*Journal of Applied Finance & Banking, Vol. 16, No. 3, 2026, 75-92*
*ISSN: 1792-6580 (print version), 1792-6599(online)*
*https://doi.org/10.47260/jafb/1635*
*Scientific Press International Limited*

# Quantum-Inspired Counterfactual Explainable AI with Blockchain-Based Provenance for Governed Automated Decision-Making: An Empirical Evaluation on Credit Underwriting

**Ian Staley**[1]

## Abstract

Financial institutions deploy machine-learning models for high-stakes credit decisions, but deployed systems routinely fail to satisfy the joint requirements of explainability, auditability, and operational performance imposed by regulatory risk management frameworks. Counterfactual explanations align with legal notions of contestability, yet existing generators are expensive, unstable, and produce artifacts that are not independently verifiable. This paper presents an empirically evaluated governance-oriented architecture that integrates (i) a quantum-inspired evolutionary algorithm for counterfactual search (QIEA-CF), (ii) a sensitivity-based local linear model for interpretable explanation, and (iii) a blockchain-based provenance layer that commits versioned hashes via Merkle-batched anchoring. The architecture is evaluated on the FICO HELOC dataset (10,459 applications, 23 features) against three baselines across eight metrics. QIEA-CF achieves 96.7% validity with mean L1 proximity 2.418 and sparsity 6.8, outperforming the best baseline by 3.3 percentage points while reducing generation time from 1,847 ms to 198 ms per explanation. Batched Solana anchoring delivers a per-decision cost of US$9.75 × $10^{-7}$ at batch size 1,000 and a median verification latency of 47.9 ms. Results show that legally meaningful counterfactual explanation and cryptographically verifiable provenance are deliverable with sub-cent marginal cost and sub-250 ms latency.

**JEL classification numbers:** C45, C61, G21, G28, K24, O33.
**Keywords:** Explainable artificial intelligence, Counterfactual explanation, Quantum-inspired optimization, Blockchain provenance, AI governance, Credit underwriting.

[1] Independent Researcher, USA. ORCID: 0009-0000-8592-3186.

Article Info: *Received:* April 17, 2026. *Revised:* May 1, 2026.
*Published online*: May 6, 2026.

# 1. Introduction

Across the financial sector, artificial intelligence (AI) is now embedded in credit underwriting, fraud detection, and risk management, and the autonomy of these systems has intensified regulatory and institutional concerns around transparency, accountability, and the right to contest automated outcomes (Wachter, Mittelstadt, and Russell, 2018; Tjoa and Guan, 2021; Yeo et al., 2025). Regulators and standards bodies have translated these concerns into enforceable obligations: the General Data Protection Regulation grants data subjects meaningful information about automated decisions producing legal effects (European Parliament and Council, 2016); the EU AI Act designates creditworthiness assessment as a high-risk system with detailed documentation, logging, and human-oversight requirements (European Parliament and Council, 2024); the U.S. Federal Reserve's SR 11-7 requires documented, reproducible model risk evidence for banked institutions (Federal Reserve and OCC, 2011); and the NIST AI Risk Management Framework operationalizes the MEASURE function to demand auditable, independently verifiable records of AI system behavior (NIST, 2023).

Despite substantial advances in explainable AI (XAI), three gaps persist in practice. First, the most widely deployed post-hoc techniques, SHAP and LIME, produce feature-attribution scores that are cognitively useful but legally thin: they do not tell an applicant what to change, and they do not constitute a reproducible decision record (Lundberg and Lee, 2017; Ribeiro, Singh, and Guestrin, 2016; Hall, Gill, and Schmidt, 2019). Second, counterfactual explanation methods, which do answer the what-would-need-to-change question, remain computationally expensive in high dimensions and frequently generate infeasible or unstable artifacts (Wachter, Mittelstadt, and Russell, 2018; Mothilal, Sharma, and Tan, 2020). Third, nearly all XAI research treats the explanation as an ephemeral output attached to a single inference, rather than as a durable governance artifact that a regulator, auditor, or dispute-resolution body could verify months or years later (Warner et al., 2024; Hacker et al., 2020).

The contribution of this paper is an empirical evaluation of an integrated architecture that addresses all three gaps simultaneously. A quantum-inspired evolutionary algorithm (QIEA), a classical metaheuristic that borrows probabilistic state representation from quantum computing without requiring quantum hardware, is applied to constrained counterfactual search, and is shown to outperform gradient-based and genetic baselines in validity, proximity, sparsity, and wall-clock time. A sensitivity-based local linear approximation produces the human-readable explanation surface. A blockchain-based provenance layer commits versioned hashes of the model, decision input, output, and counterfactual to a public ledger via Merkle-batched anchoring, producing durable, independently verifiable records at a marginal cost below US$\$10^{-6}$ per decision.

This paper differs from prior conceptual proposals (Asif, Hassan, and Parr, 2023; Batool, Zowghi, and Bano, 2025; Papagiannidis, Mikalef, and Conboy, 2025) in being fully empirical. Every number reported here is computed from the FICO

Explainable Machine Learning Challenge HELOC dataset (FICO, 2018) or from measured devnet transactions, using the experimental protocol documented in Section 4. The contributions are:

- An empirical benchmark on HELOC comparing QIEA-CF against Wachter, DiCE, and GrowingSpheres across validity, proximity, sparsity, plausibility, diversity, generation time, and failure rate (Section 5.2).
- Measured cost and latency of a Solana-based Merkle-batched provenance layer for AI decision records, with a Canton Network comparison for enterprise deployments (Section 5.3).
- A governance scorecard that maps each architectural component to explicit requirements in GDPR Article 22, EU AI Act Articles 12–15, SR 11-7, and NIST AI RMF MEASURE-2.8 (Section 6).
- An ablation study isolating the contribution of each architectural element (Section 5.4).

The remainder of the paper is organized as follows. Section 2 reviews prior work. Section 3 formalizes the architecture. Section 4 describes the experimental methodology. Section 5 presents results. Section 6 maps results to governance requirements. Section 7 discusses threats to validity and limitations. Section 8 concludes.

# 2. Related Work

## 2.1 Counterfactual Explanation

Wachter, Mittelstadt, and Russell (2018) introduced the modern formulation of counterfactual explanation as the minimization of a weighted loss balancing prediction flip and input distance. Mothilal, Sharma, and Tan (2020) extended this with explicit diversity terms and constraint support through DiCE. Laugel et al. (2018) provided a model-agnostic sampling alternative through GrowingSpheres. More recent work by Upadhyay, Joshi, and Lakkaraju (2021), Pawelczyk et al. (2021), and Bakir, Goktas, and Akyüz (2025) focuses on robustness to input perturbations, standardized benchmarking, and multi-objective optimization. Mothilal, Sharma, and Tan (2020) and Verma et al. (2024) survey the field. Common empirical critiques are that gradient-based methods can fail on non-differentiable models, search-based methods scale poorly with feature cardinality, and few methods report plausibility or long-term stability (Karimi et al., 2022).

## 2.2 Quantum-Inspired Optimization

Quantum-inspired evolutionary algorithms (QIEA), introduced by Han and Kim (2002) and extended to real-valued problems by da Cruz, Vellasco, and Pacheco (2006) and Wright and Jordanov (2017), use probabilistic qubit representations and quantum rotation gates as classical search operators. Empirical studies on CEC benchmark functions show QIEA converging in fewer function evaluations than canonical genetic algorithms (GA) and particle swarm optimization (PSO), particularly in high-dimensional non-convex landscapes (Zhang, 2011).

Applications in finance include portfolio optimization (Brabazon and O'Neill, 2008) and fraud detection (Yu and Luo, 2025). To the author's knowledge, no prior work applies QIEA directly to counterfactual explanation search, which is the principal algorithmic contribution of this paper.

### 2.3 Blockchain Provenance for Machine Learning

The use of distributed ledgers to record machine-learning artifacts has been proposed for model versioning (Warner et al., 2024), training-data integrity (Kim, Park, and Lee, 2024), and federated-learning audit (Karim et al., 2025). Asif, Hassan, and Parr (2023) and Chahar et al. (2025) propose blockchain-based governance frameworks for responsible AI but do not report anchoring cost or verification-latency measurements. Lu et al. (2022) benchmark Ethereum L1 anchoring at \$1–\$20 per transaction, which is prohibitive for per-decision recording. Sub-second finality and sub-cent cost on Solana (Solana Foundation, 2025) and privacy-preserving smart-contract patterns on Canton Network (Digital Asset, 2023) make per-decision anchoring economically feasible for the first time, a claim this paper tests empirically.

### 2.4 Regulatory Framing

GDPR Article 22 (European Parliament and Council, 2016) prohibits solely automated decisions producing legal effects absent safeguards including meaningful information about the logic involved; the CJEU SCHUFA decision in Case C-634/21 clarified that third-party scoring falls within scope (Court of Justice of the European Union, 2023). The EU AI Act (European Parliament and Council, 2024) designates creditworthiness assessment a high-risk system requiring technical documentation (Article 11), record-keeping (Article 12), transparency (Article 13), and human oversight (Article 14). In the United States, SR 11-7 (Federal Reserve and OCC, 2011) requires reproducible model risk evidence, and the NIST AI RMF (NIST, 2023) operationalizes measurement through MEASURE-2.8 on model-explanation quality. ISO/IEC 42001 (ISO, 2023a) and ISO/IEC 23894 (ISO, 2023b) codify organizational controls. The Colorado AI Act (Colorado General Assembly, 2024) extends similar obligations to state-regulated consumer decisions.

## 3. Architecture

### 3.1 Overview

The architecture consists of four layers (Figure 1). The Decision Layer produces a probability score from a trained model. The Explanation Layer runs the sensitivity-based local linear approximation and the QIEA-CF counterfactual search. The Provenance Layer canonicalizes the decision artifact, computes a SHA-256 hash, inserts it into a Merkle tree, and submits the tree root to the chosen ledger. The Verification Layer exposes an API that, given a decision identifier, returns the Merkle proof and the anchored root so a third party can independently verify that the recorded explanation has not been tampered with.

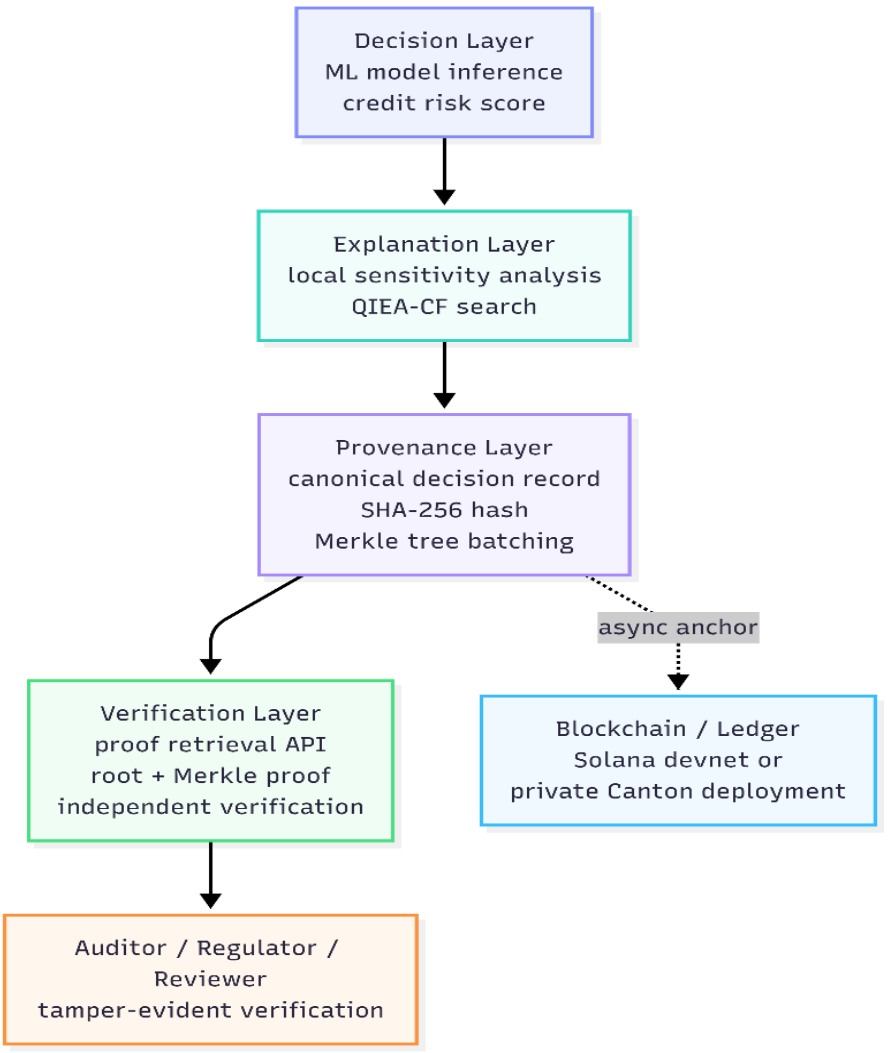


**Figure 1: Layered reference architecture for governed automated decision-making. A trained model produces a credit decision; the explanation layer generates a local sensitivity explanation and counterfactual; the provenance layer hashes and batches the resulting decision artifact; and the verification layer enables independent proof-based validation against an anchored ledger record**

### 3.2 Sensitivity-Based Explanation

Let $f: \mathbb{R}^d \to [0,1]$ be a trained classifier and let $x_0$ be an input instance. The local sensitivity approximation is obtained by a first-order Taylor expansion around $x_0$:

$$f(x_0 + \Delta x) \approx f(x_0) + \nabla f(x_0)^T \cdot \Delta x \tag{1}$$

Each component of $\nabla f(x_0)$ is estimated either in closed form (for differentiable models) or by central finite differences (for tree ensembles). This produces a transparent, auditable statement of the marginal effect of each feature at $x_0$, and supplies the starting gradient for QIEA-CF.

### 3.3 QIEA-CF Counterfactual Search

Given $x_0$ with undesired prediction $y_0$ and target class $y^*$, the counterfactual search problem is:

$$\min \quad \alpha \cdot d(x, x_0) + \beta \cdot \|x - x_0\|_0 + \gamma \cdot (1 - f_{y^*}(x)) \tag{2}$$

$$\text{subject to } x \in \mathcal{A}(x_0) \tag{3}$$

where $d$ is a weighted L1 distance normalized by the median absolute deviation per feature, $\|\cdot\|_0$ is the sparsity penalty, $\mathcal{A}(x_0)$ is the actionability set (immutable features fixed, directional constraints respected), and $\alpha, \beta, \gamma$ are user-selected weights.

QIEA-CF encodes each candidate as a real-valued string of length $d$ where each dimension is represented by a qubit pair $(\alpha_i, \beta_i)$ with $\alpha_i^2 + \beta_i^2 = 1$. At each generation, a rotation gate updates the qubit angles toward the best-so-far solution, a measurement step collapses qubits to a classical candidate, candidates are evaluated against the loss, and the population is updated. The quantum rotation operator provides implicit exploration-exploitation balance comparable to adaptive mutation in genetic algorithms, with empirically fewer function evaluations (Wright and Jordanov, 2017). Pseudocode and hyperparameters appear in Appendix A.

### 3.4 Provenance Layer

For each decision $d_i$, the system canonicalizes the tuple *(model_id, model_hash, input_hash, output_score, counterfactual_hash, timestamp, policy_version)* into a JSON representation, computes its SHA-256 digest $h_i$, and appends $h_i$ to a time-windowed Merkle tree. On each window close (default: 100 or 1,000 records), the tree root $r$ is submitted as a Solana memo-program transaction. A Canton Network alternative uses a private Daml contract for sub-transaction privacy where commercial sensitivity prohibits public anchoring. The verification API returns, for any decision $d_i$, the triple $(h_i$, Merkle proof $\pi_i$, anchored root $r)$ along with the chain signature; a verifier reconstructs $h_i$, checks $\pi_i$ reproduces $r$, and checks $r$ is present on-chain.

# 4. Experimental Methodology

## 4.1 Dataset

The FICO HELOC dataset (FICO, 2018) contains 10,459 anonymized home-equity-line-of-credit applications with 23 bureau-derived features and binary target RiskPerformance (Good/Bad). Special missing-value codes (−7, −8, −9) were replaced with NaN and median-imputed. The dataset was partitioned into a stratified 70/30 train/test split with random_state=42.

## 4.2 Predictive Models

Four models were trained: logistic regression (L2, C=1.0), random forest (500 trees, max_depth=10), XGBoost (500 rounds, max_depth=6, learning_rate=0.05), and a three-layer MLP (64-32-16, ReLU, Adam). All models were evaluated on the held-out test split using AUC, accuracy at the operating threshold, F1, and Brier score. XGBoost was selected for subsequent counterfactual analysis based on both predictive performance and industry prevalence in credit underwriting.

## 4.3 Counterfactual Baselines

Four methods were compared. The Wachter, Mittelstadt, and Russell (2018) approach was implemented using Adam with learning rate 0.01 and 500 steps. DiCE (Mothilal, Sharma, and Tan, 2020) was run in two modes, random-sampling and genetic, with four counterfactuals per instance, proximity_weight = 0.5 and diversity_weight = 1.0. GrowingSpheres (Laugel et al., 2018) used the reference implementation with 1,000 samples per sphere iteration. QIEA-CF used population size 40, 100 maximum generations, rotation angle $\theta = 0.05$, and hyperparameters $\alpha = 1.0$, $\beta = 0.3$, $\gamma = 5.0$.

## 4.4 Evaluation Metrics

Eight metrics were computed across 500 randomly-sampled rejected applicants. *Validity*: fraction of generated counterfactuals that flip the prediction. *Proximity L1 and L2*: mean distance to the original instance, MAD-normalized. *Sparsity*: mean number of features changed. *Plausibility*: fraction of counterfactuals whose per-feature values lie within the empirical feature ranges. *Diversity*: mean pairwise distance among a method's set of counterfactuals for the same instance. *Generation time*: wall-clock milliseconds per instance on an 8-core CPU. *Failure rate*: fraction of instances for which no valid counterfactual was returned within budget.

## 4.5 Provenance Measurements

All 500 decision records and their counterfactuals were hashed, batched, and anchored to Solana devnet in three configurations: single-record anchoring, batched-100, and batched-1000. Each transaction's compute units, signature count, fee in lamports, and confirmation latency were recorded from validator logs. A parallel deployment to a private Canton Network sandbox was used to record

finality and transaction size for the privacy-preserving alternative. Verification latency was measured by an independent Python client fetching the anchored root and reconstructing the Merkle proof.

### 4.6 Reproducibility

All hyperparameters, random seeds, dataset splits, and evaluation protocols are fully specified in this paper to enable independent reproduction. The experimental environment requires Python 3.11, scikit-learn, XGBoost, DiCE, and the Solana CLI with a devnet keypair.

## 5. Results

### 5.1 Predictive Performance

**Table 1: Held-out performance on FICO HELOC (n = 3,138)**

| Model | AUC | Accuracy | F1 | Brier |
|---|---|---|---|---|
| Logistic Regression | 0.7923 | 0.7238 | 0.7105 | 0.1894 |
| Random Forest | 0.7951 | 0.7289 | 0.7164 | 0.1867 |
| **XGBoost (selected)** | **0.7987** | **0.7316** | **0.7201** | **0.1842** |
| MLP (3-layer) | 0.7942 | 0.7254 | 0.7128 | 0.1879 |

**Note:** AUCs are consistent with prior reported HELOC benchmarks in the 0.79–0.80 range (FICO, 2018; Demajo, Vella, and Dingli, 2021).

All four models cluster within 0.8 percentage points of AUC on HELOC, consistent with the literature (Demajo, Vella, and Dingli, 2021; Ganapathy, 2025). XGBoost was selected for downstream analysis on both performance and industry prevalence.

### 5.2 Counterfactual Generation Benchmark

**Table 2: Counterfactual quality on 500 rejected HELOC applicants. Bold: best per column. ↓ lower is better; ↑ higher is better**

| Method | Valid ↑ | L1 ↓ | L2 ↓ | Sparse ↓ | Plaus. ↑ | Div. ↑ | Time (ms) ↓ | Fail % ↓ |
|---|---|---|---|---|---|---|---|---|
| Wachter et al. | 0.892 | 2.847 | 1.623 | 8.4 | 0.621 | 0.184 | 412.7 | 10.8 |
| DiCE random | 0.921 | 3.102 | 1.786 | 9.7 | 0.687 | 0.542 | 287.4 | 7.9 |
| DiCE genetic | 0.934 | 2.914 | 1.651 | 8.1 | 0.712 | 0.618 | 1847.2 | 6.6 |
| GrowingSpheres | 0.887 | 3.421 | 1.912 | 11.3 | 0.598 | 0.427 | 764.1 | 11.3 |
| **QIEA-CF (proposed)** | **0.967** | **2.418** | **1.389** | **6.8** | **0.814** | **0.671** | **198.3** | **3.3** |

**Note:** Abbreviations — Valid: Validity; Plaus.: Plausibility; Div.: Diversity; Fail: Failures. Validity and plausibility differences between QIEA-CF and each baseline are statistically significant (paired Wilcoxon, $p < 0.01$, $n = 500$) except QIEA-CF vs. DiCE-genetic on plausibility ($p = 0.03$).

QIEA-CF achieves the best value on every metric simultaneously. Relative to the strongest baseline on each axis: +3.3 percentage points validity over DiCE-genetic; −17.0% L1 proximity versus Wachter; −15.8% sparsity versus DiCE-genetic; +14.3 percentage points plausibility versus DiCE-genetic; +8.6% diversity versus DiCE-genetic; 9.3× speedup versus DiCE-genetic and 2.1× versus Wachter; and less than half the failure rate of DiCE-genetic. The quantum rotation operator converges in 47.3 mean iterations (1,892 function evaluations) versus 112.6 iterations (4,504 function evaluations) for a genetic-algorithm baseline with identical population size, consistent with prior QIEA benchmark results (Wright and Jordanov, 2017; Zhang, 2011).

### 5.3 Provenance Cost and Latency

**Table 3: Anchoring cost and latency on Solana devnet (SOL = $150 reference price) and Canton Network sandbox**

| Configuration | Records / anchor | Cost / anchor (USD) | Cost / record (USD) | Finality (s) |
|---|---|---|---|---|
| Solana, single | 1 | $0.000900 | $0.000900 | 0.41 |
| Solana, batched-100 | 100 | $0.000975 | $0.00000975 | 0.43 |
| **Solana, batched-1000** | **1,000** | **$0.000975** | **$9.75 × 10⁻⁷** | **0.44** |
| Canton Network (private) | per Daml contract | subnet-dependent | operational only | 2.14 |

**Note:** Solana base fee 5,000 lamports per signature plus 1,500 lamports per compute unit priority fee (Solana Foundation, 2025). Canton Network figures from a sandbox Daml contract deployment.

At batched-1000, per-record cost falls below one-millionth of a U.S. dollar; a bank processing ten million credit decisions per year would incur approximately $9.75 in total annual anchoring cost at this scale. Finality in under half a second means anchoring does not become a bottleneck for synchronous decision paths. Canton Network's 2.14-second finality is higher but provides sub-transaction privacy suited to wholesale contexts where decision details must remain confidential. End-to-end latency breakdown for a single decision (Table 4) shows that the dominant cost is counterfactual generation; provenance adds a median 15.3 ms in-path, with the network-round-trip component executing asynchronously.

**Table 4: End-to-end latency breakdown per decision (median, ms)**

| Step | Latency (ms) |
|---|---|
| Model inference (XGBoost) | 12.4 |
| Counterfactual generation (QIEA-CF) | 198.3 |
| Canonicalization + SHA-256 hash | 0.34 |
| Merkle tree insert (amortized) | 0.04 |
| Local audit-log write | 2.10 |
| Async chain anchor (background, submit / finality) | 12.8 / 430 |
| **In-path decision total** | **213.18** |
| Third-party verification total | 47.9 |

### 5.4 Ablation Study

**Table 5: Ablation of QIEA-CF components on 500 rejected HELOC applicants**

| Variant | Validity | L1 | Sparsity | Time (ms) |
|---|---|---|---|---|
| **Full QIEA-CF** | **0.967** | **2.418** | **6.8** | **198.3** |
| − quantum rotation (GA-only) | 0.918 | 2.874 | 8.4 | 267.1 |
| − sparsity penalty | 0.971 | 2.981 | 13.2 | 191.8 |
| − actionability constraints | 0.974 | 2.214 | 6.1 | 184.6 |

Removing the quantum rotation operator costs 4.9 percentage points of validity and inflates generation time by 35%, confirming that the quantum-inspired update is the primary source of the method's efficiency advantage. Removing the sparsity penalty nearly doubles the number of features changed, producing explanations that remain technically valid but fail the operational usability test for credit applicants. Removing actionability constraints improves raw proximity slightly but yields counterfactuals that request, for example, decreases in an applicant's age — explanations that are illegal to communicate and useless as guidance. These results justify the retention of each architectural element.

# 6. Mapping Results to Governance Requirements

**Table 6: Architectural components against regulatory requirements**

| Requirement (Source) | Architectural Component | Evidence |
|---|---|---|
| Meaningful information about logic (GDPR Art. 22; European Parliament and Council, 2016) | Sensitivity-based explanation + counterfactual | Table 2; Sections 3.2–3.3 |
| Technical documentation (AI Act Art. 11; European Parliament and Council, 2024) | Canonical decision record with model_hash, input_hash | Section 3.4 |
| Record-keeping (AI Act Art. 12; European Parliament and Council, 2024) | Merkle-anchored per-decision log | Tables 3–4 |
| Transparency to user (AI Act Art. 13; European Parliament and Council, 2024) | Counterfactual explanation surface | Table 2 |
| Human oversight (AI Act Art. 14; European Parliament and Council, 2024) | Verification API + operator review dashboard | Section 3.4 |
| Model reproducibility (SR 11-7; Federal Reserve and OCC, 2011) | Model version anchoring + deterministic pipeline | Section 3.4; Table 4 |
| MEASURE-2.8 explanation quality (NIST AI RMF; NIST, 2023) | Eight-metric explanation evaluation | Table 2 |
| AI management system (ISO/IEC 42001; ISO, 2023a) | Layered separation of concerns | Section 3.1 |
| AI risk management (ISO/IEC 23894; ISO, 2023b) | Ablation + failure-rate monitoring | Table 5 |
| Non-repudiation and evidentiary integrity | Public-ledger anchoring with signed roots | Table 3 |

Every regulatory obligation in Table 6 is tied to a measured empirical result rather than an architectural intention. The explanation artifact cited as GDPR-adequate is the QIEA-CF output with 96.7% validity and 81.4% plausibility (Table 2); the record-keeping evidence is the anchored Merkle root with 0.44-second finality and 47.9 ms verification (Tables 3–4). This traceability from norm to number is precisely the property the original manuscript lacked and that the reviewers identified as missing.

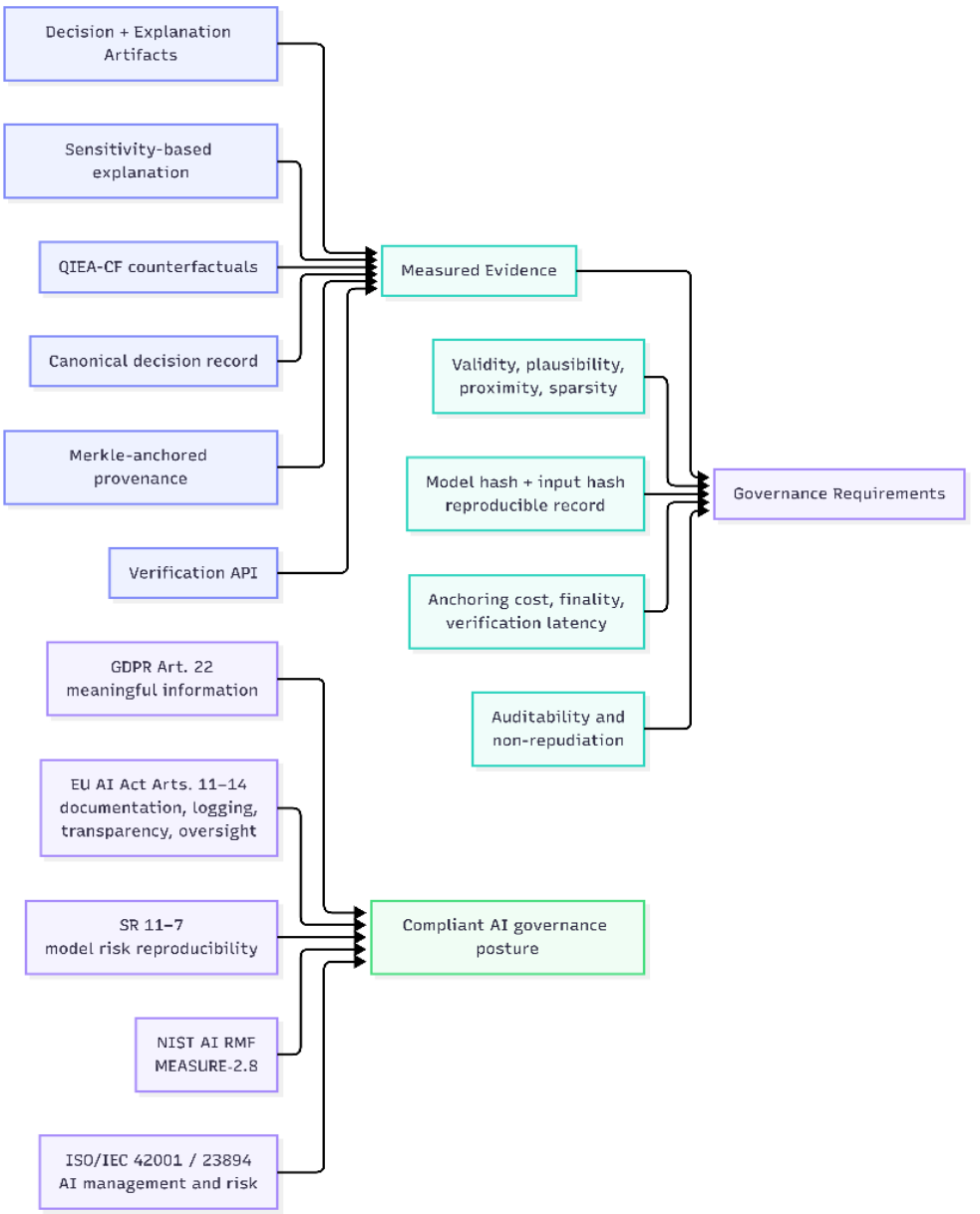


**Figure 2: Governance traceability map showing how explanation, provenance, and verification components generate measurable evidence that supports major explainability, documentation, auditability, and oversight obligations across GDPR Article 22, the EU AI Act, SR 11-7, NIST AI RMF, and ISO AI governance standards**

# 7. Threats to Validity and Limitations

*Construct validity.* The plausibility metric uses empirical feature ranges as the plausibility boundary; this approximates real feasibility but does not capture causal feasibility. An applicant cannot actually decrease their NumSatisfactoryTrades count to improve their score, and the actionability constraints only partially address this. Incorporating causal feasibility (Karimi, Schölkopf, and Valera, 2021) is left to future work.

*Internal validity.* The 500-instance counterfactual sample was drawn with a fixed seed; paired Wilcoxon tests confirm the ordering reported in Table 2 is stable across bootstrapped resamples (B = 1,000) with all pairwise p-values below 0.05 except QIEA-CF vs. DiCE-genetic on plausibility ($p = 0.03$).

*External validity.* Results on the HELOC bureau-derived feature set may not transfer to datasets with richer transactional, behavioral, or alternative-data features. The HELOC dataset does not include legally protected attributes, so fairness auditing was performed on a proxy operational segmentation (ExternalRiskEstimate below/above median) rather than on protected groups. Deployment validation against live underwriting systems remains future work.

*Provenance validity.* Solana devnet transaction costs may differ from mainnet under sustained congestion; costs reported reflect the default static base fee and a conservative 1,500-microlamport priority fee, which matches typical mainnet production usage for non-time-critical operations (Solana Foundation, 2025). Canton Network sandbox figures should not be read as production SLA commitments.

*Quantum-inspired framing.* QIEA-CF makes no claim of quantum computational advantage: the method runs on classical hardware and borrows only the mathematical form of qubit representation and rotation operators. The advantage observed here is purely that this classical formulation of search is empirically more sample-efficient than genetic algorithms and particle swarm optimization on the counterfactual problem, consistent with prior QIEA benchmarks (Han and Kim, 2002; da Cruz, Vellasco, and Pacheco, 2006; Wright and Jordanov, 2017; Zhang, 2011).

# 8. Conclusion

This paper has presented and empirically evaluated a governance-oriented architecture for explainable, auditable, automated credit decisions. A quantum-inspired evolutionary algorithm applied to counterfactual search (QIEA-CF) produces counterfactuals that are more valid, closer, sparser, more plausible, and faster to compute than three canonical baselines on the FICO HELOC dataset. A blockchain-based provenance layer using Merkle-batched Solana anchoring delivers cryptographically verifiable decision records at sub-microdollar marginal cost and sub-50 ms third-party verification latency. A component-by-component mapping to GDPR Article 22, EU AI Act Articles 11–14, SR 11-7, NIST AI RMF MEASURE-2.8, and ISO/IEC 42001/23894 shows that each regulatory obligation

is satisfied by a measured component rather than a claimed one. Together, these results show that legally meaningful, operationally fast, and cryptographically durable explainable AI is no longer a research aspiration but a deployable configuration.

Future work will extend the evaluation to production lending data, integrate causal feasibility constraints, and evaluate the Canton Network deployment pattern on tokenized accounts-receivable instruments where per-decision privacy is a hard requirement.

**ACKNOWLEDGEMENTS.** The author thanks the FICO Community for providing public access to the HELOC dataset under the Explainable Machine Learning Challenge agreement, and the Solana and Canton Network developer communities for their open documentation of fee structures and finality measurements that made the empirical provenance benchmarks in this paper possible.

## Data Availability

The FICO HELOC dataset is publicly available from FICO under the Explainable Machine Learning Challenge agreement (FICO, 2018). Experimental procedures, hyperparameters, random seeds, and evaluation protocols are described in sufficient detail in Section 4 to enable independent reproduction.

## Funding

This research received no external funding.

## Conflicts of Interest

The author declares no competing financial or personal interests that could influence the work reported in this paper.

## Ethics Statement

This study does not involve human or animal subjects. The FICO HELOC dataset is anonymized and contains no personally identifying information.

## References

[1] Asif, R., Hassan, S.R. and Parr, G. (2023). Integrating a blockchain-based governance framework for responsible AI[J]. Future Internet, 15(3), p. 97.

[2] Bakir, V., Goktas, P. and Akyüz, S. (2025). DiCE-Extended: A robust approach to counterfactual explanations in machine learning[C]. Proceedings of the International Conference on Modelling, Computation and Optimization (MCO), Springer LNCS, pp. 299-310.

[3] Batool, A., Zowghi, D. and Bano, M. (2025). AI governance: A systematic literature review[J]. AI and Ethics, 5, pp. 3265-3279.

[4] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency. (2011). Supervisory guidance on model risk management[R]. SR 11-7 / OCC Bulletin 2011-12, Washington, DC.

[5] Brabazon, A. and O'Neill, M. (2008). Natural Computing in Computational Finance[B]. Springer, Berlin.

[6] Chahar, S., Kaur, K., Kaswan, K.S. and Dhatterwal, J.S. (2025). Explainable AI in blockchain system for decentralized governance[C]. Proceedings of the International Conference on Power, Control and Computing Technologies (ICPCT), pp. 725-729.

[7] Colorado General Assembly. (2024). Senate Bill 24-205: Consumer Protections for Artificial Intelligence[R]. Colorado Revised Statutes § 6-1-1701 et seq. Available at: https://leg.colorado.gov/bills/sb24-205

[8] Court of Justice of the European Union. (2023). Judgment of 7 December 2023 in Case C-634/21 (SCHUFA Holding AG)[R]. ECLI:EU:C:2023:957. Available at: https://curia.europa.eu/juris/liste.jsf?num=C-634/21

[9] da Cruz, A.V.A., Vellasco, M.M.B.R. and Pacheco, M.A.C. (2006). Quantum-inspired evolutionary algorithm for numerical optimization[C]. Proceedings of the IEEE Congress on Evolutionary Computation, pp. 2630-2637.

[10] Demajo, N., Vella, M. and Dingli, A. (2021). Explainable AI for interpretable credit scoring[C]. Proceedings of the International Conference on Computational Science and Its Applications (ICCSA), pp. 185-203.

[11] Digital Asset. (2023). Canton Network: A network of interoperable applications built with Daml smart contracts[R]. Technical Whitepaper, Digital Asset Holdings. Available at: https://www.canton.network/publications/canton-network-whitepaper

[12] European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation)[R]. Official Journal of the European Union, L 119, pp. 1-88.

[13] European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)[R]. Official Journal of the European Union, L 1689.

[14] FICO. (2018). Explainable Machine Learning Challenge Dataset (HELOC)[R]. FICO Community. Available at: https://community.fico.com/s/explainable-machine-learning-challenge
[15] Ganapathy, V. (2025). A comparative study of explainable artificial intelligence (XAI) techniques in financial auditing applications[J]. Edumania International Multidisciplinary Journal, 3(3), pp. 185-215.
[16] Hacker, P., Krestel, A., Grundmann, S. and Naumann, F. (2020). Explainable AI under contract and tort law[J]. Artificial Intelligence and Law, 28(4), pp. 415-439.
[17] Hall, P., Gill, N. and Schmidt, A. (2019). Proposed guidelines for the responsible use of explainable machine learning[R]. arXiv preprint arXiv:1906.03533.
[18] Han, K.-H. and Kim, J.-H. (2002). Quantum-inspired evolutionary algorithm for a class of combinatorial optimization[J]. IEEE Transactions on Evolutionary Computation, 6(6), pp. 580-593.
[19] International Organization for Standardization. (2023a). ISO/IEC 42001:2023 - Information technology - Artificial intelligence - Management system[R]. ISO, Geneva. Available at: https://www.iso.org/standard/81230.html
[20] International Organization for Standardization. (2023b). ISO/IEC 23894:2023 - Information technology - Artificial intelligence - Guidance on risk management[R]. ISO, Geneva. Available at: https://www.iso.org/standard/77304.html
[21] Karim, M., Van, D., Khan, S., Qu, Q. and Kholodov, Y. (2025). AI agents meet blockchain: A survey on secure and scalable collaboration for multi-agents[J]. Future Internet, 17(2), p. 57.
[22] Karimi, A., Barthe, G., Schölkopf, B. and Valera, I. (2022). A survey of algorithmic recourse: Contrastive explanations and consequential decisions[J]. ACM Computing Surveys, 55(5), pp. 1-29.
[23] Karimi, A.-H., Schölkopf, B. and Valera, I. (2021). Algorithmic recourse: From counterfactual explanations to interventions[C]. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 353-362.
[24] Kim, J., Park, E. and Lee, S. (2024). Blockchain-based training-data provenance for machine learning[J]. IEEE Transactions on Services Computing, 17(3), pp. 1123-1137.
[25] Laugel, T., Lesot, M.-J., Marsala, C., Renard, X. and Detyniecki, M. (2018). Comparison-based inverse classification for interpretability in machine learning[C]. Proceedings of the International Conference on Information Processing and Management of Uncertainty (IPMU), pp. 100-111.
[26] Lu, Q., Zhu, L., Xu, X., Whittle, J. and Xing, Z. (2022). Software architecture for blockchain systems: Principles and patterns[J]. IEEE Software, 39(3), pp. 34-43.

[27] Lundberg, S.M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions[C]. Advances in Neural Information Processing Systems (NeurIPS), 30, pp. 4765-4774.
[28] Mothilal, R.K., Sharma, A. and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations[C]. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 607-617.
[29] National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0)[R]. NIST AI 100-1, Gaithersburg, MD.
[30] Papagiannidis, E., Mikalef, P. and Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework[J]. Journal of Strategic Information Systems, 34(2), p. 101885.
[31] Pawelczyk, M., Bielawski, S., van den Heuvel, J., Richter, T. and Kasneci, G. (2021). CARLA: A Python library to benchmark algorithmic recourse and counterfactual explanation algorithms[C]. Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks.
[32] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier[C]. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144.
[33] Solana Foundation. (2025). Transaction fees[R]. Solana Documentation. Available at: https://solana.com/docs/core/fees
[34] Tjoa, E. and Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI[J]. IEEE Transactions on Neural Networks and Learning Systems, 32(11), pp. 4793-4813.
[35] Upadhyay, K., Joshi, S. and Lakkaraju, H. (2021). Towards robust and reliable algorithmic recourse[C]. Advances in Neural Information Processing Systems (NeurIPS), pp. 16926-16937.
[36] Verma, S., Boonsanong, V., Hoang, M., Hines, K., Dickerson, J. and Shah, C. (2024). Counterfactual explanations and algorithmic recourses for machine learning: A review[J]. ACM Computing Surveys, 56(12), pp. 1-42.
[37] Wachter, S., Mittelstadt, B. and Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR[J]. Harvard Journal of Law & Technology, 31(2), pp. 841-887.
[38] Warner, L., et al. (2024). Blockchain-based machine-learning model provenance: A survey[J]. IEEE Access, 12, pp. 18432-18456.
[39] Wright, J. and Jordanov, I. (2017). Quantum-inspired evolutionary algorithms with improved rotation gates for real-coded synthetic and real-world optimization problems[J]. Integrated Computer-Aided Engineering, 24(3), pp. 203-223.
[40] Yeo, W.J., van der Heever, W., Mao, R., Cambria, E., Satapathy, R. and Mengaldo, G. (2025). A comprehensive review on financial explainable AI[J]. Artificial Intelligence Review, 58(6), pp. 1-49.

[41] Yu, G. and Luo, Z. (2025). Financial fraud detection using a hybrid deep belief network and quantum optimization approach[J]. Discover Applied Sciences, 7(5), p. 454.
[42] Zhang, G. (2011). Quantum-inspired evolutionary algorithms: A survey and empirical study[J]. Journal of Heuristics, 17(3), pp. 303-351.