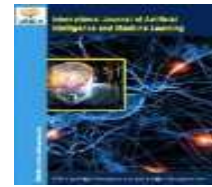




International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Governance-Integrated Architecture for Trustworthy AI in Financial Systems

Ratanchur Sarkar

AI Independent Researcher, USA, Senior Member, IEEE. E-mail: ratanchursarkar@gmail.com, ORCID: 0009-0009-2287-7464

Abstract

AI is an enterprise-wide capability in banks, applied to credit risk analysis, fraud detection, algorithmic trading, and regulatory compliance. For AI to be fully realized in regulated financial services, properties beyond predictive performance must be demonstrated: transparency, fairness, auditability, and robustness at every stage of the system lifecycle. Although the scientific literature has produced advances in AI model design and governance frameworks, these remain largely isolated, leaving practitioners without an integrated architecture for deploying AI systems to legal standards. We present a governance-integrated reference architecture for trustworthy AI in financial services, organized around four interrelated layers: (1) data governance and provenance; (2) model development and validation; (3) runtime monitoring and dynamic enforcement; and (4) enterprise AI governance and systemic-risk oversight. The architecture embeds governance controls into the AI lifecycle infrastructure and treats regulatory compliance as an architectural property rather than a post-hoc audit. We evaluate the architecture on the public Home Credit Default Risk dataset (307,511 applications), instantiating all four layers and measuring each control against conditions that arise naturally in the data. The data-quality gate exposes genuine missingness and sentinel-encoding defects; a fairness control raises the Disparate Impact Ratio for two model families at negligible accuracy cost; Population Stability Index monitoring detects a real distribution shift and triggers enforcement; and the enterprise layer surfaces correlated cross-model drift through a shared upstream dependency, raising a single systemic-risk flag. The architecture constitutes a lifecycle-oriented, technology-agnostic blueprint aligned with model risk management guidance, BCBS 239, the NIST AI Risk Management Framework, ISO/IEC 42001, and the EU AI Act.

Keyword: trustworthy AI, AI governance, financial systems, data provenance, model validation, runtime monitoring, explainability, fairness, regulatory compliance, enterprise risk.

I. INTRODUCTION

The financial sector has long operated at the intersection of data intensity and regulatory scrutiny. Recent AI applications—including credit underwriting, transaction fraud detection, anti-money laundering (AML) surveillance, algorithmic trading, and supervisory reporting—have amplified both the promise and the risks of automated decision-making at scale [1]. A 2019 global survey of 2,555 executives and managers across 29 industry sectors, conducted by MIT Sloan Management Review and the Boston Consulting Group, found that 90% perceived AI as a business opportunity for their organization, while 45% identified it as a risk—up from 37% in 2017 [13]. In regulated financial ecosystems, explainability, auditability, and behavioral governance matter more than raw predictive performance.

Regulatory expectations for financial AI are correspondingly specific. Model risk management guidance from the Office of the Comptroller of the Currency (OCC) and the Federal Reserve requires institutions to validate model logic independently, document assumptions, and monitor production models against defined benchmarks [6]. BCBS 239 mandates attribute-level data lineage for all material risk data aggregation and reporting [8]. The EU AI Act, which entered into force in August 2024 with obligations for high-risk systems applying from August 2026, lists credit scoring, fraud detection, and AML systems in Annex III under the highest risk category, imposing requirements on data transparency, human oversight, data governance, and AI robustness [7]. These frameworks make clear that governance must be built into AI architectures from the outset, not bolted on afterward.

Research has not kept pace with industry practice. A study of over 500 Microsoft ML practitioners identified data management, model customization, and AI component integration as the top engineering challenges—none well addressed by standard software engineering practices [2]. Most financial AI systems have not been evaluated against the 28-test Production Readiness Rubric [6], leaving a substantial governance gap. Existing AI governance frameworks address principles such as fairness, accountability, and transparency [3], [11], and

management standards such as the NIST AI RMF [22] and ISO/IEC 42001 [23] define what to govern—but not how to operationalize those principles in regulated financial systems.

This paper fills that gap by framing trustworthy AI in financial systems as a systems architecture problem. It makes three contributions: (1) identifying the principal technical and governance challenges for AI in regulated financial services; (2) defining trustworthy AI as a lifecycle-spanning architectural problem requiring integrated controls across data, model, operational, and enterprise layers; and (3) presenting a governance-integrated four-layer reference architecture that embeds validation, monitoring, and oversight directly into the AI infrastructure. Section II surveys related work; Section III describes the research methodology; Section IV characterizes deployment challenges; Section V presents the architecture; Section VI reports the empirical evaluation; Sections VII–IX address discussion, limitations, and future work; and Section X concludes.

II. RELATED WORK

A. Data Platforms and Lakehouse Architectures

Enterprise data infrastructure has converged on unified lakehouse architectures that combine cloud object storage economics with data warehouse capabilities—including ACID transactions, schema enforcement, and schema versioning [5]. Comparative studies confirm their scalability advantages for heterogeneous, high-volume workloads [9]. However, these contributions do not address the governance controls—lineage, auditability, and policy enforcement—that financial institutions must implement before AI systems can reliably meet regulatory compliance requirements.

B. Machine Learning Systems and MLOps

Operating production ML systems involves three pervasive classes of hidden technical debt: data dependencies, feedback loops, and distributional shift—all degrading model performance in ways that standard engineering testing cannot easily detect [1]. MLOps practitioners cite data management complexity, model customization, and AI component integration as leading challenges [2]. Existing MLOps frameworks, including the 28-Test Production Readiness Rubric [6], operationalize structured validation and monitoring processes but largely do not address the regulatory accountability obligations specific to financial AI.

C. Trustworthy AI: Fairness, Explainability, and Governance

Algorithmic bias arising from historical data distributions, feature proxies, and sampling bias can harm affected populations and expose organizations to regulatory risk [3], [15]. More than 23 distinct types of bias have been catalogued in the ML literature [12]. Explanation methods such as SHAP and LIME produce interpretable outputs suitable for auditors and regulators [4], [20]. Model cards support post-deployment accountability [11]. Fairness-aware training approaches achieve performance comparable to unconstrained baselines [10], [16]. Cross-disciplinary ethical frameworks address accountability, social value, and institutional responsibility [14], [17], [19]. Individually and collectively, however, these contributions have not produced an integrated architecture for regulated financial deployment.

D. AI Governance and Management Frameworks

The NIST AI Risk Management Framework structures practice around Govern, Map, Measure, and Manage [22], while ISO/IEC 42001 specifies requirements for an AI management system [23]. Both set valuable shared expectations, but they are deliberately domain-agnostic: they describe what to govern, not how those obligations become enforceable controls inside a financial pipeline subject to BCBS 239 and EU AI Act Annex III. The architecture presented here is complementary, supplying the technical control plane through which such frameworks can be operationalized.

E. The Gap

These bodies of work have produced valuable advances in data platforms, ML engineering, and AI governance, yet the contributions remain largely siloed, as summarized in Table I. To the best of our knowledge, no prior contribution spans attribute-level data lineage, structured pre-deployment validation, quantitative runtime-drift enforcement, and portfolio-level systemic-risk oversight within a single lifecycle model intended for regulated finance. It is precisely this gap—between individually useful techniques and an integrated architecture that operationalizes them end to end—that the proposed architecture is designed to fill.

TABLE I. Coverage of trustworthy-AI capabilities across representative prior work.

Capability	Data platforms	MLOps	Fairness / XAI	Governance fwks	This work
Attribute-level data lineage	Partial	No	No	Process-level	Yes
Structured pre-deployment validation	No	Yes	Partial	Process-level	Yes

Capability	Data platforms	MLOps	Fairness / XAI	Governance fwks	This work
Quantitative runtime-drift enforcement	No	Partial	No	No	Yes
Portfolio / systemic-risk oversight	No	No	No	Partial	Yes
Regulatory mapping (BCBS 239 / EU AI Act)	No	No	Partial	Partial	Yes

III. RESEARCH METHODOLOGY AND DESIGN APPROACH

We adopt an architecture-oriented, design-science approach in which the primary contribution is a governance-integrated architecture as a systems-level artifact, rather than a new learning algorithm. Governance, validation, monitoring, and risk-management requirements are drawn from BCBS 239 [8], the EU AI Act [7], model risk management guidance [6], and the AI governance literature [22], [23], and mapped onto a four-layer architecture so that governance is embedded throughout the system lifecycle rather than applied post-hoc.

We evaluate the architecture in two complementary ways. First, an architectural-coverage analysis (Table I) positions the design against prior work. Second, a reproducible empirical evaluation (Section VI) instantiates all four layers on the public Home Credit Default Risk dataset [27], so that each control responds to conditions that arise naturally in the data rather than being artificially injected. The architecture is technology-agnostic, with controls mapping directly onto managed cloud services such as SageMaker Clarify and Model Monitor [28]; all code, seeds, and commands are released for independent verification.

IV. CHALLENGES OF DEPLOYING AI IN REGULATED FINANCIAL SYSTEMS

A. Data Quality, Lineage, and Governance

Financial AI systems consume heterogeneous data sources—transaction, customer, market, device, third-party, and behavioral data—that differ in structure, quality, sensitivity, and terms of use. Inadequate data governance produces inaccurate features, biased predictions, and potential regulatory breaches. BCBS 239 requires completeness, accuracy, and attribute-level traceability in all material risk data flows [8], making such traceability an architectural necessity, not an operational luxury.

B. Explainability, Fairness, and Model Validation

AI models for credit scoring, fraud detection, AML classification, and risk classification are designated high-risk under Annex III of the EU AI Act, with compliance obligations applying from August 2026 [7]. The EEOC four-fifths rule defines a quantitative fairness floor: the Disparate Impact Ratio (DIR) must be at least 0.80. Systematic pre-deployment testing against this rule is a governance requirement, and the interpretability-accuracy trade-off must be structurally resolved through mandatory explainability controls [4], [20].

C. Runtime Drift and Operational Risk

Fraud patterns, macroeconomic conditions, and regulatory requirements change continuously. The Population Stability Index (PSI) quantifies distributional shift: $PSI < 0.10$ indicates no meaningful change; $0.10-0.20$ warrants investigation; $PSI > 0.20$ signals a material change requiring immediate remediation. Mitigating runtime risk requires both monitoring for these signals and pre-authorized countermeasures.

D. Enterprise Oversight and Systemic AI Risk

Multiple production models may share upstream data sources, feature pipelines, or third-party vendor dependencies. Degradation in shared components propagates into downstream systems in ways that model-level monitoring alone cannot detect. Sculley et al. identified such systemic dependency failures as among the costliest and least observable forms of technical debt in deployed AI systems [1]. Managing this exposure requires portfolio-level governance: model inventories, risk-based classification, dependency mapping, and aggregated signal analysis.

V. PROPOSED ARCHITECTURE: GOVERNANCE-INTEGRATED FOUR-LAYER FRAMEWORK

The proposed framework comprises four layers spanning the AI lifecycle (Fig. 1), governed by a principle of governance-by-design: controls are embedded into system components rather than applied as a separate review activity. Each layer contextualizes, informs, and constrains adjacent layers, while end-to-end traceability provides a complete audit trail from raw data through model output to compliance reporting. The framework is technology-agnostic: it specifies the properties a solution must satisfy, not the particular tools used at each layer.

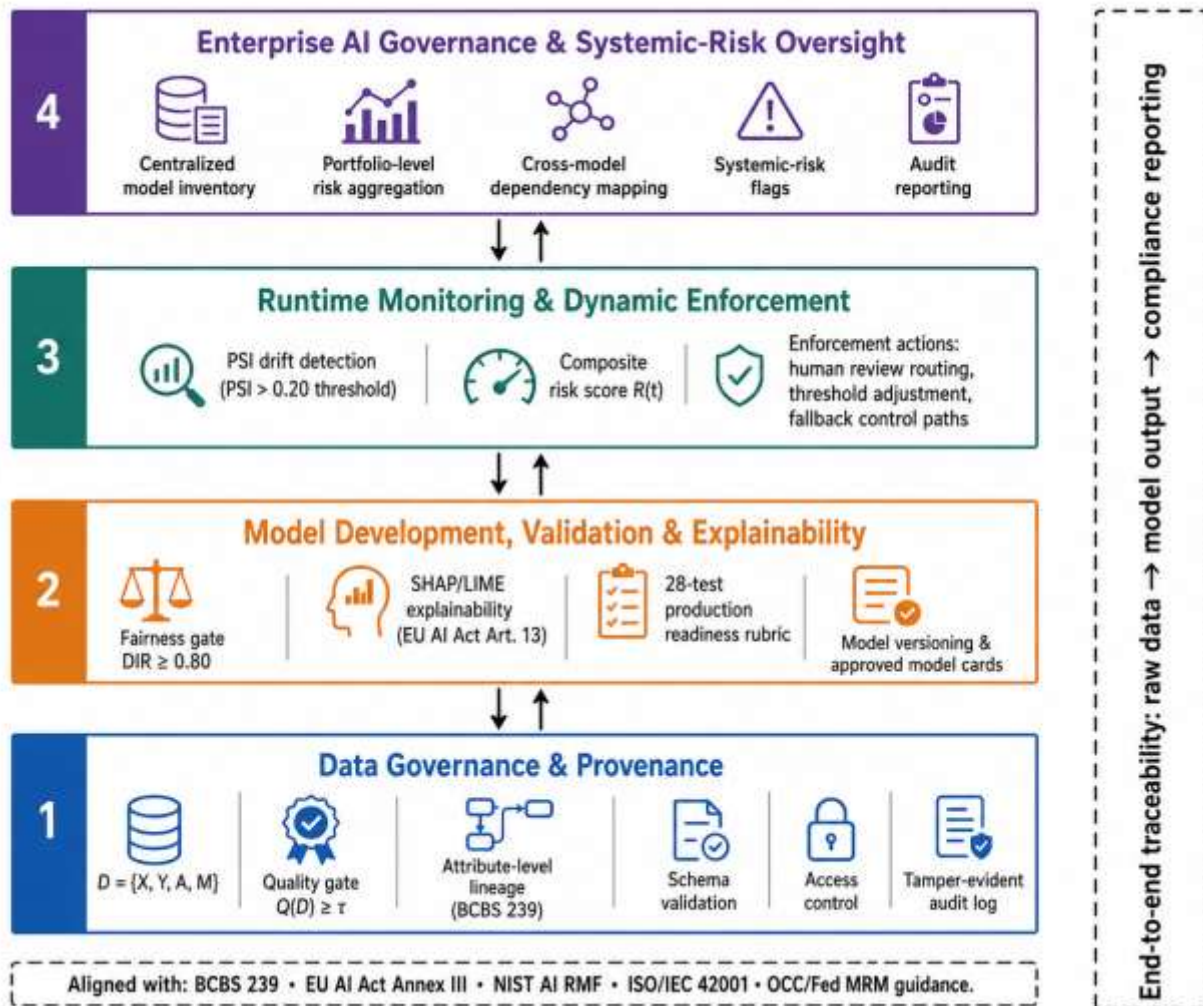


Fig. 1. Governance-integrated four-layer architecture spanning the AI lifecycle, with governance-by-design constraints applied within and across layers.

TABLE II. Architecture layers, functions, and governance controls.

Layer	Primary function	Key governance controls
Data Governance & Provenance	Govern data integrity, lineage, and policy	Schema validation, lineage, access control, tamper-evident audit, quality gate $Q(D)$
Model Development & Validation	Govern construction and pre-deployment approval	Fairness (DIR), explainability (SHAP), 28-test readiness, model cards and versioning
Runtime Monitoring & Enforcement	Govern behavior after deployment	Drift (PSI), composite risk $R(t)$, enforcement actions
Enterprise AI Governance	Govern the portfolio and systemic risk	Model inventory, risk classification, signal aggregation, audit reporting

A. Layer 1: Data Governance and Provenance

The Data Governance and Provenance layer provides the common data-access and control-plane components used by all other layers. A governed dataset version $D = \{X, Y, A, M\}$ captures the feature vector X , target variable Y , protected attributes A , and provenance metadata M . Provenance metadata M records the source identifier, event-time and ingestion-time timestamps, consent and compliance flags, lineage identifiers, and the full transformation path, ensuring that the origin of every value can be traced—in line with BCBS 239 [8] and EU AI Act data-governance obligations [7].

Data quality is assessed through a quality function $Q(D)$, evaluated over completeness, consistency, accuracy, and timeliness. Only datasets for which $Q(D) \geq \tau$ are admitted into the model pipeline. The layer implements

cataloguing, schema validation, lineage tracking, metadata management, access control, and tamper-evident audit logging, ensuring that data entering the AI lifecycle is trusted, policy-compliant, and fully documented.

B. Layer 2: Model Development, Validation, and Explainability

This layer governs how AI models are created, validated, and approved for production. Model training is framed as a constrained optimization problem in which predictive accuracy is maximized subject to fairness, interpretability, and regulatory constraints. Fairness is quantified via the Disparate Impact Ratio: $DIR = P(\text{positive outcome} \mid \text{unprivileged group}) / P(\text{positive outcome} \mid \text{privileged group})$. A DIR of at least 0.80 satisfies the EEOC four-fifths rule and serves as the quantitative acceptance criterion for deployment [12], [15]. Equalized-odds constraints further require equal true-positive and false-positive rates across groups [3], [25]. Explainability methods—SHAP and LIME—produce reviewable outputs consistent with the transparency obligations of Article 13 of the EU AI Act [4], [7], [20]. Model validation follows the 28-test production-readiness rubric [6], spanning data-validation, model-development, ML-infrastructure, and monitoring tests. Model lifecycle management records versions, training and validation datasets, approval status, and assumptions, providing a complete audit trail [11]. Validation criteria are summarized in Table III.

TABLE III. Model-validation criteria, methods, and governance outputs.

Validation criterion	Assessment method	Governance output	Regulatory basis
Predictive performance	Hold-out, cross-validation	Validation report	MRM guidelines [6]
Temporal stability	Time-series backtesting	Stability certificate	BCBS 239 [8]
Fairness (DIR $\geq$ 0.80)	Four-fifths rule, equalized odds	Fairness documentation	ECOA/Reg B [26]; EU AI Act [7]
Explainability	Local attribution (SHAP/LIME)	Interpretability report	EU AI Act Art. 13 [7]
Documentation	28-test readiness rubric [6]	Approved model card [11]	MRM guidelines [6]

C. Layer 3: Runtime Monitoring and Dynamic Enforcement

This layer governs model behavior in production. Deployed models are monitored across input-feature distributions, prediction-score distributions, error rates, latency, and exceptions. Distributional shift is quantified through the Population Stability Index: $PSI = \sum [(\text{Actual}\% - \text{Expected}\%) \cdot \ln(\text{Actual}\% / \text{Expected}\%)]$. A PSI above 0.20 mandates remediation; between 0.10 and 0.20 warrants investigation. Concept-drift monitoring tracks changes in the feature–target relationship in parallel [1].

Individual monitoring signals are combined into a composite runtime risk score: $R(t) = w_1 \cdot PSI(t) + w_2 \cdot \text{ErrorRate}(t) + w_3 \cdot \text{FairnessViolation}(t) + w_4 \cdot \text{LatencyBreach}(t)$, where weights (set to 0.4, 0.3, 0.2, and 0.1 in our evaluation) are tuned to the institution’s risk appetite. When this score or any individual metric crosses its threshold, pre-authorized enforcement actions are applied—including routing affected cases to human review, dynamically adjusting the prediction threshold, falling back to rule-based control paths, or restricting automation under high-risk conditions (Table IV). Runtime signals also feed back into adjacent layers, closing the governance loop across the lifecycle.

TABLE IV. Runtime monitoring signals, detection methods, enforcement responses, and thresholds.

Signal type	Detection method	Enforcement response	Threshold
Feature distribution drift	Population Stability Index (PSI)	Alert + threshold review	PSI > 0.20 (critical)
Prediction score shift	KL divergence, statistical comparison	Model recalibration trigger	$p < 0.05$
Performance degradation	Error rate vs. validation baseline	Manual review routing	Degradation > 10% relative
Processing latency anomaly	SLA monitoring (p99 latency)	Fallback model invocation	Breach of SLA bound
Fairness violation	Continuous DIR tracking	Automated decision suspension	DIR < 0.80 for any subgroup

D. Layer 4: Enterprise AI Governance and Systemic Risk

This layer provides institution-level oversight across the full AI portfolio through a centralized model inventory that tracks each deployed model's purpose, owner, target data, validation status, risk classification, and deployment context. Risk classification embeds governance requirements proportional to business impact, regulatory risk exposure, model complexity, data sensitivity, and EU AI Act Annex III tier [7].

Outputs from the three underlying layers are aggregated to yield a portfolio-wide view of AI risk, including systemic risks such as correlated PSI exceedances across models sharing a common upstream data dependency [1]. Governance processes encompass model approval and deployment, periodic revalidation, incident management, and audit reporting linking model outputs, governed datasets, validation artifacts, and runtime records into a complete audit trail aligned with BCBS 239 [8] and the EU AI Act [7].

VI. EMPIRICAL EVALUATION ON PUBLIC CREDIT DATA

We instantiate all four layers on the Home Credit Default Risk dataset [27], comprising 307,511 loan applications with 122 attributes and a binary default target. The dataset contains a legally protected attribute (applicant sex), substantial genuine data-quality defects, and real group disparity. Crucially, the conditions each control detects arise naturally in the data—not artificially injected—strengthening the evidential value relative to a synthetic proof-of-concept. We use a 70/30 stratified split, seed all randomness, and report results for regularized logistic regression and gradient boosting.

A. Layer 1: Data-Quality Admission Gate

Multiple attributes are 50–70% empty (e.g., COMMONAREA_MODE), and the DAYS_EMPLOYED field encodes an invalid sentinel value of 365,243 for approximately 18% of records. Applying worst-feature aggregation, the quality function $Q(D)$ scores the raw feed at 0.534, well below the admission threshold $\tau = 0.95$, and the Layer-1 gate blocks the dataset. After governed remediation (sentinel conversion, median imputation, out-of-range clipping), the score rises to 0.999 and the data is admitted. The gate enforces the completeness and traceability obligations of BCBS 239 on real data before any model consumes it.

B. Layer 2: Fairness Gate and Explainability

We adopt the perspective of a lender approving only the lowest-risk 50% of applicants, excluding the protected attribute from model inputs so that measured disparity flows solely through proxy features. Baseline models achieve AUC values of 0.746 (logistic regression) and 0.753 (gradient boosting), with DIRs of 0.822 and 0.810 respectively—both above the 0.80 legal floor, reflecting a mild but genuine disparity (approval rates of 0.437 versus 0.532).

Applying Kamiran–Calders reweighing [24] to reach an institutional internal target of 0.90, the selected model attains a DIR of 0.912 at an AUC of 0.745—an accuracy cost of approximately 0.1 percentage point (Figs. 2–3 and Table VI). Local SHAP attributions provide Article-13-consistent explanations. Reporting results for both a linear and a nonlinear family ensures the finding does not depend on a single estimator.

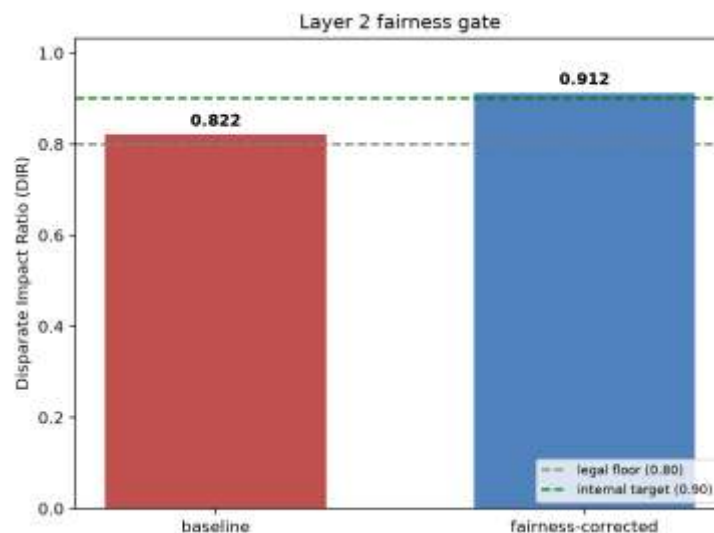


Fig. 2. Layer-2 fairness gate: DIR before (0.822) and after reweighing (0.912), against the 0.80 legal floor and 0.90 internal target.

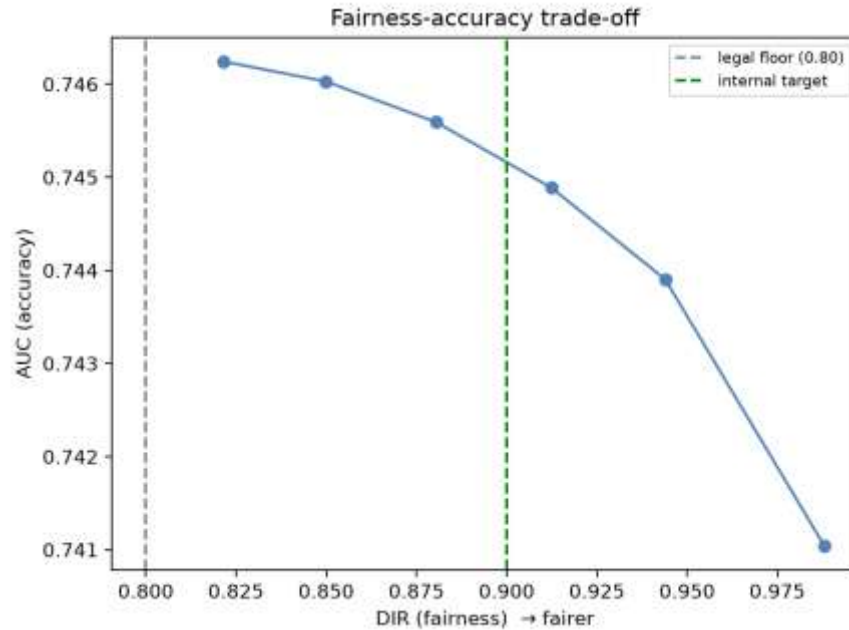


Fig. 3. Layer-2 fairness-accuracy trade-off (efficient frontier): AUC declines only marginally as DIR rises toward the internal target.

C. Layer 3: Drift Detection and Enforcement

We construct a 12-week deployment stream exhibiting a real, progressive covariate shift assembled from real applicants to emulate a macroeconomic income shock. The PSI remains below 0.01 through week 4, enters the investigation band at week 5 (PSI = 0.115), and exceeds the 0.20 critical threshold at week 7 (PSI = 0.211), triggering enforcement: affected cases are routed to human review and recalibration is initiated. The composite risk score $R(t)$ crosses its 0.20 enforcement threshold at week 11 (Figs. 4–5).

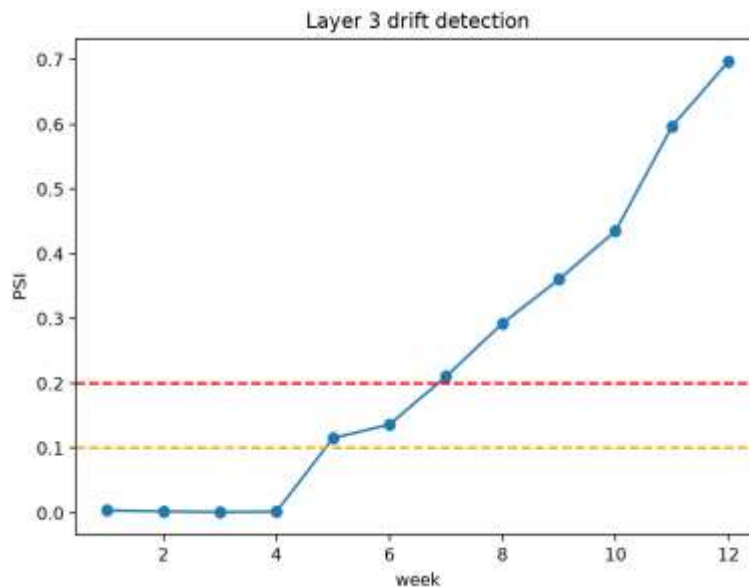


Fig. 4. Layer-3 drift detection: PSI by week against investigation (0.10) and critical (0.20) thresholds; enforcement triggers at week 7.

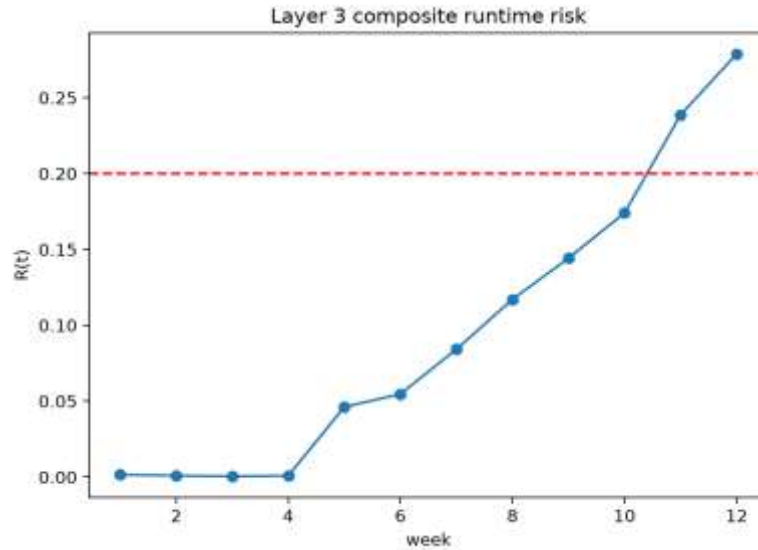


Fig. 5. Layer-3 composite runtime risk $R(t)$, a weighted blend of drift, error, fairness, and latency signals; crosses the enforcement threshold as drift sets in.

D. Layer 4: Portfolio Systemic Risk

We model a portfolio of three services—credit scoring, fraud detection, and AML monitoring—each monitoring a different but correlated upstream feature (income, credit amount, and annuity) fed by the same economic shock. Per-model PSI series are strongly correlated (cross-model correlation = 0.999), and all three breach the 0.20 threshold within weeks 6–7 (Fig. 6). The enterprise layer raises a single systemic-risk flag—an exposure that would be invisible to individual model monitoring and otherwise addressed as three unrelated incidents.

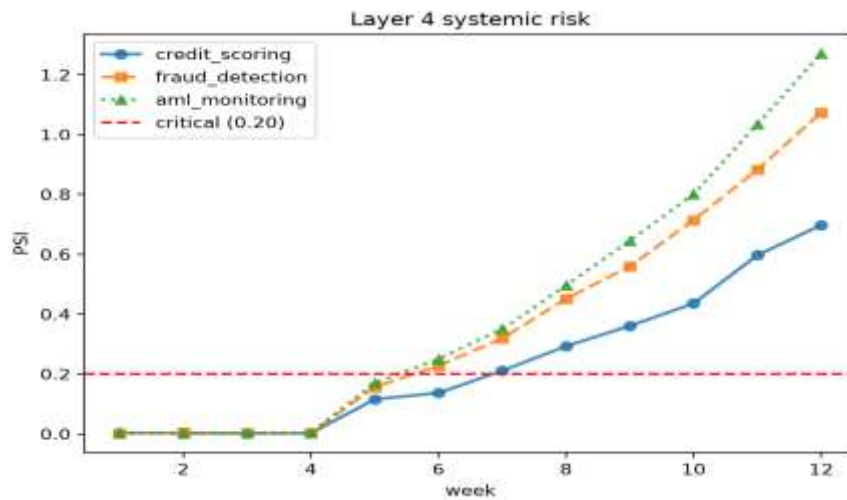


Fig. 6. Layer-4 systemic risk: correlated PSI exceedance across three services sharing an upstream dependency; a single systemic flag is raised.

TABLE V. Summary of the four-layer empirical evaluation on public credit data.

Layer	Control	Real condition (not injected)	Observed result	Outcome
L1	Data-quality gate $Q \geq 0.95$	Native missingness + DAYS_EMPLOYED sentinel	$Q = 0.534$ then 0.999 after remediation	Blocks, then admits
L2	Fairness gate $DIR \geq 0.80$	Real disparate impact on protected attribute	DIR 0.822 to 0.912; AUC 0.746 to 0.745	Gate enforced; trade-off quantified

Layer	Control	Real condition (not injected)	Observed result	Outcome
L3	PSI drift + R(t) enforcement	Real covariate shift (income shock)	PSI > 0.20 at week 7; R(t) crosses week 11	Enforcement triggered
L4	Portfolio systemic-risk flag	Shared upstream features under real shift	Cross-model PSI corr 0.999; breach weeks 6-7	Systemic flag raised

TABLE VI. Fairness-accuracy trade-off: DIR and AUC at baseline versus selected operating point.

Metric	Baseline	Selected (DIR ≥ 0.90 target)
DIR	0.822	0.912
AUC	0.746	0.745

E. Reproducibility and Implementation Pathway

Every quantity reported above is produced by a single seeded driver that writes results to a metrics file, from which all tables and figures are regenerated deterministically. The controls map directly onto managed cloud services—SageMaker Clarify for bias and SHAP, Model Monitor for drift, and a model registry serving as the portfolio inventory [28]. No proprietary or personal data is used at any stage.

VII. DISCUSSION

Under this architecture, trustworthy AI becomes a property of the financial system as a whole rather than of any single model. The performance-explainability trade-off is particularly acute in high-stakes financial modeling: richer models capture more structure but resist explanation, complicating validation and the justification of adverse actions under Article 13 [4], [7]. Our evaluation makes this trade-off explicit and measurable—the approximately 0.1-point AUC cost of raising the DIR to the internal target is precisely the kind of quantity a model-risk committee can deliberate over.

Governance placement raises a second tension. Centralized governance ensures consistency and portfolio-level awareness, yet it must not impede domain teams operating under different risk profiles. Governance should therefore be layered: core acceptance criteria (DIR ≥ 0.80, PSI < 0.20, 28-test compliance) applied uniformly, with implementation detail delegated to individual teams. Because the architecture is incrementally deployable one layer at a time and technology-agnostic, each layer maps onto services an institution may already operate, lowering the barrier to adoption.

VIII. LIMITATIONS

The evaluation demonstrates that governance controls respond correctly to conditions arising naturally in a public credit dataset, but several limitations should be acknowledged. First, the study uses a single public dataset and two model families; live institutional data, richer architectures, and longitudinal deployment remain to be examined. Second, the measured gender disparity is mild (baseline DIR = 0.822, already above the 0.80 floor), so reweighing is applied to reach a stricter internal target rather than to remedy a severe violation. Third, because the Home Credit data lacks calendar timestamps, the Layer-3 shift is a constructed covariate shift from real records rather than a naturally temporal one, and the Layer-4 portfolio is a shared-dependency monitoring demonstration rather than three independently trained models. Finally, the Layer-1 evaluation exercises the admission gate; provenance, lineage, access-control, and audit-logging components are realized through standard catalog tooling rather than measured empirically. Thresholds adopted here are industry and regulatory conventions and should be calibrated locally before institutional adoption [21].

IX. FUTURE DIRECTIONS

Three directions follow naturally from this work. First, deploying the architecture on real institutional systems would test practical feasibility and yield empirical governance metrics—lineage completeness, PSI-exceedance frequencies, DIR-compliance rates, and 28-test score distributions—that would validate the framework at scale. Second, extending the architecture to generative and multi-model pipelines raises additional questions around output provenance, hallucination risk, and cross-model dependencies not yet addressed by current model risk management frameworks or the present EU AI Act [7], [18]. Third, reconciling cross-jurisdictional requirements—particularly the interplay between the EU AI Act and US model risk management regulation—remains a key challenge for internationally operating institutions.

X. CONCLUSION

For regulated financial services, trustworthy-AI governance must be built into AI systems from the outset to avoid material regulatory risk. The four-layer architecture presented here provides a structured and quantitatively grounded way to do so: data governance and provenance with attribute-level lineage satisfying BCBS 239; model development and validation with quantitative fairness thresholds and 28-test readiness; runtime monitoring and enforcement with PSI-based drift detection and DIR-triggered suspension; and enterprise governance with portfolio-level risk aggregation. Evaluated on a public credit dataset, the architecture behaved as intended at every layer: the data-quality gate blocked a defective feed ($Q = 0.534$), the fairness control raised the DIR from 0.822 to 0.912 at negligible accuracy cost, the drift monitor triggered enforcement at week 7, and a shared-dependency shock raised a single systemic flag (cross-model correlation = 0.999). These results recast trustworthy AI as an architectural problem—one addressed through disciplined governance-by-design across the full AI lifecycle.

DATA AND CODE AVAILABILITY

The Home Credit Default Risk dataset is publicly available under its competition license [27]. All evaluation code will be made available upon request, with a README specifying exact dependencies, random seeds, and commands required to regenerate every table and figure. No proprietary or personal data is used.

REFERENCES

- [1] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 28, 2015.
- [2] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann, "Software engineering for machine learning: A case study," in Proc. IEEE/ACM 41st Int. Conf. Software Engineering: SEIP, 2019, pp. 291-300. DOI: 10.1109/ICSE-SEIP.2019.00042
- [3] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities. Cambridge, MA: MIT Press, 2023.
- [4] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.
- [5] M. Armbrust, A. Ghodsi, R. Xin, and M. Zaharia, "Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics," in Proc. 11th Annu. Conf. Innovative Data Systems Research (CIDR), 2021.
- [6] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley, "The ML test score: A rubric for ML production readiness and technical debt reduction," in Proc. IEEE Int. Conf. Big Data, 2017, pp. 1123-1132. DOI: 10.1109/BigData.2017.8258038
- [7] European Parliament and Council, "Regulation (EU) 2024/1689 - Artificial Intelligence Act," Official Journal of the European Union, Aug. 2024.
- [8] Basel Committee on Banking Supervision, "BCBS 239: Principles for Effective Risk Data Aggregation and Risk Reporting," Bank for International Settlements, 2013.
- [9] A. A. Harby and F. Zulkernine, "From data warehouse to Lakehouse: A comparative review," in Proc. IEEE Int. Conf. Big Data, 2022, pp. 389-395. DOI: 10.1109/BigData55660.2022.10020719
- [10] Y. Roh, K. Lee, S. E. Whang, and C. Suh, "FairBatch: Batch selection for model fairness," in Proc. Int. Conf. Learning Representations (ICLR), 2021.
- [11] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in Proc. ACM Conf. Fairness, Accountability, Transparency (FAcT), 2019, pp. 220-229. DOI: 10.1145/3287560.3287596
- [12] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Computing Surveys, vol. 54, no. 6, pp. 1-35, 2021. DOI: 10.1145/3457607
- [13] S. Ransbotham, S. Khodabandeh, R. Fehling, B. LaFountain, and D. Kiron, "Winning with AI," MIT Sloan Management Review & BCG, Oct. 2019.
- [14] V. Dignum, Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Cham: Springer, 2019.
- [15] S. Barocas and A. D. Selbst, "Big data's disparate impact," California Law Review, vol. 104, pp. 671-732, 2016.
- [16] S. Bird, M. Dudik, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft Research, Tech. Rep. MSR-TR-2020-32, 2020.

- [17] L. Floridi et al., "An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689-707, 2018. DOI: 10.1007/s11023-018-9482-5
- [18] R. Bommasani et al., "On the opportunities and risks of foundation models," arXiv:2108.07258, 2022.
- [19] M. Wieringa, "What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability," in *Proc. ACM FAccT*, 2020, pp. 1-18. DOI: 10.1145/3351095.3372833
- [20] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed., 2022.
- [21] A. Thiebes, S. Lins, and A. Sunyaev, "Trustworthy artificial intelligence," *Electronic Markets*, vol. 31, pp. 447-464, 2021. DOI: 10.1007/s12525-020-00441-4
- [22] NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, Jan. 2023.
- [23] ISO/IEC 42001:2023, "Information technology - Artificial intelligence - Management system," ISO, 2023.
- [24] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, vol. 33, no. 1, 2012.
- [25] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Proc. NeurIPS*, vol. 29, 2016.
- [26] Consumer Financial Protection Bureau, "Equal Credit Opportunity Act (Regulation B), 12 CFR Part 1002," 2023.
- [27] Home Credit Group, "Home Credit Default Risk," Kaggle dataset, 2018. [Online]. Available: <https://www.kaggle.com/c/home-credit-default-risk>
- [28] Amazon Web Services, "Amazon SageMaker Clarify and Model Monitor," AWS Documentation, 2024. [Online]. Available: <https://docs.aws.amazon.com/sagemaker/>