Journal of Business and Management Studies (JBMS)
ISSN: 2709-0876
DOI: 10.32996/jbms
Journal Homepage: www.al-kindipublisher.com/index.php/jbms

JBMS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Auditable Credit Risk Intelligence for the U.S. Financial System: A Scalable Explainable-AI Framework Reconciling Predictive Performance with ECOA, FCRA, and Model Risk Governance

Hasibur Rahman[1] ✉, Sarder Abdulla Al Shiam[2] and Md Sibbir Hossain[3]
[12] *Department of Business Analytics, St. Francis College, 179 Livingston Street, Brooklyn, NY 11201, USA*
[3] *Department of Computer Science, The City College of New York, 160 Convent Ave, New York, NY 10031, USA*
Corresponding Author: Hasibur Rahman, **E-mail:** hrahman@sfc.edu

| ABSTRACT

Machine-learning credit underwriting in the United States operates under three families of obligation that are not reducible to one another: a statutory prohibition on discrimination, a statutory duty to disclose the specific principal reasons for an adverse action, and prudential expectations for model risk management. Prevailing practice reconciles these demands by sacrificing model capacity, deploying low-capacity scorecards whose interpretability is structural rather than earned. This study argues that the United States regulatory realignment of 2025 and 2026 makes that settlement less defensible, not more. The Consumer Financial Protection Bureau withdrew the interpretive circulars governing algorithmic adverse-action practice and amended Regulation B to disclaim disparate-impact liability under the Equal Credit Opportunity Act, while leaving the statutory disclosure duty untouched and preserving liability for the intentional use of facially neutral proxies. The federal banking agencies simultaneously replaced the prescriptive 2011 interagency model risk guidance with a principles-based instrument that expressly disclaims enforceable standards. Fewer obligations are now externally specified, and more must be self-specified, self-justified, and self-evidenced. Because effects-based exposure persists under the Fair Housing Act and state analogues, and private litigation is unaffected by federal guidance withdrawal, the value of verifiable self-generated evidence rises as external specification recedes. The paper develops ACRIS, the Auditable Credit Risk Intelligence Stack, a five-layer architecture in which admissibility, explainability, disparity control, and governance evidence are enforced during training and serving rather than audited afterwards. Its layers comprise a provenance-gated feature-admissibility screen bounding residual proxy information through conditional mutual information; a shape-constrained predictive core; a constrained-optimization layer recording an entire searched alternative-model frontier, including rejected candidates and rejection rationales; an explanation engine deriving principal reasons from a counterfactual approval baseline and gating disclosure on a per-decision stability margin; and an append-only, tamper-evident governance ledger. A finite-sample sufficient condition for top-k reason-set preservation under parameter resampling is proved, with its assumptions and failure modes stated explicitly. The framework is not empirically validated. A pre-registered protocol of falsifiable propositions, corpora, temporal validation design, baselines, metrics, statistical plan, and pre-committed disconfirmation criteria is specified so that every central claim can be tested and, if wrong, refuted.

| KEYWORDS

Credit risk analytics; explainable artificial intelligence; algorithmic fairness; model risk governance; regulatory compliance; adverse action notices; monotonic machine learning; auditability; financial technology; business analytics

| ARTICLE INFORMATION

ACCEPTED: 02 January 2026 **PUBLISHED:** 11 March 2026 **DOI:** 10.32996/jbms.2026.8.3.5

1. Introduction

Consumer credit allocation is among the highest-volume consequential decision processes in the modern economy. Every underwriting decision is simultaneously a statistical prediction of a latent default propensity, a

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

legally consequential act governed by anti-discrimination law, and an artifact potentially subject to supervisory examination, private litigation, or state enforcement. A decade of benchmarking has established that gradient-boosted decision trees and, to a lesser degree, deep tabular architectures deliver material improvements in discriminatory power over the logistic-regression scorecard that has anchored industry practice since the 1960s (Gunnarsson et al., 2021; Lessmann et al., 2015). Adoption in regulated first-lien and unsecured consumer portfolios nonetheless remains conservative. The obstacle is not predictive; it is evidentiary. A model that scores well but cannot state why it declined a particular applicant, cannot show which alternative designs were considered, and cannot prove that the explanation shown to a consumer was generated by the model version claimed at the time of the decision is not deployable in a regulated portfolio, irrespective of its area under the curve. This study therefore treats the evidentiary deficit, rather than the predictive gap, as the binding constraint on responsible adoption.

Three families of obligation bear on a United States consumer credit model, and they are close to orthogonal. The Equal Credit Opportunity Act (ECOA), 15 U.S.C. § 1691, prohibits discrimination on a prohibited basis in any aspect of a credit transaction, and its implementing regulation, Regulation B, 12 C.F.R. part 1002, gives that prohibition operational content. Section 615(a) of the Fair Credit Reporting Act, 15 U.S.C. § 1681m(a), together with Regulation B § 1002.9, requires that an applicant subject to adverse action receive a statement of the specific principal reasons for that action; a reason code is therefore not a user-experience artifact but an attested claim about the basis of one particular decision, and the duty attaches individually rather than in aggregate. Federal banking agencies expect regulated institutions to identify, measure, and control model risk across the model lifecycle, with independent validation and effective challenge (Board of Governors of the Federal Reserve System, 2026; Office of the Comptroller of the Currency, 2026); this family is indifferent to accuracy as such and asks instead whether the institution can demonstrate that it understands and controls the model.

The orthogonality of these demands defeats single-axis solutions. A model can be accurate and statistically even-handed yet fail its disclosure duty because its explanations are unstable across refits. It can be interpretable and well-governed yet remain exposed under ECOA because a facially neutral feature functions as a proxy. It can be even-handed and explainable yet be indefensible in an examination because no verifiable record links a production decision to the model version, feature vector, and validation evidence in force when the decision was made. Frameworks that optimize a single axis, the dominant pattern in the applied explainable artificial intelligence literature, therefore do not solve the deployment problem.

The regulatory environment against which such frameworks must be designed changed materially during 2025 and 2026, and in a direction that is easy to misread. In May 2025 the Consumer Financial Protection Bureau withdrew sixty-seven guidance documents, among them the circulars addressing adverse-action notification in connection with credit decisions based on complex algorithms and the proper use of the Bureau's sample forms (Consumer Financial Protection Bureau, 2025); those circulars had been the most frequently cited authority for the proposition that algorithmic complexity does not excuse a creditor from stating specific principal reasons. In April 2026 the Bureau published a final rule amending Regulation B that removes the longstanding effects-test language and states the Bureau's conclusion that ECOA does not authorize disparate-impact liability, while preserving disparate-treatment liability, including liability for the intentional use of facially neutral criteria as proxies, and expressly disclaiming any bearing on disparate-impact liability under the Fair Housing Act or state-law analogues (Consumer Financial Protection Bureau, 2026). Also in April 2026, the federal banking agencies jointly issued revised model risk management guidance superseding the 2011 interagency guidance (Board of Governors of the Federal Reserve System, 2026; Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011; Office of the Comptroller of the Currency, 2026). The revised guidance is substantially shorter and more principles-based, narrows the definition of a model to require complexity, allows validation rigor to vary with model use, and states expressly that it does not set forth enforceable standards or prescriptive requirements.

It would be a mistake to read these developments as a relaxation of the design problem. The statutory disclosure duty is untouched; only the interpretive gloss was withdrawn. Proxy use remains actionable, now routed through disparate treatment rather than effects. Fair Housing Act and state-law exposure is undisturbed, and private litigation and state supervision are unaffected by federal guidance withdrawal. What has changed is the locus of specification: fewer obligations are spelled out by the supervisor, and more must be specified, justified, and evidenced by the institution itself. This observation is the pivot of the present study. Under a prescriptive regime, compliance can be discharged in substantial part by conformance to an enumerated checklist, and self-generated evidence adds comparatively little. Under a principles-based regime with intact statutory duties and undiminished private and state exposure, an institution's position rests on the quality of the record it can produce about its own choices. A modeling stack that emits verifiable evidence as a by-product of ordinary operation is therefore more valuable after 2026 than before it, not less.

Formally, a lender seeks a scoring function, an explanation operator, and a governance record that jointly solve the

$$\max_{\theta,\Phi,\mathcal{G}} \mathrm{AUC}(f_\theta) \quad \text{s.t.} \quad \mathrm{Disp}(f_\theta,\tau)\le \bar{d},\ \ \mathcal{S}(\Phi)\ge\sigma,\ \ \mathcal{F}(\Phi,f_\theta)\ge\varphi,\ \ \mathcal{G}\vDash\mathcal{C}$$

constrained problem in Equation 1, in which Disp is a declared disparity functional evaluated at approval threshold tau, S and F are explanation-stability and explanation-fidelity functionals, and C is a set of auditable control predicates.

Two features of this formulation deserve emphasis. The constraint set is declared rather than derived: the choice of disparity functional, of its tolerance, and of the stability threshold is a normative and legal act that no algorithm can make, and the framework's task is to surface and record that act rather than to conceal it inside a loss function. Furthermore, the quantity of practical interest is not the optimum itself but its degradation relative to the unconstrained maximum, which is to say the price of auditability. The literature routinely asserts this price to be large, and to our knowledge it has never been measured under a specification that encodes all three obligation families jointly. Four design objectives follow: constraints should bind jointly during training and serving rather than being applied as successive filters whose optima degrade one another; explanation semantics should align with the statutory question, which is a boundary counterfactual rather than a deviation from a population mean; assurance should attach per decision rather than at portfolio level, because the disclosure duty attaches individually; and governance artifacts should be independently verifiable without trusting the first or second line of defence.

Against this background, the study makes five contributions. First, it provides a current and precisely delimited regulatory synthesis characterizing the United States obligation landscape for credit models as of mid-2026, separating binding statute and regulation from expressly non-binding supervisory guidance, from withdrawn guidance, and from the authors' own proposals, and argues that the 2025-2026 retrenchment strengthens rather than weakens the case for self-evidencing model architectures. Second, it specifies ACRIS, the Auditable Credit Risk Intelligence Stack, a five-layer reference architecture in which each layer discharges a named obligation family and emits a named artifact. Third, it introduces conditional-mutual-information proxy screening, defining a residual proxy load that bounds the information a candidate feature carries about a protected attribute after conditioning on legitimate risk factors, adjudicated against a stratified permutation null with false-discovery-rate control, and placing features that are both proxy-laden and predictively material into an explicit quarantined state requiring a documented necessity record rather than automatic exclusion. Fourth, it reformulates principal reasons as attributions against a counterfactual approval baseline and proves a finite-sample sufficient condition under which the top-k reason set is preserved under bootstrap resampling, yielding a deployable gate that routes low-margin decisions to manual adjudication, with the proposition's assumptions stated candidly including where they fail. Fifth, because the framework is not yet empirically validated, it specifies in advance the propositions to be tested, the corpora, the temporal validation design, the baselines, the metrics, the statistical analysis plan, and the results that would disconfirm each proposition.

The epistemic status of this work should be stated without ambiguity. This article reports no experimental results. It contains no measured area under the curve, no measured disparity ratio, no measured explanation-stability statistic, no latency or throughput measurement, and no human-subjects evaluation. Every figure is a conceptual schematic and none plots data. The framework is a design proposal supported by an analytical argument and a formal proposition, presented together with the protocol required to test it; claims of validation would be unwarranted and are not made. Nothing in this article constitutes legal advice. The remainder of the paper proceeds as follows. Section 2 reviews the relevant literature and isolates five research gaps. Section 3 establishes the regulatory baseline as of mid-2026. Section 4 develops the ACRIS framework layer by layer. Section 5 maps obligations to controls, artifacts, and verification methods. Section 6 specifies the pre-registered evaluation protocol. Section 7 discusses implications and anticipated objections, Section 8 states limitations, and Section 9 concludes.

2. Literature Review

2.1 Machine Learning for Credit Scoring

The benchmarking literature is mature and broadly consistent. Lessmann et al. (2015) evaluated forty-one classifiers across eight credit datasets and established heterogeneous ensembles as the dominant family. Gunnarsson et al. (2021) found that deep learning offers no systematic advantage over gradient-boosted decision trees on tabular credit data once hyperparameter budgets are equalized. A parallel strand pursues intrinsically interpretable high-capacity models: neural additive models recover smooth univariate structure while retaining decomposability (Agarwal et al., 2021), and penalized logistic tree regression recovers much of a tree ensemble's nonlinearity within a scorecard-compatible functional form (Dumitrescu et al., 2022). Rudin (2019) argues that for high-stakes decisions such models should be preferred outright to post-hoc explanation of opaque ones, a position developed at length in Rudin et al. (2022).

2.2 Post-Hoc Explanation and Its Fragility

Local surrogate methods (Ribeiro et al., 2016) and Shapley-value attribution (Lundberg & Lee, 2017) dominate applied explainable artificial intelligence in finance, with tree-specific algorithms making exact Shapley computation tractable for ensembles (Lundberg et al., 2020). The fragility of this apparatus is equally well documented. Alvarez-Melis and Jaakkola (2018) showed that interpretability methods are not robust to small input perturbations. Slack et al. (2020) demonstrated that both local surrogate and Shapley explanations can be adversarially manipulated to conceal bias, a result with obvious implications for any compliance regime that treats an explanation as self-authenticating. Kumar et al. (2020) provided a foundational critique of Shapley values as feature-importance measures, noting that the additivity axiom does not license the causal reading thatpractitioners impose on it. Krishna et al. (2022) quantified the disagreement problem, in which distinct attribution methods produce divergent top-k rankings on identical instances,.

2.3 Algorithmic Fairness in Consumer Credit

The conceptual foundations of the field are set out by Barocas and Selbst (2016), who analyze how facially neutral data-driven processes reproduce disparity. Interventions are conventionally partitioned into pre-processing (Kamiran & Calders, 2012), in-processing (Agarwal et al., 2018), and post-processing (Hardt et al., 2016). Impossibility results establish that calibration and equalized odds cannot generally be satisfied jointly under differing base rates (Chouldechova, 2017; Kleinberg et al., 2017), which means that criterion selection is a normative act requiring documentation. This is precisely the kind of act that the post-2026 regime pushes onto the institution, and it is an important reason why the framework proposed here records the choice rather than making it silently.

Kozodoi et al. (2022) provide the definitive empirical study in credit scoring, benchmarking fairness processors and quantifying profit implications. Hurlin et al. (2024) develop testing procedures aligned to credit-market institutions. Fuster et al. (2022) and Bartlett et al. (2022) document heterogeneous, group-dependent effects of algorithmic underwriting in United States markets,.

A distinctive United States complication is that Regulation B restricts the collection of race and ethnicity in most non-mortgage credit, so the protected attribute is typically unobserved. Surname and geography based imputation (Elliott et al., 2009) is the standard operational workaround. Chen et al. (2019) show formally that disparity estimates computed on imputed attributes are generally biased, and that the direction and magnitude of the bias are not guaranteed to be conservative, a point the applied fairness literature propagates far too rarely. Rubin's rules for multiple imputation (Rubin, 1987) provide the appropriate machinery for propagating that uncertainty into interval estimates, and are adopted here.

2.4 Model Risk Governance and Auditability

Governance scholarship remains largely practitioner-facing. Model cards propose a documentation standard for reporting model provenance, intended use, and performance disaggregated by relevant subgroup (Mitchell et al., 2019). Tamper-evident logging and provenance infrastructure is mature in the systems community (Crosby & Wallach, 2009) but has not been integrated into supervised credit-risk pipelines.

2.5 Synthesis of Research Gaps

Five gaps emerge from the review, and the framework developed in Section 4 is designed to close them jointly rather than serially.

- G1. Sequential rather than joint constraint handling. Accuracy, disparity, interpretability, and governance are optimized in separate stages, so each stage's optimum is degraded by the next and the joint frontier is never characterized.
- G2. Marginal rather than conditional proxy detection. Screening in practice uses marginal association with a protected attribute, conflating illegitimate proxy leakage with legitimate correlation with creditworthiness, and thereby both over-excluding and under-excluding features.
- G3. Explanation semantics misaligned with the disclosure duty. Attributions are computed against a population baseline where the duty poses a boundary counterfactual; stability is measured in aggregate if at all, and never certified per decision.
- G4. The alternative-model search is unoperationalized. No reproducible procedure exists for demonstrating which less disparate designs were considered and on what grounds each was rejected.
- G5. Governance evidence is assertive rather than verifiable. No tamper-evident linkage exists between production decisions and the model and validation state that produced them.

3. The U.S. Regulatory and Governance Landscape as of Mid-2026

Framework papers in this area routinely conflate statutory command, regulatory text, supervisory expectation, and

scholarly recommendation. The distinction is not pedantic: it determines what an institution must do, what it may be criticized for not doing, and what is merely good practice. This section fixes the baseline against which the framework of Section 4 is specified.

3.1 ECOA and Regulation B After the 2026 Amendment

ECOA makes it unlawful for a creditor to discriminate against an applicant on a prohibited basis, comprising race, colour, religion, national origin, sex, marital status, age, receipt of public assistance income, and the good-faith exercise of rights under the Consumer Credit Protection Act, in any aspect of a credit transaction. The statute is unchanged. Regulation B is not. The final rule published in April 2026, effective July 2026, amended Sections 1002.2(p), 1002.4(b), 1002.6(a), and 1002.8 (Consumer Financial Protection Bureau, 2026). It deleted language indicating that disparate-impact liability may attach under ECOA and stated the Bureau's position that the Act does not recognize the effects test.

Three consequences follow for model design, and they must be stated exactly. First, proxy use remains actionable. The rule preserves disparate-treatment liability and states that facially neutral criteria used as proxies for a prohibited basis continue to be analyzed as intentional disparate treatment. A feature that functions as a proxy is therefore not made safe by the amendment; the theory of liability has shifted, not disappeared, and screening for proxy load remains a first-order design requirement. Second, effects-based exposure persists outside ECOA. The rule expressly disclaims any bearing on disparate-impact liability under the Fair Housing Act or state-law analogues. For mortgage and housing-related credit in particular, the effects-based question survives intact, and a lender's ability to show that it searched for and evaluated alternative model designs retains its evidentiary value. Third, statistical evidence retains a role. Disparity statistics do not become irrelevant under a disparate-treatment theory; they become evidence bearing on intent and on inconsistent application rather than freestanding grounds for liability. Measuring and recording them is therefore still the prudent course.

3.2 FCRA Section 615(a) and Regulation B Section 1002.9

The adverse-action disclosure duty is statutory and unchanged. Section 615(a) of the Fair Credit Reporting Act requires a person who takes adverse action based in whole or in part on information in a consumer report to provide notice and specified disclosures, and Regulation B Section 1002.9 requires that a statement of reasons be specific and indicate the principal reasons for the adverse action. Neither instrument contains a complexity exception. This obligation is the single most demanding constraint on high-capacity credit models, because it attaches to each individual decision and requires that the stated reasons reflect the actual basis of that decision rather than a generic checklist selection. It is also the obligation that the existing explainable artificial intelligence toolkit is least well equipped to discharge, for the reasons set out in Section 2.2.

3.3 Model Risk Governance: From the 2011 Guidance to the 2026 Revision

The 2011 interagency guidance organized a generation of model risk practice around conceptual soundness, ongoing monitoring, outcomes analysis, independent validation with effective challenge, and a governed model inventory with change control (Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011). In April 2026 the agencies replaced it (Board of Governors of the Federal Reserve System, 2026; Office of the Comptroller of the Currency, 2026). The revised guidance is principles-based and comparatively concise. It narrows the model definition to require complexity and excludes deterministic rule-based processes and simple arithmetic calculations. It allows the timing, nature, and frequency of validation to vary with a model's purpose and use, including phased validation where a business need is urgent, provided limitations are communicated and compensating controls applied. It retains effective challenge, defined as critical analysis by objective, competent parties with the standing to compel change. And it states in terms that it does not set forth enforceable standards or prescriptive requirements, so that non-compliance will not itself result in supervisory criticism. Generative and agentic artificial intelligence are placed outside its scope pending further guidance.

For a credit-scoring model the practical effect is threefold. The model remains squarely in scope, since it is neither deterministic nor arithmetically simple. The institution now bears more of the burden of specifying what adequate validation means for its own use case. And the value of evidence that an outside party can verify without trusting the institution's own assertions increases correspondingly, because there is less external specification to point to.

3.4 What the 2025 Withdrawals Did and Did Not Do

The withdrawal of the adverse-action circulars removed interpretive authority; it did not amend Section 1002.9 or Section 1681m(a), which are the operative sources of the disclosure duty (Consumer Financial Protection Bureau, 2025). A framework that had been designed around the circulars' language would now rest on a vacated foundation. ACRIS is deliberately specified against the statutory and regulatory text, and records reliance on any interpretive

material as a dated, reviewable assumption in the governance ledger rather than embedding it in a control. This is a small design decision with a large consequence for durability, and it generalizes: in a period of interpretive volatility, the assumptions a system makes about guidance should be first-class, versioned, and revisited on a schedule.

3.5 Residual Enforcement Channels

Federal supervisory retrenchment does not exhaust the relevant exposure. Fair Housing Act effects liability persists for housing-related credit. State regulators and state analogue statutes operate independently of federal guidance. Private litigation is unaffected. State artificial intelligence statutes reaching algorithmic decisions that affect access to lending are arriving, with Colorado's framework having been postponed and subsequently replaced by a successor statute with a 2027 effective date. An institution designing to the federal floor alone is designing to the least demanding of its several audiences,.

3.6 A Taxonomy of Instruments

Table 1 records the taxonomy used throughout the remainder of the paper. Every claim made here about an obligation is tagged to one of these categories, and the framework's controls are designed to be robust to change in the weaker categories. In particular, no ACRIS control depends solely on an instrument in category C or category D.

Table 1. Taxonomy of instruments used in this paper. No ACRIS control depends solely on an instrument in category C or D.

	Category	Instances relevant to this paper
A	Statute (binding)	Equal Credit Opportunity Act, 15 U.S.C. § 1691; Fair Credit Reporting Act § 615(a), 15 U.S.C. § 1681m(a); Fair Housing Act, 42 U.S.C. § 3601
B	Implementing regulation (binding)	Regulation B, 12 C.F.R. pt. 1002, as amended at 91 Fed. Reg. 21620, effective 21 July 2026
C	Supervisory guidance (expressly non-binding)	SR Letter 26-2; OCC Bulletin 2026-13; FIL-15-2026 (17 April 2026)
D	Withdrawn guidance (no longer authority)	CFPB Circulars 2022-03 and 2023-03, withdrawn 12 May 2025
E	Voluntary standards and frameworks	NIST AI Risk Management Framework 1.0 (NIST AI 100-1)
F	This paper's proposals	Residual proxy load; boundary-anchored reason codes; per-decision margin gate; ledger schema

The design discipline this taxonomy enforces is worth making explicit. A control that depends on a withdrawn circular fails silently when the circular is withdrawn, and the institution discovers the failure during an examination or a deposition rather than during a design review. A control that depends on a statutory duty fails only if the statute changes, which is a slower and far more visible process. Wherever a choice existed, ACRIS anchors the control to the more durable instrument and records the dependence on the less durable one as an assumption.

4. The ACRIS Framework

4.1 Notation and Architectural Overview

Let X denote the admissible feature space, y a binary default indicator defined over a stated performance window, and a the protected attribute. Write p-hat for the predicted default probability and let the approval decision be the indicator that this probability does not exceed a cutoff tau. Partition the features into a monotone set carrying a domain-mandated direction and a free set, and identify separately a consumer-mutable subset admissible in recourse recommendations, as distinct from immutable or time-only-mutable features such as the age of the oldest account.

Figure 1 gives the architecture. Five layers are composed in sequence. Each discharges a named obligation family, emits a named artifact, and writes that artifact to the append-only ledger constituted by Layer 5. The right-hand column of the figure records which instrument each layer is designed to address; it is a design mapping chosen by the authors, not a compliance opinion.

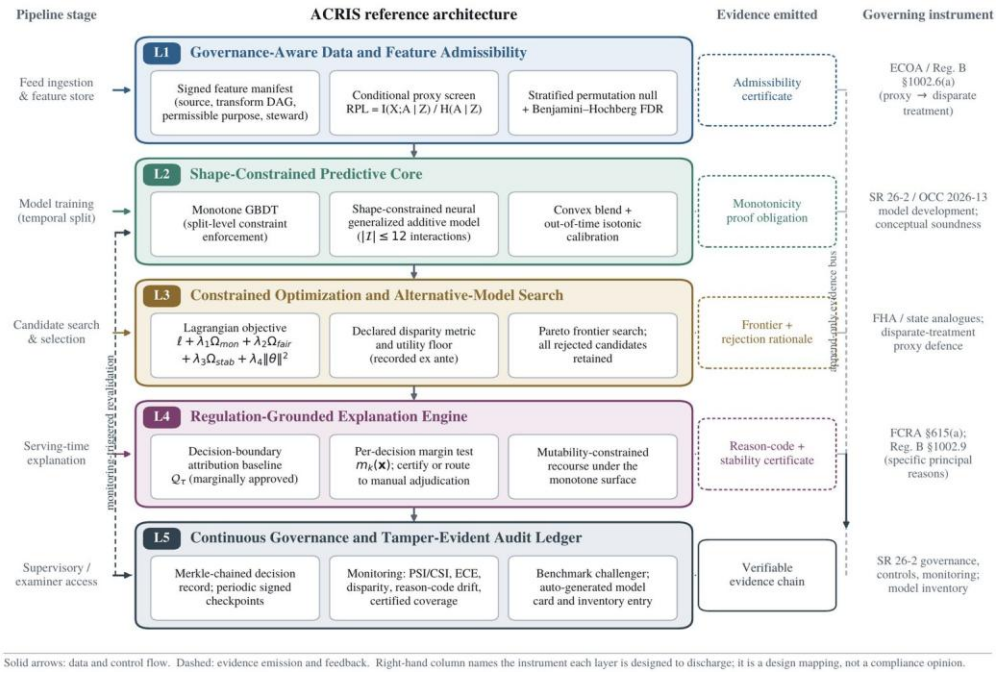

Figure 1. The ACRIS reference architecture. Each layer is paired with the instrument family it is designed to discharge and with the artifact it emits to the append-only ledger. Solid arrows denote data and control flow; dashed arrows denote evidence emission and feedback. This is a conceptual schematic and plots no data.

4.2 Layer 1: Governance-Aware Data and Feature Admissibility

4.2.1 Provenance and lineage

Every feature is registered with a signed manifest recording its source, transformation directed acyclic graph, refresh cadence, permissible purpose, data steward, and content hash. Features lacking a complete manifest are inadmissible by construction.

4.2.2 Conditional proxy screening

Marginal dependence between a candidate feature and a protected attribute is a poor admissibility criterion. Income correlates with race in United States data and is nonetheless an unambiguously legitimate underwriting factor; excluding on marginal association would gut the model while leaving subtler proxies untouched. The operative question is whether a feature carries information about the protected attribute beyond what is already conveyed by an established core of legitimate risk factors Z, such as debt-to-income, payment history, and credit-file depth. We therefore define the residual proxy load as in Equation 2.

$$\mathrm{RPL}(X_j) \;=\; \frac{I(X_j; A \mid \mathbf{Z})}{H(A \mid \mathbf{Z})} \;\in\; [0, 1]$$

This quantity is the fraction of the residual uncertainty in the protected attribute that a candidate feature resolves after conditioning on the legitimate core. Conditional mutual information is estimated using a nearest-neighbour estimator of the Kraskov family for continuous variables (Kraskov et al., 2004) and by plug-in estimation with bias correction for discretized variables. Because the estimator is positively biased in finite samples, admissibility is adjudicated against a permutation null in which the protected attribute is shuffled within strata of Z, with false-discovery-rate control across features (Benjamini & Hochberg, 1995).

Two criteria are required jointly, and the reason is worth stating. Statistical significance alone is uninformative at the sample sizes typical of consumer credit, where trivially small dependencies are reliably detectable. Effect size alone is uncalibrated against sampling noise. The decision rule that results is shown in Figure 2.

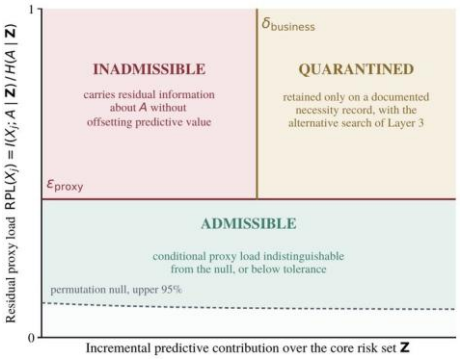

Figure 2. The feature-admissibility decision plane. A feature is admitted when its residual proxy load is indistinguishable from the stratified permutation null or falls below the declared tolerance; it is quarantined for documented necessity review when it is proxy-laden but materially predictive; otherwise it is excluded. Conceptual schematic: axis scales are unlabelled and no feature positions are plotted.

4.3 Layer 2: Shape-Constrained Predictive Core

4.3.1 Monotone gradient boosting

Monotonicity is the most examiner-legible property a credit model can possess. A reviewer can verify that increasing delinquency never decreases predicted risk without reading a line of code, and the constraint eliminates the class of locally counterintuitive behaviours that make post-hoc explanations indefensible. The requirement is that the partial derivative of the score with respect to every constrained feature, multiplied by that feature's mandated sign, be non-negative at every point of the admissible feature space.

It is enforced exactly in the boosted component by constraining split-candidate acceptance: a split on a monotone feature is admissible only if the resulting child-node value ordering respects the mandated direction, with sibling values clipped to the running bound. This yields global rather than in-sample monotonicity, which is the property the constraint must have to survive validation on data drawn from a later period.

4.3.2 Shape-constrained additive component

To recover smooth low-order structure that axis-aligned trees approximate with staircases, a neural generalized additive component with a restricted set of pairwise interactions is added (Agarwal et al., 2021), with the interaction budget capped so that every shape function remains visually auditable. Monotonicity is imposed on the univariate shape functions by a penalty evaluated on a quasi-random grid, supplemented at certification time by exact verification (Liu et al., 2020), since a penalty evaluated on a finite grid establishes monotonicity on that grid and not everywhere.

4.3.3 Blending and calibration

The core prediction is a convex blend of the boosted and additive components, with the mixing weight selected on a validation fold. Blending preserves monotonicity, since a convex combination of monotone functions is monotone, and preserves the auditability of the additive component. Scores are mapped to probabilities by isotonic regression fitted out of time: isotonic rather than parametric scaling, because a monotone calibration map cannot invert the shape constraints and because calibration drift is itself a monitored quantity in Layer 5.

4.4 Layer 3: Constrained Optimization and Alternative-Model Search

4.4.1 Composite objective

Training minimizes the composite objective in Equation 4, which combines log-loss with monotonicity, disparity, stability, and complexity penalties.

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, f_\theta(\mathbf{x}_i)) + \lambda_1\Omega_{mon} + \lambda_2\Omega_{disp} + \lambda_3\Omega_{stab} + \lambda_4\|\theta\|_2^2$$

Because thresholded approval rates are non-differentiable, the disparity term uses a smoothed approval-rate surrogate, given in Equation 5, in which the first term is a hinge on the declared threshold and the second is an equalized-odds relaxation.

$$\Omega_{disp} = \left[\max\left(0,\ \alpha - \frac{\min_a \pi_a}{\max_a \pi_a}\right)\right]^2 + \gamma \sum_{a \in \mathcal{A}} \left|\mathrm{TPR}_a - \overline{\mathrm{TPR}}\right|$$

The stability term penalizes attribution volatility under input perturbation, following the robustness formulation of Alvarez-Melis and Jaakkola (2018). It is estimated on Monte Carlo minibatches and is the mechanism by which explanation reliability is purchased during training rather than hoped for afterwards.

4.4.2 Documented alternative-model search

Where a lender must justify a design against less disparate alternatives, whether under the Fair Housing Act, under state analogues, or in support of a business-necessity position, the operative artifact is a record of what was searched. ACRIS produces that record as a by-product of model selection rather than as a retrospective reconstruction. Define a candidate space generated by admissible feature subsets from Layer 1, disparity weights, monotone-constraint configurations, and threshold policies. For a declared business-utility floor, the selection rule minimizes the declared disparity functional over the candidate space subject to expected utility remaining within a declared tolerance of that floor.

The search is executed by multi-objective Bayesian optimization over the frontier, and the entire frontier, including dominated and rejected candidates, is written to the ledger together with the reason for rejection. The evidentiary value lies not in the selected model but in the reproducible demonstration that alternatives were enumerated, evaluated against pre-declared criteria, and rejected for stated reasons.

4.5 Layer 4: Regulation-Grounded Explanation Engine

4.5.1 Global evidence

Evidence for conceptual soundness comprises shape-function plots with bootstrap bands readable directly by validators, global importance aggregated over a stratified sample, model-class reliance intervals bounding the range of importance consistent with near-optimal models (Fisher et al., 2019), and the monotonicity verification certificate emitted by Layer 2.

4.5.2 Boundary-anchored principal reason codes

Conventional practice reports the top-k features by Shapley value computed against the training-population mean baseline. That answers the question of why this applicant differs from the average applicant. The disclosure duty asks a different question: what were the principal reasons this applicant was declined, which is to say what drove the gap between the applicant's score and the approval boundary. We therefore define the boundary-anchored attribution in Equation 7.

$$\phi_j^{db}(\mathbf{x}) \;=\; \mathbb{E}_{\mathbf{x}' \sim \mathcal{Q}_\tau}\left[\phi_j(\mathbf{x}) - \phi_j(\mathbf{x}')\right]$$

Here the reference distribution comprises marginally approved applicants, those whose scores fall in a narrow band immediately below the cutoff, restricted where appropriate to a matched reference cohort so that reasons are comparative against genuinely similar approved applicants rather than against the population at large. Reason codes are the top-k features by this quantity, mapped through a version-controlled dictionary to plain-language statements.

4.5.3 A per-decision stability condition

A reason code that changes when the model is refit on a resample is a poor basis for a disclosure purporting to state the reasons for a particular decision. Let R denote the top-k reason set and let RBO denote rank-biased overlap, a similarity measure for indefinite rankings that weights agreement at the head of the list (Webber et al., 2010). An explanation operator is (epsilon, delta)-stable at a point if the probability that the rank-biased overlap between two bootstrap-refit reason sets is at least one minus epsilon exceeds one minus delta.

Assumption 1 (composite Lipschitz attribution). There exists a finite constant Lambda such that, for all parameter vectors in the neighbourhood explored by resampling, the supremum-norm distance between the corresponding boundary-anchored attribution vectors is bounded by Lambda times the Euclidean parameter distance.

Proposition 1 (finite-sample top-k set stability). Let Assumption 1 hold, let the top-k attribution margin at a point be the difference between the k-th and (k+1)-th largest attributions and be strictly positive, and let bootstrap refits satisfy a bound rho on expected parameter deviation. Then the probability that the top-k reason set changes is bounded as in Equation 8. Consequently, if the margin is at least twice Lambda times rho divided by delta, the operator is (epsilon, delta)-stable at that point for every non-negative epsilon.

$$\Pr\left[R_k(\mathbf{x};\theta^{(b)}) \neq R_k(\mathbf{x};\theta)\right] \;\leq\; \frac{2\Lambda\varrho}{m_k(\mathbf{x})}$$

Proof. Suppose the supremum-norm attribution perturbation is strictly less than half the margin. Take any feature in the top-k set and any feature outside it. The perturbed attribution of the former exceeds its original value less half the margin, which is at least the k-th largest original attribution less half the margin. The perturbed attribution of the latter is less than its original value plus half the margin, which is at most the (k+1)-th largest original attribution plus half the margin, and this equals the k-th largest original attribution less half the margin. The former therefore strictly exceeds the latter. Since this holds for every such pair, the top-k set is unchanged. Contrapositively, a change in the set requires the supremum-norm perturbation to reach half the margin, which by Assumption 1 requires the parameter deviation to reach the margin divided by twice Lambda. Markov's inequality applied to the parameter deviation yields the stated bound. When the set is unchanged the rank-biased overlap of two identical sets is one, so the stability condition holds for any non-negative epsilon. This completes the proof.

Three qualifications must be stated plainly, because the proposition is easily over-read. First, it bounds set stability, not rank stability: the argument establishes that the same k features are selected and says nothing about their internal ordering, so where the ordering itself carries disclosure significance a separate argument is required. Second, Assumption 1 does not hold universally. A gradient-boosted ensemble is piecewise constant in its inputs and is not Lipschitz in its parameters under the usual tree parametrization. The assumption is defensible for the differentiable additive branch and for the blended score under a smoothed surrogate, but for the tree branch the constant is not available analytically and must be estimated empirically from the resampling distribution, so that the resulting quantity is an estimate with its own sampling error rather than a certificate in the strict sense. Third, the bound is sufficient rather than necessary and is loose for small margins; where twice Lambda times rho exceeds the margin the bound is vacuous.

4.5.4 Actionable recourse

Beyond stating why, the framework can state what would change the outcome, subject to mutability and plausibility constraints, as in Equation 9.

$$\mathbf{x}^{cf} = \arg\min_{\mathbf{x}'} \; d_{MAD}(\mathbf{x},\mathbf{x}') + \mu\,\mathcal{P}(\mathbf{x}') \quad \text{s.t.} \quad f_\theta(\mathbf{x}') \leq \tau, \;\; \mathbf{x}'_{\bar{\mathcal{M}}} = \mathbf{x}_{\bar{\mathcal{M}}}, \;\; \mathbf{x}' \in \mathcal{C}$$

The distance is a median-absolute-deviation-normalized norm (Wachter et al., 2018), the immutable feature set is hard-constrained, and a density term restricts recommendations to the data manifold. The recourse literature has established the importance of actionability and plausibility (Karimi et al., 2023); the contribution here is narrower and structural. Global monotonicity from Layer 2 guarantees that the prescribed direction of change is outcome-improving everywhere, not merely locally. Under an unconstrained response surface an applicant who over-executes a recommendation can re-cross the boundary into denial, which is a recommendation that becomes indefensible once disclosed. Shape constraints eliminate that class of failure by construction rather than by solver design.

4.6 Layer 5: Continuous Governance and Tamper-Evident Ledger

4.6.1 The decision record

Each production decision emits a record comprising the decision identifier, a hash of the feature vector, the predicted probability, the applied cutoff, the model, feature and policy versions, the explanation payload, the attribution margin, and a timestamp, appended to a hash chain in which each root incorporates its predecessor, with periodic signed checkpointing (Crosby & Wallach, 2009).

Retrospective alteration of a decision, an explanation, or the model version attributed to a decision invalidates all subsequent roots and is therefore detectable. This supplies the property that effective challenge presupposes but that conventional logging does not deliver: a reviewer can verify, rather than trust, that the explanation shown to a consumer in a given month was generated by the model version claimed.

4.6.2 Ongoing monitoring

Monitored quantities, with two-tier warning and breach escalation, comprise population and characteristic stability indices, out-of-time discrimination decay, expected calibration error, the declared disparity metric with intervals that propagate protected-attribute imputation uncertainty by Rubin's rules where the attribute is imputed (Chen et al., 2019; Rubin, 1987), reason-code distribution drift measured against the validation baseline, and certified coverage, defined as the share of decisions clearing the stability margin. The last two are, to our knowledge, novel as monitored controls. Reason-code drift detects semantic drift in the justification of decisions even when aggregate disparity

is stable, a failure mode invisible to conventional monitoring. Certified coverage is the leading indicator for adjudication staffing and is itself drift-sensitive.

4.6.3 Support for effective challenge

A regularized scorecard challenger and the champion model are scored in parallel on every batch, and divergence beyond a governed tolerance triggers documented review. Model cards (Mitchell et al., 2019) are generated from ledger state rather than authored separately, which removes the documentation lag that examinations routinely surface. Figure 3 places these controls in the governed lifecycle and identifies which line of defence acts at each stage.

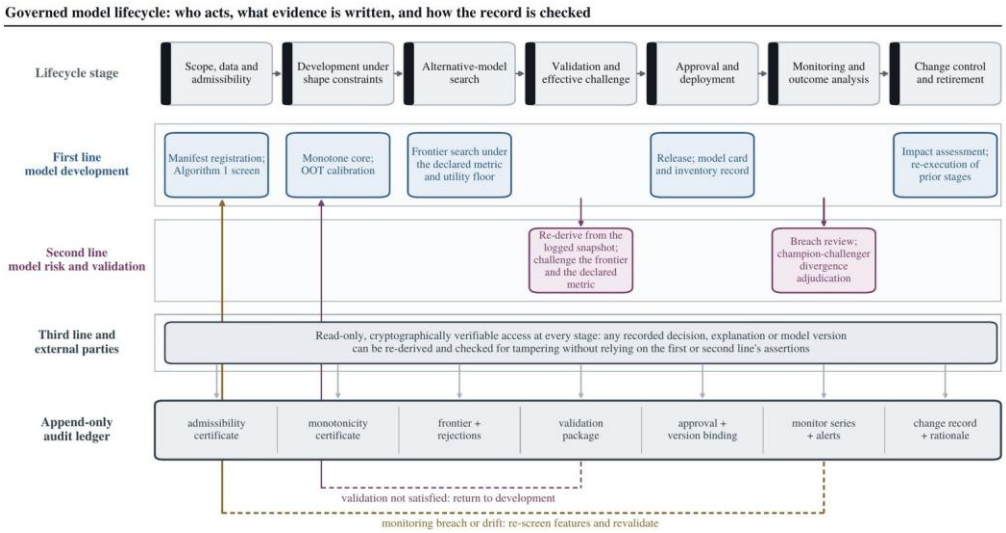

Figure 3. The governed model lifecycle under ACRIS, showing the three lines of defence, the evidence written to the append-only ledger at each stage, and the two feedback paths that return control to development. The third-line band expresses the design's central governance claim: verification does not require trusting the assertions of the first or second line. Conceptual schematic; no data are plotted.

4.7 Computational Complexity and Scalability

Training scales horizontally through data-parallel histogram aggregation, and serving scales through stateless replicas behind a shared model registry and an append-only ledger writer. Whether the resulting serving latency is compatible with real-time underwriting is an empirical question and is posed as such in Section 6; no latency figure is asserted here.

5. Obligation-to-Control Traceability

The framework's governance claim is not that it satisfies the instruments of Section 3. No artifact can satisfy an instrument; only a supervised institution can. The claim is narrower and more useful: that ACRIS renders each obligation into a predicate that is computable, a control that enforces the predicate, an artifact that records the enforcement, and a method by which a third party can check the artifact. Figure 4 presents this mapping in full.

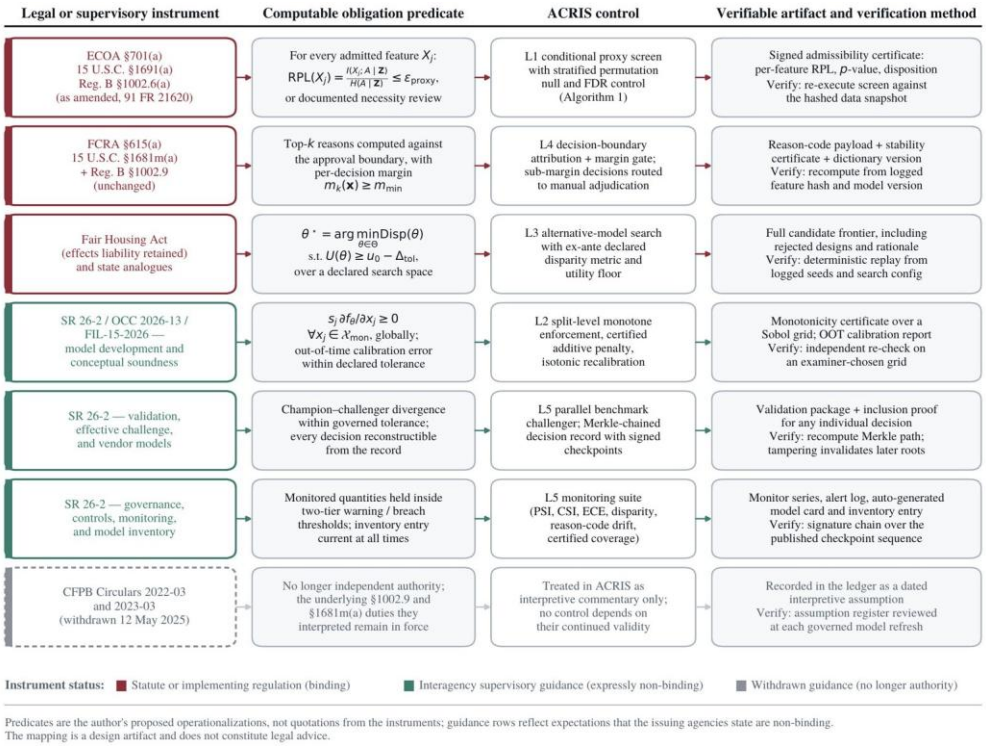

Figure 4. Obligation-to-control traceability. Each row renders an instrument into a computable predicate, the ACRIS control that enforces it, and the artifact together with the method by which an independent party can verify it. Instrument status is colour-coded per Table 1. The predicates are the authors' proposed operationalizations, not quotations from the instruments.

Two design commitments visible in that figure deserve explicit statement. First, no control depends on an instrument in category C or category D of Table 1. Withdrawing a circular or revising guidance changes what the institution must document about its assumptions; it does not change whether its controls continue to function. Second, every artifact is paired with a verification method that does not require trusting the producer. An admissibility certificate can be re-derived by re-executing the screen against a hashed data snapshot; a reason code can be recomputed from the logged feature hash and model version; a search frontier can be replayed deterministically from logged seeds and configuration; and a decision's presence in the ledger can be checked by recomputing a hash path, since any tampering invalidates every subsequent root. This is the operational content of design objective O4, and it is the respect in which the framework differs most sharply from documentation-based approaches, which record what a team asserts rather than what an independent party can confirm.

A practical corollary follows for managers. The artifacts in the right-hand column of Figure 4 are precisely the materials that an examiner, a state regulator, or an opposing expert will request, and they are produced automatically at the moment the underlying activity occurs rather than assembled under time pressure after a request arrives. The reduction in response latency is itself a governance benefit, and the elimination of retrospective reconstruction removes a well-known source of inconsistency between what an institution did and what it later reports having done.

6. Pre-Registered Evaluation Protocol

6.1 Epistemic Status of the Framework

The framework of Section 4 is a design proposal. Its component techniques, comprising monotone boosting, additive models, conditional mutual information estimation, Shapley attribution, and hash-chained logging, are individually established in the literature cited. Their composition, the boundary-anchored reason-code formulation, the per-decision margin gate, and the claim that joint constraint satisfaction is inexpensive are not established, and this paper does not establish them.

6.2 Falsifiable Propositions

- P1. The joint constraint penalty is small. Imposing monotonicity, disparity, and explanation-stability constraints jointly costs less discriminatory power, relative to an unconstrained gradient-boosted baseline, than the gap between that baseline and a weight-of-evidence logistic scorecard. Disconfirmed if the constrained model's out-

of-time discrimination is closer to the scorecard's than to the unconstrained baseline's on a majority of corpora.

- P2. Conditional screening dominates marginal screening. Excluding features by residual proxy load yields a better disparity-discrimination trade-off than excluding the same number of features by marginal association with the protected attribute. Disconfirmed if marginal screening attains equal or lower measured disparity at equal or higher discrimination.
- P3. The boundary baseline improves reason-code stability. Attributions computed against the marginal-approval cohort exhibit higher top-k rank-biased overlap under bootstrap refit than attributions computed against the population mean, holding the model fixed. Disconfirmed if the difference in mean overlap is not positive with a lower confidence bound above zero.
- P4. Shape constraints act as an explainability mechanism. Removing monotonicity degrades reason-code stability more than removing the explicit stability regularizer does. Disconfirmed if the ablation ordering reverses on a majority of corpora.
- P5. The governance apparatus is operationally viable. The marginal serving latency of attribution, margin evaluation, and ledger append is a minority of end-to-end decision latency at production throughput. Disconfirmed if governance overhead exceeds half of end-to-end latency, or if throughput scaling efficiency across nodes falls below a pre-declared floor.

6.3 Corpora

Four public corpora relevant to United States consumer credit are specified, chosen to span secured and unsecured credit, differing label definitions, and the presence versus absence of an observed protected attribute. These comprise the Freddie Mac Single-Family Loan-Level Dataset, the Fannie Mae Single-Family Loan Performance Data, the LendingClub accepted-loan data, and the Home Credit Default Risk data with bureau linkage. The first two are linkable to Home Mortgage Disclosure Act data at census-tract level, which permits an internal validation of the imputation used on the latter two against an observed attribute, a comparison the non-mortgage fairness literature is generally unable to perform. Record counts, feature counts, positive-class rates, and temporal spans are properties of the released files and are to be reported as measured; no such figures are stated in this paper.

6.4 Validation Design

All splits are strictly temporal. Models are trained on originations through a given period, validated on the following period, and tested from the period after that onward, with no intersection between the test window and training. Random cross-validation is additionally executed for the sole purpose of quantifying the inflation it induces relative to the temporal design; the magnitude of that inflation is itself a reportable quantity and bears directly on how the existing benchmarking literature should be read (Bergmeir & Benitez, 2012).

Selective labels are addressed where a rejected-application file exists, by a two-stage sample-selection correction whose exclusion restriction is stated and defended explicitly. Where booked-loan universes provide no such file, results are reported as conditional on the historical approval policy and are not extrapolated (Banasik & Crook, 2007). This limitation is structural to the public-data ecosystem and is not remediable within the protocol.

6.5 Baselines

Twelve baselines across four families are specified. The regulatory-incumbent family comprises a weight-of-evidence logistic scorecard and an elastic-net logistic regression. The tree-ensemble family comprises random forest, unconstrained gradient boosting under two widely used implementations, a third boosting implementation with ordered target statistics, and monotone-constrained gradient boosting. The interpretable-by-design family comprises explainable boosting machines, neural additive models (Agarwal et al., 2021), and penalized logistic tree regression (Dumitrescu et al., 2022). The deep tabular family comprises an attention-based tabular architecture and a transformer variant for tabular data. Fairness-processing comparators applied to the unconstrained tree baseline comprise reweighing (Kamiran & Calders, 2012), exponentiated-gradient reduction (Agarwal et al., 2018), and threshold post-processing (Hardt et al., 2016).

6.6 Metrics

Discrimination is measured by area under the receiver operating characteristic curve, the Kolmogorov-Smirnov statistic, and area under the precision-recall curve. Calibration is measured by Brier score, expected calibration error over equal-mass bins, and calibration slope. Economic performance is measured by expected portfolio profit under a stated pricing function and by approval rate at a fixed loss budget. Disparity is measured by approval-rate ratio, statistical parity difference, equalized-odds difference, and standardized mean difference of scores; where the protected attribute is imputed, all intervals must propagate imputation uncertainty by Rubin's rules over multiply-imputed draws, since point estimates on imputed attributes are generally biased and nominal intervals

anticonservative (Chen et al., 2019; Rubin, 1987).

6.7 Statistical Analysis Plan

Paired differences in area under the curve on a common test set are assessed by the DeLong test (DeLong et al., 1988). Multi-corpus comparisons use the corrected resampled t-test to account for train-test overlap dependence (Nadeau & Bengio, 2003). Family-wise error is controlled by the Holm procedure across a pre-registered comparison family that is fixed in advance and not expanded post hoc. Effect sizes with confidence intervals are reported alongside every p-value, and the primary inferential emphasis is on effect magnitude.

6.8 Threats to the Protocol

The protocol has known weaknesses that its executors should not paper over. Public credit corpora are booked-loan universes or curated competition datasets and are not representative of any institution's application flow. Protected-attribute imputation validated on mortgage data need not transfer to non-mortgage populations with different surname and geography joint distributions (Elliott et al., 2009). The monotonicity directions supplied to the constrained models are domain judgments, and a misspecified direction imposes a real and undetected cost. Finally, a protocol executed by the framework's own proponents carries an obvious risk of favourable analytic choices; the pre-registration of propositions, disconfirmation criteria, and comparison families is intended to constrain that risk, and independent replication would constrain it further.

7. Discussion

7.1 Why Regulatory Retrenchment Raises the Value of Self-Evidencing Models

The natural reading of the 2025 and 2026 developments is that the compliance burden on United States credit modelers has fallen, and that the case for elaborate governance machinery has weakened with it. This reading is mistaken, for three reasons. First, the binding obligations that most constrain model architecture were not relaxed: the duty to state specific principal reasons is statutory and untouched, proxy use remains actionable under a disparate-treatment theory, and effects-based exposure survives in housing credit and under state law. What was withdrawn was interpretation, and what was revised was guidance that now expressly disclaims enforceability.

Second, the removal of external specification shifts the specification burden inward. When guidance enumerated validation activities, an institution could discharge much of its obligation by conformance to the enumeration. When guidance states that appropriate practice varies with the institution and declines to prescribe, the institution must decide what adequate validation means for its own model and must be able to defend that decision. The artifact that defends it is a record of the choice, its rationale, and its execution, which is exactly what Layers 3 and 5 produce as a by-product of ordinary operation rather than as a special project. Third, the audiences have multiplied even as the federal supervisor has stepped back: state regulators, state artificial intelligence statutes, Fair Housing Act plaintiffs, and private litigants are not bound by federal guidance withdrawal, and their information demands are adversarial rather than supervisory in character. An adversarial audience does not accept assertions; it tests them.

The general principle, which we offer as the paper's central conceptual claim, is that the value of self-generated, independently verifiable evidence varies inversely with the specificity of external regulation. It is highest precisely where a regulator has declined to specify and where liability nonetheless remains.

7.2 Anticipated Objections

Four objections merit direct response. That the framework optimizes against metrics and no metric is a safe harbour is correct, and we do not claim otherwise; an approval-rate ratio is a coarse, threshold-dependent statistic, and a stable reason may still be misleading. The framework's defensible claim is narrower: it makes the choice of metric explicit, records it as a governed decision, and preserves the evidence needed to revisit it when challenged. That publishing a rejected-candidate frontier creates discoverable material is also true; but an institution that searched alternatives and recorded the search is in a materially stronger position than one that cannot show it searched at all, and we regard the deterrent effect on deploying a knowingly more disparate design as a feature rather than a defect. That monotonicity is a domain assumption dressed up as a guarantee is partly true, since directions are supplied rather than learned and a wrong direction costs accuracy silently; what monotonicity buys is legibility and local stability rather than correctness, and proposition P4 exists precisely to test whether the second effect is real. That the stability proposition assumes what it needs is likewise partly true, and the qualifications following Proposition 1 state it: the contribution is the form of the control, a per-decision margin test with an explicit exception path, rather than the tightness of the constant.

7.3 Managerial and Implementation Implications

Three organizational rather than technical frictions should be anticipated. The first concerns governance sequencing: the alternative-model search requires the model-risk and fair-lending functions to declare a disparity criterion and a

utility tolerance in advance of the search rather than after seeing its results, and the impossibility results guarantee that this choice cannot be delegated to the algorithm. The second concerns legal posture: writing a rejected-candidate frontier to a durable record will meet resistance on litigation-exposure grounds and should be discussed with counsel at design time, since the realistic alternative is not an absent record but a partial, reconstructed, and internally inconsistent one. The third concerns operating model and staffing: the manual-adjudication path for low-margin decisions requires staffing forecast from the certified-coverage rate, which is drift-sensitive and therefore belongs in the monitoring suite and in a rolling capacity forecast rather than in an annual plan.

A fourth implication is strategic rather than operational. Because the framework's artifacts are produced continuously, an institution that adopts it acquires the capacity to answer regulatory and litigation inquiries on a timescale that competitors relying on retrospective reconstruction cannot match. In a period when supervisory posture is shifting and private enforcement is comparatively more salient, that responsiveness is a defensible source of advantage rather than a pure cost centre, and it should be evaluated on that basis in the business case.

8. Limitations

The framework is unvalidated. No component of ACRIS has been empirically evaluated in this work, and no claim of validation is made. The propositions of Section 6 are hypotheses. It is entirely possible that the joint constraint penalty proves large, that conditional screening offers no advantage over marginal screening, or that the margin gate rejects an operationally intolerable share of decisions. Establishing which of these obtains is the purpose of the protocol, and readers should treat the framework accordingly.

The selective-labels problem is not solved. Public credit corpora observe outcomes only for approved applicants. No claim about counterfactual approvals can be supported without an exclusion restriction whose validity is untestable, and this bounds every performance claim the protocol can produce.

Jurisdictional scope is deliberately narrow. The framework is specified against United States federal consumer-credit law as of mid-2026. It does not address the European Union artificial intelligence regulation, data-protection provisions on automated decision-making, commercial-credit contexts, or the full range of state requirements. Mapping statutory text to computable predicates involves interpretive choices that counsel must ratify, and nothing in this paper is legal advice.

Finally, the regulatory baseline is itself volatile. The 2026 Regulation B amendment states an agency position on statutory interpretation that may be tested in litigation. The framework is deliberately specified so that its controls retain force under either reading, but readers should not assume that the baseline described in Section 3 is stable, and any institution adopting the framework should treat that baseline as a versioned assumption subject to periodic review.

9. Conclusion and Future Research

This paper has argued that the perceived conflict between predictive performance and regulatory auditability in United States consumer credit is substantially an artifact of sequential, post-hoc compliance engineering rather than an intrinsic statistical trade-off, and that the regulatory realignment of 2025 and 2026 strengthens rather than weakens the case for models that generate their own verifiable evidence. Where external specification recedes while statutory duties and private exposure remain, the institution's position rests on the quality of the record it can produce about its own choices.

The paper developed ACRIS, a five-layer framework in which feature admissibility is adjudicated by conditional rather than marginal dependence, the predictive core is shape constrained, the alternative-model search is recorded in full including rejected candidates and their rejection rationales, principal reasons are anchored to the approval boundary and gated on a per-decision stability margin, and every governance artifact is bound into a tamper-evident ledger that a third party can check without trusting its producer. A finite-sample sufficient condition for top-k reason-set preservation was proved, with its assumptions and limits stated candidly, including the cases in which it does not apply. Because the framework is a proposal rather than a validated system, the paper specified in advance the propositions that would test it and the observations that would refute them.

Five directions for future research follow. The first is causal admissibility: replacing conditional-mutual-information screening with identified causal criteria under an explicit structural model, which would support counterfactual rather than associational claims about proxy use and would strengthen the evidentiary value of the admissibility certificate. The second is privacy-preserving governance: extending the ledger and the admissibility screen to cross-institutional consortium models, where proxy screening must operate without pooling protected attributes across participants. The third is selective-label correction at scale: integrating principled exploration policies capable of identifying counterfactual approval outcomes at acceptable expected loss, which would relax the most binding constraint on empirical work in this area. The fourth is jurisdictional generalization: re-specifying the predicate set for emerging

European and state artificial intelligence regimes and characterizing whether the predicates compose or conflict. The fifth is shape-constrained tabular foundation models: determining whether in-context tabular learners admit shape constraints and per-decision stability gating at all, which is presently an open architectural question with direct bearing on whether the framework survives the next generation of modelling technology.

The authors intend to release the admissibility screen, the margin test, and the ledger schema as an open reference implementation, so that the protocol of Section 6 can be executed independently of its proponents. Independent replication is the appropriate standard for a framework of this kind, and the design of the protocol has been shaped throughout by that expectation.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

ORCID iD: Hasibur Rahman — [ORCID iD, if available]; Sarder Abdulla Al Shiam — [ORCID iD, if available]; Md Sibbir Hossain — [ORCID iD, if available].

Data Availability Statement: No new data were generated or analysed in this study. The public corpora identified in the evaluation protocol are available from their respective providers under their published licences.

Author Contributions: All authors contributed to the conceptualisation of the framework, the regulatory analysis, and the drafting and critical revision of the manuscript. All authors have read and agreed to the submitted version.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

1. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. In Proceedings of the 35th International Conference on Machine Learning (pp. 60–69). PMLR.
2. Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., & Hinton, G. E. (2021). Neural additive models: Interpretable machine learning with neural nets. In Advances in Neural Information Processing Systems 34 (pp. 4699–4711). Curran Associates.
3. Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. In Proceedings of the ICML Workshop on Human Interpretability in Machine Learning. arXiv. https://arxiv.org/abs/1806.08049
4. Banasik, J., & Crook, J. (2007). Reject inference, augmentation, and sample selection. European Journal of Operational Research, 183(3), 1582–1594. https://doi.org/10.1016/j.ejor.2006.06.072
5. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671–732.
6. Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics, 143(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047
7. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289–300.
8. Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192–213. https://doi.org/10.1016/j.ins.2011.12.028
9. Board of Governors of the Federal Reserve System. (2026). Revised guidance on model risk management (SR Letter 26-2). https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm
10. Board of Governors of the Federal Reserve System, & Office of the Comptroller of the Currency. (2011). Supervisory guidance on model risk management (SR Letter 11-7 / OCC Bulletin 2011-12). https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm
11. Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 339–348). ACM. https://doi.org/10.1145/3287560.3287594
12. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163. https://doi.org/10.1089/big.2016.0047
13. Consumer Financial Protection Bureau. (2025). Interpretive rules, policy statements, and advisory opinions: Withdrawal. Federal Register. https://www.federalregister.gov/documents/2025/05/12/2025-08286
14. Consumer Financial Protection Bureau. (2026). Equal Credit Opportunity Act (Regulation B): Final rule. 91 Fed. Reg. 21620 (April 22, 2026), RIN 3170-AB54. https://www.federalregister.gov/documents/2026/04/22/2026-07804
15. Crosby, S. A., & Wallach, D. S. (2009). Efficient data structures for tamper-evident logging. In Proceedings of the 18th USENIX Security Symposium (pp. 317–334). USENIX Association.

16. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44(3), 837–845. https://doi.org/10.2307/2531595
17. Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research, 297(3), 1178–1192. https://doi.org/10.1016/j.ejor.2021.06.053
18. Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., & Lurie, N. (2009). Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. Health Services and Outcomes Research Methodology, 9(2), 69–83. https://doi.org/10.1007/s10742-009-0047-1
19. Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research, 20(177), 1–81.
20. Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. The Journal of Finance, 77(1), 5–47. https://doi.org/10.1111/jofi.13090
21. Gunnarsson, B. R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M., & Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? European Journal of Operational Research, 295(1), 292–305. https://doi.org/10.1016/j.ejor.2021.03.006
22. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems 29 (pp. 3315–3323). Curran Associates.
23. Hurlin, C., Pérignon, C., & Saurin, S. (2024). The fairness of credit scoring models. Management Science. Advance online publication. https://doi.org/10.1287/mnsc.2022.03888
24. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8
25. Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2023). A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. ACM Computing Surveys, 55(5), Article 95. https://doi.org/10.1145/3527848
26. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Proceedings of the 8th Innovations in Theoretical Computer Science Conference (Article 43). Schloss Dagstuhl.
27. Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. European Journal of Operational Research, 297(3), 1083–1094. https://doi.org/10.1016/j.ejor.2021.06.023
28. Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. Physical Review E, 69(6), Article 066138. https://doi.org/10.1103/PhysRevE.69.066138
29. Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv. https://arxiv.org/abs/2202.01602
30. Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. In Proceedings of the 37th International Conference on Machine Learning (pp. 5491–5500). PMLR.
31. Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030
32. Liu, X., Han, X., Zhang, N., & Liu, Q. (2020). Certified monotonic neural networks. In Advances in Neural Information Processing Systems 33 (pp. 15427–15438). Curran Associates.
33. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9
34. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765–4774). Curran Associates.
35. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). ACM. https://doi.org/10.1145/3287560.3287596
36. Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. Machine Learning, 52(3), 239–281. https://doi.org/10.1023/A:1024068626366
37. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1
38. Office of the Comptroller of the Currency. (2026). Model risk management: Revised guidance (OCC Bulletin 2026-13). https://www.occ.gov/news-issuances/bulletins/2026/bulletin-2026-13.html

39. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778
40. Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. Wiley. https://doi.org/10.1002/9780470316696
41. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
42. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys, 16, 1–85. https://doi.org/10.1214/21-SS133
43. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 180–186). ACM. https://doi.org/10.1145/3375627.3375830
44. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.
45. Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. ACM Transactions on Information Systems, 28(4), Article 20. https://doi.org/10.1145/1852102.1852106