conditioning.

F.11 Fairness-aware feature selection

This subsection illustrates one use of the certificates, fairness-aware feature selection. The selection score is $\text{score}_i = \phi_i(v_{A,CMI}) - \alpha\phi_i(v_D)$, where v_D is the SP- or EO-aligned certificate under evaluation and $\alpha = 1$, on ACS Income and ACS Public Coverage. Three of four SP-aligned certificates and all EO-aligned certificates recover the correct feature ranking under this score.

G Extended Benchmark Tables

This section reports the extended benchmark results behind Section 5, evaluated against the NN reference predictor; all entries are attribution error (%) and lower is better.

Table 7: Alternative-dependence-metric certificates: attribution error (%, mean ± s.d. across 3 seeds) against the NN reference predictor, followed by relative error change (%) in parentheses and average runtime; lower error is better. Objective is accuracy (top), SP (middle), and EO (bottom).

Measure (Avg. runtime)	ACS Income	ACS Pub. Cov.	ACS Emp.	COMPAS	Income (KDD)	Adult	Health
$v_{A,DC}$ (1.0s)	41.6 ± 0.3 (744)	45.9 ± 1.1 (341)	42.3 ± 0.4 (3331)	34.5 ± 0.5 (445)	38.7 ± 1.4 (318)	42.2 ± 0.6 (401)	44.0 ± 4.2 (318)
$v_{A,MM}$ (31.3m)	**6.7 ± 0.1** (37)	20.1 ± 0.7 (93)	**1.6 ± 0.2** (33)	**3.4 ± 1.3 (-46)**	23.8 ± 1.0 (156)	**12.8 ± 0.7** (51)	15.6 ± 5.2 (48)
$v_{A,HS}$ (12.1m)	18.5 ± 0.1 (275)	**10.7 ± 0.7** (2)	33.3 ± 0.3 (2602)	27.7 ± 0.4 (338)	35.0 ± 0.9 (277)	42.3 ± 0.5 (401)	27.0 ± 1.5 (156)
$v_{A,NO}$ (46.4m)	10.8 ± 0.3 (119)	13.6 ± 1.1 (30)	4.7 ± 0.3 (282)	10.4 ± 1.1 (64)	**12.4 ± 1.4** (34)	16.8 ± 0.6 (99)	**15.4 ± 2.7** (46)
$v_{SP,DC}$ (1.0s)	40.1 ± 1.9 (223)	54.3 ± 2.6 (457)	43.0 ± 2.0 (942)	36.7 ± 3.8 (154)	43.5 ± 0.2 (87)	54.0 ± 3.8 (787)	45.8 ± 5.9 (170)
$v_{SP,MMD}$ (37.8m)	**8.1 ± 2.3 (-35)**	15.9 ± 1.8 (64)	**2.6 ± 1.6 (-37)**	**10.7 ± 3.1 (-26)**	27.1 ± 1.0 (17)	7.1 ± 2.3 (17)	34.5 ± 5.9 (103)
$v_{SP,HS}$ (11.6m)	23.4 ± 0.6 (89)	17.4 ± 1.0 (79)	33.4 ± 1.9 (708)	32.9 ± 3.2 (128)	32.8 ± 1.0 (41)	37.5 ± 3.3 (516)	26.8 ± 4.0 (58)
$v_{SP,NO}$ (47.3m)	18.8 ± 1.9 (52)	**11.6 ± 2.2** (19)	9.5 ± 1.2 (131)	15.8 ± 3.7 (10)	**24.1 ± 0.4** (4)	**6.0 ± 2.4 (-1)**	**18.5 ± 7.1** (9)
$v_{EO,CDC}$ (3.9s)	49.7 ± 2.9 (105)	48.0 ± 3.2 (278)	38.9 ± 3.9 (454)	40.1 ± 5.2 (134)	**28.6 ± 1.0 (-7)**	35.1 ± 3.4 (469)	43.7 ± 2.1 (123)
$v_{EO,CMMD}$ (43.8m)	27.8 ± 2.9 (14)	**12.0 ± 2.0 (-5)**	**4.6 ± 2.8 (-34)**	**11.6 ± 3.5 (-32)**	30.1 ± 2.3 (**-2**)	7.4 ± 2.9 (20)	36.1 ± 8.7 (84)
$v_{EO,CHS}$ (17.0m)	28.8 ± 3.3 (18)	19.4 ± 1.9 (53)	32.0 ± 3.8 (356)	34.2 ± 3.7 (99)	36.3 ± 2.0 (19)	37.1 ± 3.4 (502)	29.0 ± 2.5 (48)
$v_{EO,CNO}$ (44.4m)	**23.9 ± 2.6 (-2)**	13.7 ± 2.5 (8)	12.8 ± 2.1 (82)	18.9 ± 5.0 (10)	34.4 ± 3.2 (12)	**7.3 ± 3.6** (18)	**21.5 ± 6.5** (10)

Table 8: Prediction-aware accuracy certificates: attribution error (%, mean ± s.d. across 3 seeds) against the NN reference predictor; pre-hoc gap (%) is in parentheses and average runtime follows the measure. Lower error is better.

Measure (Avg. runtime)	ACS Income	ACS Pub. Cov.	ACS Emp.	COMPAS	Income (KDD)	Adult	Health
$v^{\star}_{A,MI}$ (1.1m)	2.7 ± 0.1 (81.2)	2.7 ± 0.9 (288.4)	1.3 ± 0.1 (**-2.8**)	3.3 ± 1.2 (91.5)	7.7 ± 1.7 (20.3)	4.9 ± 1.3 (73.6)	6.8 ± 4.8 (54.0)
$v^{\star}_{A,CMI}$ (1.5m)	2.5 ± 0.0 (109.4)	3.0 ± 1.0 (279.5)	0.7 ± 0.1 (34.2)	6.1 ± 0.5 (30.7)	6.5 ± 1.6 (47.5)	5.7 ± 0.9 (49.4)	8.3 ± 4.7 (38.7)
$v^{\star}_{A,MI_g}$ (12.3s)	2.6 ± 0.1 (173.4)	2.7 ± 1.0 (336.1)	1.2 ± 0.1 (55.7)	**3.2 ± 1.2** (20.0)	6.9 ± 1.6 (135.1)	**4.4 ± 1.2** (257.7)	6.8 ± 4.8 (96.7)
$v^{\star}_{A,CMI_g}$ (15.4s)	**2.1 ± 0.1** (250.3)	5.1 ± 0.8 (143.3)	0.8 ± 0.1 (187.3)	3.4 ± 0.9 (5.5)	**4.5 ± 1.0** (251.7)	5.4 ± 1.0 (158.3)	6.6 ± 4.6 (105.7)
$v^{\star}_{A,DC}$ (1.3m)	3.5 ± 0.2 (1076.1)	6.4 ± 1.9 (612.9)	**0.5 ± 0.3** (7625.9)	3.3 ± 1.1 (952.8)	26.6 ± 10.8 (45.6)	23.4 ± 0.6 (80.6)	8.3 ± 4.2 (432.0)
$v^{\star}_{A,MM}$ (2.0h)	4.7 ± 0.3 (44.8)	3.2 ± 2.0 (532.1)	1.6 ± 0.2 (3.7)	6.3 ± 1.3 (**-45.0**)	9.0 ± 1.8 (164.0)	6.7 ± 1.2 (90.8)	8.2 ± 5.1 (91.5)
$v^{\star}_{A,HS}$ (39.9m)	30.9 ± 0.2 (**-40.2**)	30.9 ± 0.7 (**-65.6**)	51.1 ± 1.3 (**-34.8**)	35.4 ± 1.1 (**-21.8**)	31.9 ± 1.5 (9.5)	31.6 ± 0.8 (33.6)	33.0 ± 6.3 (**-18.2**)
$v^{\star}_{A,NO}$ (1.3h)	2.8 ± 0.1 (282.8)	**2.5 ± 0.8** (437.9)	1.2 ± 0.1 (282.2)	3.6 ± 1.3 (190.7)	7.0 ± 1.6 (76.0)	4.5 ± 1.2 (276.6)	**6.5 ± 5.1** (137.9)

G.1 Robustness across downstream predictors

The information-theoretic certificates are evaluated against six separately tuned downstream predictors on the three largest datasets. These are the per-predictor errors summarized in Table 2 (Section 5); they show whether certificate rankings and errors persist when the reference predictor changes. Each cell is attribution error (%, mean ± s.d.); the final column is the average error across predictors.

G.2 Certificates instantiated by role

A *data auditor* should compute $v_{A,MI}$ for accuracy (best on 6/7 datasets, cheapest, most stable across model families) and $v_{SP,MI}$ for SP: $v_{SP,SI:2}$ is more faithful but is the only certificate costing hours and, like every PID or kernel certificate, carries no Proposition 1 guarantee. For EO, $v_{D,Prod}$ leads on fidelity and $v_{EO,CMI}$ is the guaranteed, far cheaper alternative; among distributional substitutes MMD/CMMD is best and NOCCO safest. By (P1), an *alignment auditor* holding model access should not reuse these instruments.

G.3 Why privacy and safety push auditing upstream

Model-agnostic auditing is better understood as what privacy and data-governance constraints force than as a convenience. Sensitive attributes carry the strictest collection, retention, and sharing limits, so the party holding A is frequently not the party holding h. Under data minimization, federation, or siloing (McMahan et al., 2017), the joint distribution over $(X_{[m]}, A, Y, \hat{Y})$ that post-hoc auditing presumes may not be assemblable by any single party, and differential privacy adds noise to precisely the group-conditional statistics fairness auditing consumes (Dwork & Roth, 2014). Stage-local certificates are compatible with this regime by construction: each is a functional of one stage's marginals plus the supervision signal, computable inside a silo, and Proposition 1 keeps the bound valid for a model trained elsewhere by someone else. We claim no more than compatibility: we add no DP noise and evaluate no federated deployment (Appendix J). The safety reading is parallel: an audit runnable only on a deployed model arrives after the harm and can be influenced by the audited

Table 9: Prediction-aware SP certificates: attribution error (%, mean ± s.d. across 3 seeds) against the NN reference predictor; pre-hoc gap (%) is in parentheses and average runtime follows the measure. Lower error is better.

Measure (Avg. runtime)	ACS Income	ACS Pub. Cov.	ACS Emp.	COMPAS	Income (KDD)	Adult	Health
$v^*_{SP,MI}$ (1.3m)	**5.2 ± 1.8** (138.1)	8.7 ± 2.6 (12.4)	3.7 ± 0.9 (11.3)	4.9 ± 3.0 (192.1)	3.5 ± 0.7 (563.0)	4.2 ± 1.4 (45.9)	19.2 ± 4.9 (**-11.7**)
$v^*_{SP,SI:1}$ (1.2m)	**5.2 ± 1.8** (145.4)	7.5 ± 2.3 (18.6)	3.3 ± 0.8 (25.1)	11.4 ± 5.1 (43.4)	3.4 ± 0.7 (479.8)	3.5 ± 1.7 (171.1)	13.6 ± 5.5 (28.8)
$v^*_{SP,SI:2}$ (1.2m)	**5.2 ± 1.8** (13.7)	8.7 ± 2.6 (162.8)	3.3 ± 0.8 (2.2)	5.0 ± 2.7 (41.7)	3.5 ± 0.7 (434.8)	**3.3 ± 1.1** (105.2)	14.4 ± 5.2 (**-1.7**)
$v^*_{D,Prod}$ (1.2m)	7.9 ± 2.9 (99.1)	6.0 ± 1.0 (66.2)	3.8 ± 1.6 (3.7)	**3.3 ± 1.4** (409.5)	4.5 ± 1.2 (421.0)	5.6 ± 3.1 (4.3)	19.5 ± 3.7 (8.0)
v^*_{SP,MI_g} (12.6s)	**5.2 ± 1.8** (151.6)	8.8 ± 2.6 (144.8)	4.5 ± 1.0 (**-47.5**)	5.0 ± 3.0 (77.2)	3.4 ± 0.8 (759.7)	3.9 ± 1.5 (60.9)	21.6 ± 4.4 (42.1)
$v^*_{SP,SI_g:1}$ (15.1s)	**5.2 ± 1.8** (151.6)	8.8 ± 2.6 (144.8)	4.5 ± 1.0 (**-48.8**)	13.0 ± 6.1 (30.7)	3.4 ± 0.8 (759.7)	3.7 ± 2.1 (227.1)	21.6 ± 4.4 (42.1)
$v^*_{SP,SI_g:2}$ (14.6s)	**5.2 ± 1.8** (151.6)	8.8 ± 2.6 (115.4)	4.5 ± 1.0 (**-47.2**)	5.0 ± 2.6 (71.3)	3.4 ± 0.8 (760.1)	3.5 ± 0.9 (154.3)	15.5 ± 5.3 (80.5)
$v^*_{D,Prod_g}$ (18.8s)	8.6 ± 2.9 (90.9)	9.0 ± 1.4 (132.1)	6.1 ± 2.5 (**-39.6**)	3.5 ± 1.5 (142.6)	5.8 ± 1.3 (292.8)	8.1 ± 2.5 (**-20.1**)	24.2 ± 4.0 (23.9)
$v^*_{SP,DC}$ (1.2m)	**5.2 ± 1.7** (673.8)	**3.2 ± 1.8** (1584.6)	**1.8 ± 0.7** (2254.1)	4.9 ± 2.8 (643.6)	25.2 ± 13.2 (72.8)	8.1 ± 10.0 (569.4)	**6.2 ± 3.0** (638.2)
$v^*_{SP,MMD}$ (2.2h)	6.5 ± 1.2 (24.9)	7.2 ± 3.0 (122.6)	3.6 ± 1.2 (**-27.9**)	8.2 ± 3.7 (31.5)	**3.2 ± 1.3** (757.4)	3.7 ± 0.9 (95.4)	8.9 ± 4.4 (287.3)
$v^*_{SP,HS}$ (40.8m)	33.1 ± 2.4 (**-29.2**)	25.8 ± 2.6 (**-32.4**)	50.8 ± 1.7 (**-34.2**)	35.3 ± 3.7 (**-6.9**)	43.3 ± 0.5 (**-24.3**)	33.8 ± 4.0 (10.7)	35.2 ± 11.4 (**-23.9**)
$v^*_{SP,NO}$ (1.3h)	6.3 ± 0.8 (197.2)	10.2 ± 2.1 (13.3)	4.8 ± 1.1 (100.6)	5.6 ± 2.6 (180.4)	3.7 ± 0.8 (545.9)	3.7 ± 1.5 (61.4)	20.6 ± 3.4 (**-9.9**)

Table 10: Prediction-aware EO certificates: attribution error (%, mean ± s.d. across 3 seeds) against the NN reference predictor; pre-hoc gap (%) is in parentheses and average runtime follows the measure. Lower error is better.

Measure (Avg. runtime)	ACS Income	ACS Pub. Cov.	ACS Emp.	COMPAS	Income (KDD)	Adult	Health
$v^*_{EO,CMI}$ (1.2m)	10.5 ± 3.0 (130.8)	8.6 ± 3.3 (47.9)	11.0 ± 4.5 (**-36.0**)	6.6 ± 4.0 (158.2)	**9.9 ± 2.4** (210.4)	5.8 ± 2.8 (7.0)	20.6 ± 2.0 (**-4.9**)
$v^*_{D,Prod}$ (1.2m)	**9.0 ± 3.4** (137.0)	**5.3 ± 1.5** (116.4)	**4.0 ± 1.1** (37.0)	**4.5 ± 2.8** (285.0)	**9.9 ± 2.3** (220.2)	**4.5 ± 1.8** (34.5)	15.4 ± 5.5 (50.3)
v^*_{EO,CMI_g} (14.1s)	9.9 ± 3.0 (99.3)	9.1 ± 3.5 (102.8)	15.8 ± 6.2 (**-71.8**)	6.6 ± 4.0 (58.6)	10.6 ± 1.9 (168.9)	6.0 ± 2.7 (2.7)	22.7 ± 3.0 (58.1)
$v^*_{D,Prod_g}$ (18.8s)	9.7 ± 3.5 (63.4)	8.5 ± 2.8 (141.5)	5.9 ± 1.9 (**-5.8**)	4.7 ± 2.7 (99.7)	10.1 ± 1.9 (210.4)	6.9 ± 1.8 (**-7.2**)	20.7 ± 4.5 (45.0)
$v^*_{EO,CDC}$ (1.2m)	11.1 ± 3.2 (349.1)	6.2 ± 2.5 (668.9)	11.8 ± 6.6 (229.9)	7.1 ± 4.3 (468.1)	30.0 ± 7.6 (**-4.5**)	15.2 ± 11.5 (130.5)	**10.2 ± 2.7** (327.2)
$v^*_{EO,CMMD}$ (2.7h)	13.5 ± 3.3 (106.3)	5.9 ± 2.4 (102.7)	5.2 ± 2.0 (**-11.4**)	9.6 ± 5.2 (21.2)	11.3 ± 2.5 (165.0)	**4.5 ± 2.7** (64.9)	10.3 ± 1.3 (251.1)
$v^*_{EO,CHS}$ (56.8m)	29.7 ± 3.5 (**-3.0**)	26.1 ± 3.5 (**-25.7**)	47.6 ± 3.2 (**-32.7**)	34.3 ± 4.1 (**-0.3**)	53.0 ± 2.5 (**-31.5**)	36.4 ± 4.5 (2.1)	36.1 ± 8.9 (**-19.7**)
$v^*_{EO,CNO}$ (1.2h)	18.5 ± 4.2 (29.4)	10.4 ± 2.6 (31.8)	11.4 ± 4.7 (12.3)	10.7 ± 5.5 (77.0)	14.3 ± 3.8 (141.3)	5.8 ± 3.3 (24.8)	21.4 ± 1.2 (0.4)

party, whereas a stage-local certificate is computable before training, is not a function of the model, and under (P2) errs conservatively.

H Cadence Study Details

This section gives the details of the cadence study of Section 5.4: exact costs, sampling, the declared causal constraints, and the monotonicity check.

H.1 Sampling-based Shapley approximation

The protocol follows the companion submission (Anonymous Authors, 2026). The study uses ACS Coverage with 20 input features: the 18 defined for the dataset plus GCL and COW. It evaluates v_{A,MI_g}, v_{SP,MI_g}, and v_{EO,CMI_g}. Exact Shapley values require 71 hours for v_{A,MI_g} and v_{SP,MI_g} and 150 hours for v_{EO,CMI_g}; Monte Carlo sampling converges below 1% approximation error within 30–70 minutes. Monte Carlo sampling has the lowest error across the full horizon, outperforming a multi-armed-bandit allocation, with accuracy, SP, and EO attribution error below 17%, 15%, and 16% after 30 minutes.

H.2 Causal constraints and the sliding-window check

The declared ancestor–descendant pairs are C1: (AGEP, SCHL), C2: (AGEP, PINCP), C3: (SCHL, COW), C4: (SCHL, PINCP), C5: (ESR, PINCP), and C6: (MIG, PINCP). The nested sets are CK1= {C1}, CK2= {C1,C2}, CK3= {C1,C2,C3}, CK4= {C1–C5}, and CK5= {C1–C6}. At 6000 seconds, CK3 has the lowest error for v_{A,MI_g} and v_{SP,MI_g}; CK1 or CK2 has the lowest error for v_{EO,CMI_g}; CK5 is worst for the accuracy and SP measures. For the monotonicity check, all admissible pairs are ranked by the error each induces alone, and groups of four are formed with a sliding window over this ranking: groups built from individually strong pairs give the lowest error, and the error increases monotonically as their constituent pairs weaken. A third party holding the declared constraint set and the per-pair errors can therefore check the issuer's constraints for consistency, which makes the discretion in (P5) reviewable rather than merely disclosed.

Table 11: Robustness across six downstream predictors on ACS Income.

Measure	CatBoost	LGBM	XGB	RandomForest	Logistic	HistGradientBoosting	Avg. error
$v_{A,MI}$	4.8 ± 0.1	8.4 ± 0.1	7.5 ± 0.4	6.1 ± 0.4	7.9 ± 0.2	4.9 ± 0.1	6.6 ± 1.4
$v_{A,CMI}$	5.1 ± 0.1	8.7 ± 0.1	7.7 ± 0.4	6.3 ± 0.4	8.2 ± 0.2	5.2 ± 0.1	6.9 ± 1.4
$v_{SP,MI}$	12.8 ± 0.3	20.1 ± 0.7	19.6 ± 1.0	14.4 ± 0.7	18.8 ± 1.8	14.3 ± 0.7	16.7 ± 2.9
$v_{SP,SI:1}$	13.6 ± 0.3	21.2 ± 0.7	20.3 ± 0.8	15.6 ± 0.7	16.9 ± 1.6	15.4 ± 0.6	17.2 ± 2.7
$v_{SP,SI:2}$	6.9 ± 0.8	9.8 ± 1.3	13.3 ± 1.5	6.0 ± 1.7	13.0 ± 2.3	7.8 ± 0.8	9.5 ± 2.8
$v_{D,Prod}$	14.6 ± 0.5	21.9 ± 0.7	21.1 ± 0.8	16.7 ± 0.6	21.5 ± 1.5	16.2 ± 0.6	18.7 ± 2.9
$v_{EO,CMI}$	23.6 ± 1.0	29.1 ± 0.9	29.4 ± 2.0	25.8 ± 1.8	41.7 ± 7.5	28.1 ± 1.6	29.6 ± 5.8
$v_{D,Prod}$	20.1 ± 0.9	28.0 ± 2.5	28.1 ± 2.3	22.6 ± 1.8	43.5 ± 6.8	24.8 ± 1.6	27.9 ± 7.5

Table 12: Robustness across six downstream predictors on ACS Public Coverage.

Measure	CatBoost	LGBM	XGB	RandomForest	Logistic	HistGradientBoosting	Avg. error
$v_{A,MI}$	10.3 ± 0.4	15.2 ± 0.5	15.6 ± 1.0	11.7 ± 1.3	17.4 ± 0.1	10.1 ± 0.5	13.4 ± 2.8
$v_{A,CMI}$	11.1 ± 0.4	16.0 ± 0.5	16.5 ± 1.0	12.6 ± 1.3	18.2 ± 0.1	11.0 ± 0.4	14.2 ± 2.8
$v_{SP,MI}$	9.3 ± 1.3	18.0 ± 2.1	18.3 ± 3.1	9.1 ± 1.6	22.6 ± 0.4	9.4 ± 1.2	14.5 ± 5.4
$v_{SP,SI:1}$	8.5 ± 1.3	17.2 ± 2.1	17.5 ± 3.1	10.9 ± 1.6	21.8 ± 0.4	8.6 ± 1.3	14.1 ± 5.0
$v_{SP,SI:2}$	21.9 ± 1.4	19.9 ± 2.1	19.6 ± 3.1	25.0 ± 3.1	15.3 ± 0.4	22.3 ± 0.9	20.7 ± 3.0
$v_{D,Prod}$	10.4 ± 1.6	17.6 ± 2.1	17.9 ± 3.1	11.3 ± 1.5	22.2 ± 0.4	9.9 ± 0.7	14.9 ± 4.6
$v_{EO,CMI}$	13.5 ± 1.4	17.9 ± 2.5	18.3 ± 3.8	15.1 ± 1.8	23.2 ± 0.6	13.0 ± 1.0	16.8 ± 3.5
$v_{D,Prod}$	12.6 ± 1.4	16.2 ± 2.6	16.7 ± 3.8	13.5 ± 1.7	21.6 ± 0.6	11.9 ± 1.1	15.4 ± 3.3

I Two-Stage Study: Design, Hypotheses, and Per-State Results

This section specifies the two-stage study of Section 5.5 and reports every state.

Data and provenance. We use the 2018 ACS one-year person records for the 50 states and Puerto Rico, as distributed through folktables (Ding et al., 2021). Before running anything we checked the extracts against the public record: they contain 3,236,107 person rows, the published folktables sample size; the ACS Income filter (age > 16, income $> \$100$, hours worked > 0) returns exactly the published 1,664,500 rows, and the ACS Public Coverage filter returns exactly the published 1,138,289 once not-applicable entries are excluded. Codes are raw integers from the PUMS data dictionary, not normalized or rescaled; the state FIPS code agrees with the file in every row. The only departure from the raw record is that not-applicable entries are stored as 0 rather than missing; none survives the ACS Income filter in the features we use. The extracts also carry a derived label column thresholded at \$25k; we do not use it, and derive $Y = \mathbf{1}[\text{income} > \$50\text{k}]$ from personal income as in the main benchmark.

State selection. States were ranked by the number of positive examples in the minority group ($A = 0, Y = 1$) in the full ACS Income population of the state, which reflects both population size and demographic composition and governs whether an EO gap can be estimated. The top 26 were used; the rule was applied before any results were inspected, and no state was added or removed afterwards.

Upstream block. Two-sided label bias: $Y_2 = 1$ where $Y = 0$, $A = 1$ and $U < \lambda$; $Y_2 = 0$ where $Y = 1$, $A = 0$ and $U < \lambda$; otherwise $Y_2 = Y$, with $U \sim \mathrm{Uniform}[0, 1]$ drawn independently per row. A one-sided variant that only relabels minority positives was piloted first and discarded before the reported runs, because its effect is bounded by $P(A = 0, Y = 1)$, which is under 2% of rows in a typical state and too small to detect.

Estimation. Hours worked (`WKHP`) is quantized into 10 equal-frequency bins, and the categorical features `COW`, `SEX`, and `SCHL` keep their categories. Certificates use exact Shapley aggregation (Eq. (1)) over all 2^4 coalitions; MI and CMI are computed from the empirical joint distribution of the quantized variables, which is the quantity `dit` returns, and were checked against a closed-form evaluation. Downstream models (a two-layer MLP, HistGradientBoosting, logistic regression) are trained on a stratified 67% split and evaluated on the rest; the EO gap is $\max_y |P(\hat{Y} = 1 \mid A =$

Table 13: Robustness across six downstream predictors on ACS Employment.

Measure	CatBoost	LGBM	XGB	RandomForest	Logistic	HistGradientBoosting	Avg. error
$v_{A,MI}$	1.1 ± 0.0	0.9 ± 0.0	0.9 ± 0.0	1.1 ± 0.1	2.6 ± 0.1	1.0 ± 0.0	1.3 ± 0.6
$v_{A,CMI}$	0.8 ± 0.0	0.6 ± 0.0	0.6 ± 0.0	0.8 ± 0.1	2.3 ± 0.1	0.8 ± 0.0	1.0 ± 0.6
$v_{SP,MI}$	5.2 ± 0.1	3.0 ± 0.0	2.7 ± 0.0	4.0 ± 0.5	3.5 ± 0.2	5.5 ± 0.3	4.0 ± 1.1
$v_{SP,SI:1}$	5.0 ± 0.1	3.0 ± 0.1	2.7 ± 0.0	4.0 ± 0.4	3.4 ± 0.2	5.4 ± 0.3	3.9 ± 1.0
$v_{SP,SI:2}$	3.9 ± 0.1	2.4 ± 0.1	2.6 ± 0.0	2.6 ± 0.4	2.0 ± 0.1	4.2 ± 0.3	2.9 ± 0.8
$v_{D,Prod}$	5.2 ± 0.1	3.0 ± 0.1	2.6 ± 0.0	3.8 ± 0.6	3.5 ± 0.2	5.5 ± 0.3	3.9 ± 1.1
$v_{EO,CMI}$	10.1 ± 0.8	3.5 ± 0.1	3.0 ± 0.1	6.8 ± 1.4	6.0 ± 0.4	9.7 ± 0.9	6.5 ± 2.7
$v_{D,Prod}$	8.8 ± 0.8	2.3 ± 0.1	2.1 ± 0.1	4.8 ± 1.4	4.0 ± 0.4	8.5 ± 0.9	5.1 ± 2.7

$1, Y = y) - P(\hat{Y} = 1 \mid A = 0, Y = y)|$ against the true Y. The harm-direction statistic is the Spearman correlation between λ and the mean realised EO gap, averaged over the three models.

Synthetic pipeline. $A \sim \mathrm{Bern}(0.5)$; $X_1 = 1.2A + \varepsilon_1$, $X_2 = 0.4A + \varepsilon_2$, $X_3 = \varepsilon_3$, $X_4 = \varepsilon_4$; $Y = \mathbf{1}[0.9X_1 + 0.6X_2 + 0.8X_3 + 0.5X_4 + 0.8\varepsilon_5 > q_{0.6}]$, with $\varepsilon_j \sim \mathcal{N}(0,1)$ independent and $q_{0.6}$ the 60th percentile of the score. Each feature is quantized into four equal-frequency bins; $n = 10{,}000$; 10 seeds.

Hypotheses fixed before the 26-state runs, and outcomes. **H1**: the EO certificate on Y_2 falls with λ. Supported: it falls at $\lambda = 0.5$ in 26/26 states (sign test $p = 3.0 \times 10^{-8}$), significantly in 23/26 (paired t-test over seeds, $p < 0.05$), with Spearman correlation against λ at or below -0.8 in 23/26. **H2**: the accuracy certificate on Y_2 falls with λ. Supported: strictly monotonically in 26/26. **H3**: across states, proxy strength $I(A; X_{[m]})$ predicts the direction of realised harm. Not supported: $\rho = +0.05$ ($p = 0.82$) over all λ, and $\rho = -0.17$ ($p = 0.40$) over $\lambda \leq 0.3$, where downstream accuracy has not yet collapsed. An exploratory analysis, not pre-specified, found no association between the loss of downstream accuracy and harm direction either ($\rho = -0.03$, $p = 0.90$). The SP certificate on Y_2 equals that on Y in every state and seed, as Remark 3 requires.

J Ideas and Reproducibility Summary

This section states what the certificates do not establish and what further experiments would be needed.

A stage-local certificate cannot establish that a downstream model will cause harm: Proposition 1 is one-sided, and slack arises when the downstream model suppresses the dependence a stage makes available. It also does not replace a model-level audit, does not establish formal privacy loss, and does not validate a high-dimensional learned representation outside the tabular, global-attribution scope of the reported experiments. Role separation nevertheless extends naturally to privacy and safety governance: an upstream certificate can be computed within the data holder's access set before training, while a model-side audit requires assembling $(X, A, Y, \hat{Y})$ across organizational boundaries.

A DP-budget experiment is outside the scope of this paper and is not reported here. To test the compatibility claim quantitatively, it would need to vary a declared privacy budget, quantify noise-induced changes in the group-conditional statistics used by SP and EO, and report both certificate fidelity and the Proposition 1 bound under the same protocol. The two-stage study of Section 5.5 instantiates the staged formulation for one upstream block with a known label-bias mechanism (Appendix I). Longer chains, learned upstream blocks whose bias is not known in advance, and bias entering through features rather than labels remain stress cases rather than completed experiments; each would require recording which block emitted each Y_k, preserving the stage boundary, and holding a reference supervision against which the change can be measured.

J.1 Reproducibility Summary

This subsection records the datasets, splits, seeds, hardware, and package settings used for the reported experiments.

Table 14: Two-stage study, per state, ordered by proxy strength. Minority positives: $A = 0, Y = 1$ in the full state population (selection criterion). Flips: share of labels changed by the upstream block at $\lambda = 0.5$. ΔAcc and ΔEO: relative change of $v_{A,MI}$ and $v_{EO,CMI}$ on Y_2 from $\lambda = 0$ to 0.5 (%). Seeds $\downarrow$ and p: seeds in which the EO certificate fell, and the paired t-test p-value. Harm: Spearman correlation between λ and the realised EO gap against the true Y, averaged over NN, HGB and LR, over all λ and over $\lambda \leq 0.3$; negative means the gap falls. NN acc.: downstream NN accuracy against the true Y at $\lambda = 0$ and 0.5.

State	Minority pos.	$I(A;X)$	Flips (%)	ΔAcc	ΔEO	Seeds $\downarrow$	p	Harm (all)	Harm ($\lambda \leq .3$)	NN acc.
WA	345	0.045	27	−53.3	−4.4	4/5	0.054	−0.10	+0.40	0.73→0.51
AZ	332	0.048	31	−53.8	−2.6	4/5	0.077	−0.16	+0.87	0.75→0.46
IN	371	0.051	32	−55.7	−7.3	5/5	0.013	−0.52	+0.13	0.76→0.46
NV	244	0.055	28	−52.1	−6.5	5/5	<0.001	−0.73	−0.20	0.72→0.53
MO	468	0.057	32	−53.3	−5.2	5/5	0.004	−0.35	+0.20	0.76→0.48
KY	236	0.057	32	−51.6	−6.4	5/5	0.007	−0.60	−0.07	0.76→0.47
PA	849	0.059	30	−53.7	−4.8	4/5	0.049	−0.89	−0.60	0.74→0.52
CA	2,951	0.063	27	−53.4	−3.9	5/5	0.008	−0.79	−0.47	0.74→0.55
MI	766	0.064	31	−54.6	−7.8	5/5	0.003	−0.85	−0.47	0.75→0.50
OH	975	0.064	30	−54.3	−9.2	5/5	<0.001	−0.66	−0.13	0.75→0.53
TX	3,422	0.065	29	−53.2	−2.1	4/5	0.260	−0.79	−0.33	0.75→0.53
MA	653	0.067	24	−53.7	−7.5	5/5	0.009	−0.89	−0.87	0.74→0.58
TN	824	0.070	31	−51.5	−12.4	5/5	<0.001	−0.71	−0.20	0.75→0.53
IL	1,496	0.075	27	−51.3	−11.7	5/5	<0.001	−0.83	−0.47	0.74→0.58
FL	2,183	0.080	29	−48.6	−11.3	5/5	<0.001	−0.92	−0.73	0.74→0.51
CT	454	0.080	24	−52.1	−11.0	5/5	<0.001	−0.92	−0.73	0.73→0.60
VA	1,920	0.087	25	−50.6	−10.4	5/5	0.002	−0.89	−0.60	0.74→0.63
NJ	1,529	0.091	23	−48.9	−11.6	5/5	<0.001	−0.98	−0.93	0.73→0.61
AL	807	0.092	28	−47.0	−13.2	5/5	<0.001	−0.71	−0.07	0.75→0.60
NY	3,764	0.093	26	−48.3	−10.9	5/5	0.001	−0.70	+0.07	0.72→0.56
NC	1,594	0.096	27	−46.6	−11.5	5/5	<0.001	−0.87	−0.87	0.75→0.58
GA	2,998	0.103	25	−48.5	−10.3	5/5	<0.001	−0.37	−0.47	0.75→0.67
MD	3,124	0.113	22	−50.9	−8.7	5/5	<0.001	−0.31	−0.40	0.74→0.67
MS	626	0.114	25	−41.7	−17.9	5/5	<0.001	+0.50	+0.00	0.76→0.71
LA	893	0.119	25	−41.8	−17.5	5/5	<0.001	+0.70	+0.40	0.74→0.64
SC	776	0.122	27	−42.9	−15.4	5/5	<0.001	−0.30	+0.33	0.75→0.61
Synthetic	–	0.235	17	−36.9	−34.8	10/10	<0.001	+1.00	+1.00	0.83→0.80

The seven datasets are ACS Income, ACS Employment, ACS Public Coverage, Adult, Census Income KDD, COMPAS, and Heritage Health. The benchmark uses the 67% training partition, a validation split for tuning the NN, and held-out data for the reference; errors are averaged over 3 seeds. The NN is a fully connected classifier over quantized features, and CatBoost, LightGBM, XGBoost, Random Forest, Logistic Regression, and HistGradientBoosting are tuned on the same partition. Each measure is a single task on one node with two CPU cores and 16 GB RAM. Continuous variables are quantized into 10 equal-frequency bins in the benchmark; MI and DC use `dit`, and shared information uses `BROJA_2PID` with MOSEK and ECOS backup. Kernel measures use a Gaussian RBF kernel with $\sigma = 0.5$; MMD and HSIC use variance-controlled subsampling and NOCCO uses batches of 512. The relevant estimator protocols are permutation sampling (Castro et al., 2009) and causally constrained asymmetric Shapley aggregation (Frye et al., 2020).

K Conclusion

Alignment auditing is a multi-party process in which no party sees the whole system. We took that as a specification rather than an obstacle: a data auditor holding no model can still issue certificates that provably bound the fairness gap of every downstream model and that transfer across seven predictor families. The pre-hoc attribution gap prices the information asymmetry—modest for well-chosen instruments, enormous for badly chosen ones, and traceable to whether an instrument's scale moves with coalition size. A two-stage study adds a condition: a stage-local certificate detects that its supervision changed, but is interpretable only together with the provenance of that supervision. The resulting guidelines concern construction rather than any single measure: build for the access regime, validate the instrument where it will be used, publish the discretion it leaves open.