

Research Article

Beyond Suspicion Scores: Competing-Explanation Inference for Fraud-Shaped Legitimate Activity

Ankur Garg¹, Jeffrey Esposito¹¹. Rockland Trust Company, United States

Fraud detection systems commonly score suspiciousness and act when a score crosses a threshold. This is unreliable when legitimate activity exhibits fraud-like structure: in our diagnostic corpus, structurally suspicious legitimate cases carry more fraud-shaped relational structure than genuine fraud does, and every purely structural statistic we evaluated ranked them above fraud. We present AFIE, a knowledge-based evidentiary architecture that treats fraud recognition as adjudication between competing explanations rather than accumulation of suspicion. A case likelihood is composed on a bounded log-odds scale as $L0 + F + M - B - C$, combining affirmative fraud-semantic evidence (F) grouped into coherent mechanisms, gated mature recognition (M), affirmative exculpatory evidence (B), and contradiction evidence (C). The architecture is recomputed from the current evidence state at each observation prefix, has no absorbing state, and treats the absence of exculpatory evidence as contributing exactly zero. Its principal contribution is an evidentiary admissibility rule for mature recognition: a completed exemplar is admitted only when it is persistent, positively informative relative to locally estimated ordinary opportunity, a member of the coherent mechanism selected by the affirmative channel, situated in a mechanism whose net evidence supports fraud, and not covered by applicable exculpatory evidence. On a 160-case authored all-comers corpus with semantically adequate documentary evidence, the architecture attains an area under the ROC curve of 0.904 at the final prefix, with final-prefix AUC 0.904 (95% bootstrap CI [0.847, 0.952]), complete mature-fraud detection, zero observed escalation across both legitimate families, 4.3% dangerous under-detection, and 0/1,600 prefix-isolation leakage violations. We report negative findings in full, including the failure of graph-coherence formulations and of an emerging-before-mature hypothesis. The evaluation is a local mechanism proof on an authored corpus and is not independent validation.

1. Introduction

Automated fraud systems are predominantly discriminative: a model or rule set assigns a suspiciousness score, and a threshold converts that score into an action. Reviews of the field consistently describe this scoring-and-classification framing as the dominant one, across statistical, data-mining and machine-learning families alike^{[1][2]}. This framing carries an implicit assumption, that fraud-like observations are rare in legitimate activity, so that accumulating them is informative. The assumption is frequently false. Structured legitimate behaviour – escrow settlement, account transfer on sale or probate, control overrides during operational incidents, supplier onboarding after dormancy – produces the same observable structure that fraud produces, and can produce more of it.

The operational cost of this is well documented. Practitioner accounts of deployed card-fraud systems describe alert volumes far in excess of investigative capacity, with review burden rather than raw accuracy setting the binding constraint^{[3][4]}, and interviews with anti-money-laundering specialists attribute the resulting false-positive load specifically to rule-based monitoring firing on legitimate customer behaviour^[5]. The consequence is not merely a higher false-positive rate, however. It is that the ordering itself inverts. In the diagnostic corpus reported here, cases constructed as structurally suspicious but legitimate carry a higher mean relational-coherence measure than emerging-fraud cases at every observation prefix. A system that scores suspicion will therefore rank the wrong cases highest, and tightening the threshold suppresses genuine detections at the same rate.

This suggests a different question. Rather than asking whether suspicious evidence exists, a fraud system should ask whether that evidence survives the competing legitimate explanations available in the same case, and whether it supports a coherent fraud mechanism strongly enough to justify a consequential conclusion. That is an evidentiary question rather than a scoring question, and it requires exculpatory evidence to be a first-class channel rather than a post-hoc suppression rule.

We present AFIE, an architecture built on that premise. Fraud-semantic observations are grouped into coherent mechanisms and weighted by likelihood ratios against locally estimated ordinary opportunity. Affirmative legitimate explanations are admitted only when positively evidenced and are matched by content to the specific mechanism carrying the fraud evidence. Contradiction evidence invalidates the

specific observation it targets. Mature recognition – the consequential conclusion that a known fraud mechanism is present – is admitted only through a conjunctive evidentiary gate.

The research question is therefore narrower than fraud detection and more specific than anomaly scoring:

Can a fraud-inference architecture distinguish genuine fraud from legitimate activity exhibiting the same suspicious structure by adjudicating positively evidenced competing explanations, rather than by accumulating suspiciousness alone?

Contributions

The contributions are stated at the level the evidence and the literature support. Where an underlying principle is established elsewhere, we claim its application and integration rather than the principle.

1. ARCHITECTURAL INTEGRATION. AFIE combines affirmative fraud evidence, affirmative legitimate explanation, targeted contradiction, mechanism-scoped aggregation, reversible evidence-state recomputation, and mature-recognition admissibility in one fraud-inference architecture. The constituent ideas have antecedents in evidential and defeasible reasoning, but AFIE operationalises them together around a single requirement: a consequential fraud conclusion must remain coherent with the mechanism and evidence that support it.
2. MECHANISM-CONSISTENT ADMISSIBILITY. Mature recognition is admitted only through a conjunctive rule (M1-M8). In the final refinements, M7 requires the completing exemplar to belong to the mechanism selected by the affirmative channel, and M8 requires that mechanism to be positive on balance. Each gate was introduced after a reproducible failure and is reported with the failure it prevents.
3. EXCULPATORY-EVIDENCE CONSTRAINTS. Legitimate explanation contributes only when positively evidenced and content-scoped to the mechanism it explains. Absence of exculpatory evidence contributes exactly zero. These constraints were adopted after observing the specific degenerations produced by negation-defined or coverage-free formulations.
4. FRAUD-SHAPED LEGITIMATE ACTIVITY AS A STRESS CONDITION. On the corpus reported here, structurally suspicious legitimate activity carries more relational coherence than fraud at every observation prefix, producing sign inversion rather than merely weak discrimination in three successive relational formulations. Fraudster camouflage is an adjacent graph-fraud precedent, but

it addresses fraud acquiring benign appearance rather than legitimate activity out-signalling fraud on a suspicious structural signature^[6].

5. MECHANISM-PROOF EVIDENCE AND NEGATIVE FINDINGS. We report a local mechanism proof on a 160-case all-comers corpus with semantically adequate documentary evidence, together with the failed graph-coherence formulations, the semantically inadequate corpus diagnosis, and the failure of the emerging-before-mature hypothesis that originally motivated the work.

2. Related work and background

2.1. *Fraud detection, anomaly detection and relational methods*

Financial-fraud detection spans rule-based and knowledge-based systems as well as statistical and machine-learning approaches. Reviews of the field describe classification, clustering, outlier detection, neural models, Bayesian approaches and rule-based techniques as recurring families of methods^{[1][2]}. Within that landscape the dominant research effort is directed at representation and estimation rather than at what a detection should license: Baesens et al.^[7] characterise the practical performance gains in transaction-fraud systems as arising chiefly from feature and instance engineering, with interpretability treated as a requirement management imposes on the resulting classifier. AFIE belongs to the knowledge-based side of this landscape because its evidence primitives are declared from a domain frame rather than learned; its question, however, is narrower than benchmark classification accuracy: what should a system conclude when the suspicious observation is real but a legitimate explanation is also positively evidenced?

Anomaly detection treats departures from expected behaviour as evidence and is therefore sensitive to how normality is defined^[8]. Relational fraud methods add graph structure, shared attributes and neighbourhood information, and the last several years have seen this become the principal direction of work in the area; a systematic review of graph neural networks for financial fraud detection catalogues the resulting families of models^[9]. Representative contributions learn inductive node representations that generalise to transactions unseen at training time^[10], construct network representations tailored to credit-card transaction graphs^[11], combine strong-node signals with graph topology^[12], encode transaction context with adaptive neighbourhood aggregation^[13], or distribute learning across institutions that cannot share raw transactions^[14]. These methods differ from AFIE in what their evidence is: the discriminating quantity is a learned function of graph position and attributes, whereas

AFIE's evidence primitives are declared semantic conditions and its comparator is an explicitly estimated ordinary-opportunity rate. Dou et al.^[6] show that graph-fraud detectors can be degraded by feature and relation camouflage, in which fraudulent nodes acquire benign-looking local characteristics. Our stress condition is the converse: legitimate activity independently exhibits the suspicious relational signature, in this corpus more strongly than fraud. That distinction motivated the move from generic topology toward semantic evidence primitives. A related but separate line of work treats the fraud environment as non-stationary, so that the mapping from observation to conclusion must be revised as behaviour evolves^[15]. AFIE addresses a different aspect of change: within a single case, the evidence set grows as observation proceeds, and Section 3 requires the conclusion to be recomputed from that set rather than carried forward.

2.2. Likelihood-ratio evidential reasoning

Likelihood-ratio reasoning evaluates an observation under competing propositions rather than under a single model. In forensic evidence evaluation, the denominator depends on a case-relevant or relevant population; the choice of that comparator is an established part of evidential interpretation rather than a new statistical principle^[16]. AFIE applies the same logic to fraud inference by comparing each fraud-semantic observation with the rate at which comparable ordinary opportunity produces it.

2.3. Belief functions and evidence combination

Belief-function approaches distinguish lack of support from support for the opposite proposition and provide formal machinery for combining uncertain evidence^[17]. AFIE enforces a related but narrower operational rule: absence of exculpatory evidence contributes zero; it is not converted into support for fraud. Within an evidence family, AFIE uses geometric discounting to reduce repeated contributions from correlated observations. We claim no optimality for that engineering choice.

2.4. Defeasible reasoning and competing explanations

Defeasible reasoning formalises conclusions that can be withdrawn when defeating information arrives. Pollock^[18] distinguishes rebutting defeat, which supports an incompatible conclusion, from undercutting defeat, which attacks the connection between evidence and conclusion. Later structured argumentation frameworks formalise related attack and preference relations^{[19][20]}. AFIE operationalises this distinction in a quantitative fraud setting: B supports a positively evidenced legitimate explanation

of the mechanism, while C targets and invalidates a specific affirmative observation. The architecture does not claim a new theory of argumentation; it uses defeat as part of an integrated evidentiary decision process.

2.5. Alert verification and pattern completion

The proposition that a matched pattern does not automatically justify the nominal conclusion has direct precedent in intrusion detection. Alert-verification work distinguishes a successful attack from a correctly detected attempt that could not succeed against the target, even though the signature matched^{[21][22]}. AFIE adopts the same general caution but applies different evidentiary tests. M6 requires the completed exemplar itself to carry positive evidential weight relative to ordinary opportunity; M7 requires it to belong to the mechanism selected by the affirmative channel; and M8 requires that selected mechanism to be positive on balance.

2.6. Synthesis

The prior literature therefore supplies important parts of AFIE's intellectual lineage: suspiciousness modelling, relevant-population likelihood ratios, defeasible defeat, evidence combination and verification after pattern match. AFIE's contribution is the way these ideas are configured for fraud-shaped legitimate activity. Suspicious evidence is evaluated inside a coherent mechanism, legitimate explanation is represented affirmatively and scoped to that mechanism, contradiction can remove targeted support, and mature recognition is admitted only when the completing exemplar and the affirmative channel agree about both mechanism identity and net evidential direction.

3. The AFIE architecture

A case is observed as a sequence of nested prefixes. At each prefix the architecture computes a likelihood on a bounded log-odds scale.

$$\Lambda(p) = L_0 + F(p) + M(p) - B(p) - C(p)$$

$$LK(p) = 100 \cdot \sigma(\text{clip}(\Lambda(p), \pm 6)), \sigma(x) = 1/(1 + e^{(-x)})$$

L_0 is a declared standing prior. In the frozen M8 configuration, $L_0 = -2.1972$ (10% prior odds); $\lambda B = 3.00$; $\lambda C = 1.00$; $\lambda M = 1.25$ with mature scale $s_{\text{cap}} = 2.40$, so an admitted mature contribution is $M = 3.000$; the per-observation LLR cap is 4.0; $\kappa_{\text{min}} = 0.60$; within-family redundancy discount $\rho = 0.5$; and exculpatory

blocking threshold $\zeta = 0.70$. The bounded log-odds composition clips Λ to ± 6 before conversion to likelihood. These quantities were declared rather than fitted and were held fixed through the final M8 evaluation. Table 1 summarises the architectural commitments that distinguish this composition from an accumulated score.

Property	Statement	Consequence
Evidence-state recomputation	$\Lambda(p)$ is a function of the evidence set at prefix p. It does not read $\Lambda(p-1)$.	No fixed point, no convergence to an absorbing value. A case can move in either direction on new evidence.
Absence contributes zero	No term reads the absence of an exculpatory item. $B = \lambda_B \log(1+0) = 0$.	A case with no legitimate explanation sits exactly where it would sit had exculpation never been considered. Absence of exoneration is not inculpatory.
Reversibility	No smoothing, rate limiting, hysteresis or latched flag is permitted anywhere in the composition.	A case that has crossed the action threshold may fall below it when exculpatory or contradiction evidence arrives.
Mechanism coherence	Fraud evidence is aggregated only within a connected component of live atoms linked through shared entities; each atom has already satisfied its family-specific semantic and temporal constraints.	Independent suspicious observations with no evidenced entity path cannot be summed to clear a threshold.
Local opportunity conditioning	Each observation is weighted against the rate at which ordinary activity in comparable circumstances produces it, estimated label-free.	An observation common in ordinary activity carries little or negative weight regardless of how suspicious it appears.
Content-scoped exculpation	An exculpatory item covers fraud evidence only when its subject, effective interval and assertion type all match the mechanism carrying that evidence.	Exculpatory weight cannot be earned by the mere existence of documentation.
Contradiction invalidates	Contradiction evidence removes the specific observation it targets from the affirmative channel and contributes a residual term.	A challenged observation stops supporting the conclusion it was supporting.

Table 1. Architectural commitments of the AFIE composition.

4. Affirmative fraud evidence (F)

F is an accumulator of affirmative evidence, not a measure of structural unusualness. Four evidence families are derived from the frozen domain frame. Each is defined by a semantic, accounting or control discriminator rather than by topology, which is the central design decision of this component: the topological shape of each family occurs in legitimate activity at comparable rates, and it is the semantic condition that carries the evidence.

	Observable	Semantic discriminator	Legitimate analogue it must survive
EF-1	Value leaves a source to two or more intermediaries and reconverges at a common destination within a bounded interval	Value conservation across the split	Escrow or split settlement, where value legitimately conserves
EF-2	An account control change precedes an outflow from the same account within a bounded interval	The controlling party differs before and after, distinguishing a control transfer from a status change	Account sale, executor appointment, mandated signatory change
EF-3	A control fires, does not pass or is overridden, and the triggering activity nevertheless completes	The authorisation state of the override	Operational incident with retrospective authorisation
EF-4	A dormancy interval is followed by a burst of outflows	Counterparty novelty against the prior-history window	Supplier onboarding after a seasonal gap

Table 2. The four affirmative evidence families and the legitimate behaviour each must be distinguished from.

4.1. Weighting

Each observation a of family f is weighted by a capped log-likelihood ratio against the rate at which applicable cases in ordinary circumstances produce that observation:

$$\text{LLR}(a) = \min(\text{LLR}_{\text{cap}}, \log(\pi_f/p_{\text{opp}}(f))) \cdot \kappa(a)$$

π_f is a declared probability of the observable under an active fraud mechanism (EF-1 0.75, EF-2 0.85, EF-3 0.90, EF-4 0.80). The frozen label-free ordinary-opportunity rates are $p_{opp} = \{\text{EF-1: } 0.612, \text{EF-2: } 0.680, \text{EF-3: } 0.581, \text{EF-4: } 0.997\}$. They were estimated once over applicable cases at the final corpus boundary and then held fixed across all ten prefixes. Because the pool includes fraud cases, the denominator is conservative with respect to the authored corpus. $\kappa(a)$ is the fraction of five required exemplar elements satisfied: three semantic elements plus two relationship conditions, so $\kappa \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$; atoms contribute to F only at $\kappa \geq 0.60$. The ordinary-opportunity denominator plays the role of a relevant comparator: AFIE applies the established relevant-population principle to the locally applicable activity represented in the corpus^[16].

4.2. Aggregation under mechanism coherence

Operationally, live affirmative atoms are grouped into connected components through shared entities. Each atom has already satisfied its family-specific semantic and temporal constraints before grouping; the component therefore binds observations through an evidenced entity path rather than through case-level co-occurrence. Within a mechanism, weights of the same family are discounted geometrically to handle redundancy between correlated observations; distinct families are summed. The case-level value is the maximum over mechanisms, not the sum:

$$F(p) = \max \text{ over mechanisms } g \text{ of } \sum_{\text{families}} \Sigma_j \text{LLR}_j \cdot \rho^{(j-1)}$$

The maximum rather than the sum is deliberate. Summing across mechanisms would permit unrelated suspicious activity in different parts of a case to accumulate into a conclusion about none of them.

4.3. The bounded role of F

We state plainly that F does not by itself separate fraud from fraud-shaped legitimate activity on this corpus. Mean F at the final prefix is +0.312 for emerging fraud and +0.237 for structurally suspicious legitimate cases (Table 5). The four families are semantically correct and contribute real evidence, but on a corpus where legitimate activity produces the same semantics, that evidence is small and only weakly separating. F 's role in the architecture is to identify and weight a candidate mechanism, not to reach a conclusion; the conclusion is reached by the interaction of F with M , B and C .

5. Mature recognition as evidentiary admissibility (M1–M8)

Mature recognition is the consequential conclusion that a known fraud mechanism is present. Pattern completion alone is not treated as entitlement to conclude. That general caution has precedent in alert verification, where a signature match may correctly identify an attempted attack that nevertheless could not succeed against the target^{[21][22]}. AFIE places that principle inside a different evidentiary architecture: completion must survive conjunctive conditions on persistence, exculpatory compatibility, informativeness, mechanism identity and net mechanism support. The final rule set was reached through failure-driven refinement, and we report that progression because each added gate corresponds to an observed, reproducible failure mode.

Gate	Condition	Failure mode prevented	Evidence
M1	The observation is a coherent mechanism: a live atom inside a connected entity-linked component, with the atom satisfying its family-specific semantic and temporal conditions	Unrelated observations across a case combining into an apparent mechanism	Aggregation constraint, Section 4.2
M2	$\kappa = 1.0$: every required element of the exemplar is satisfied	Partial instantiation being treated as the mechanism itself	Definitional
M3	The completing exemplar persists across two consecutive prefixes	Coincidental single-prefix completion; transient conjunctions	Definitional
M5	No applicable exculpatory item covers the exemplar at or above the coverage threshold	A fully explained mechanism being recognised as fraud	Section 6
M6	The completing exemplar carries positive evidential weight, $LLR > 0$	An exemplar that is completed but ordinary, triggering recognition. EF-4 occurred in 99.7% of cases, giving $LLR = -0.220$, and fired mature recognition on cases with no affirmative evidence at all	Five legitimate cases fired M with $F = 0$, $B = 0$, $C = 0$. After M6: zero such firings; benign-drift escalation fell from 2.5% to 0%; AUC rose from 0.884 to 0.896
M7	The exemplar belongs to the mechanism selected by the affirmative channel	Recognition maturing one structure while the affirmative channel evaluates another, so the two channels disagree about what the case is	33 case-prefixes across 9 cases held a completed exemplar outside the selected mechanism. After M7: zero such cases; AUC rose to 0.896
M8	The selected mechanism carries positive net evidential weight	A mechanism containing one informative exemplar while being, on balance, evidence against fraud	25 firings across 9 cases occurred inside mechanisms with net weight ≤ 0 . After M8: zero; AUC rose to 0.904; no valid mature case was lost

Table 3. The mature-recognition admissibility conditions, with the failure mode each prevents. Gate numbering follows the development sequence; M4 was subsumed into M3 during consolidation.

M6, M7 and M8 answer different questions. M6 asks whether the completing exemplar is evidentially informative at all; in that sense it is a fraud-specific implementation of the broader alert-verification principle. M7 asks whether the completed exemplar belongs to the same mechanism the affirmative channel selected. M8 asks whether that mechanism, taken as a whole, supports fraud. A mechanism can contain one positively informative exemplar and still be net evidence against fraud because other observations in the same structure are more common under ordinary opportunity. Recognition in that situation would make the recognition component disagree with the architecture's own affirmative channel.

In the final configuration, all 547 mature firings satisfy every gate: minimum exemplar weight +0.204, minimum net mechanism weight +0.018, and every firing inside the selected mechanism.

6. Exculpatory and contradiction evidence (B and C)

B and C are affirmative channels. Neither is a suppression rule, and neither is derived from the fraud channel.

6.1. Admissibility of exculpatory evidence

An exculpatory item must assert something positively evidenced: a documented business purpose, a documented legitimate relationship, an authorised exception, a known operational event, a customer-initiated change, counterparty legitimacy evidence, expected periodic behaviour, historical behaviour continuity, or independent corroboration. Predicates defined over fraud-support statistics – of the form 'fraud feature $\leq k$ ' – are inadmissible. This restriction is load-bearing. In an earlier configuration the exculpatory channel was defined by such predicates and consequently collapsed exactly when fraud support rose, so the subtraction amplified the accumulation instead of opposing it.

6.2. Content-scoped coverage

An item covers a fraud observation only when its subject entity belongs to the observation's entity set, its effective interval covers the observation, and its assertion type is applicable to that evidence family. Coverage is therefore driven by content, and can be zero while F is large or large while F is zero. Under a coverage-free formulation the exculpatory term became a function of whether any mechanism existed at all, and so carried no independent information.

$$B(p) = \lambda_B \cdot \log(1 + \Sigma_g b_g \cdot \text{cov}(g, p))$$

6.3. Persistence

An admitted item remains active without time or density decay. A document is inactive when its identifier is named by a superseding document; otherwise it remains eligible subject to its effective interval and the current prefix boundary. Density of activity is not a state-transition input.

6.4. Contradiction

Contradiction evidence is represented by a challenge keyed to an evidence family and subject entity. At prefix p, any affirmative atom containing that subject entity in the challenged family is removed from the live affirmative set. The contradiction term is $C = \lambda_C \log(1 + n_{\text{invalidated}})$, where $n_{\text{invalidated}}$ is the number of affirmative atoms removed. A challenged observation therefore stops supporting the conclusion and contributes a residual contradiction penalty.

6.5. Resolution of fraud-shaped legitimate activity

This is the architecture's intended behaviour on the hardest cases. Of 50 structurally suspicious legitimate cases, 26 carry positive affirmative fraud evidence at the final prefix and 31 carry positive exculpatory evidence; mean mature contribution is +1.380, so many of them do reach mature recognition. None escalates. They are resolved downward by mean $B = 0.694$ against mean $B = 0.075$ in emerging-fraud cases, a ratio of approximately nine to one. The rejection is therefore attributable to the presence of legitimate explanation, not to the absence of fraud evidence.

7. Experimental method

The evaluation is a mechanism proof. Its purpose is to determine whether the architecture behaves as specified when both affirmative fraud evidence and affirmative legitimate evidence are present and semantically expressible. It is not independent validation, and the corpus was authored by the same parties who specified the architecture.

7.1. Corpus

Family	n	Construction
EMERGING_FRAUD	40	Fraud mechanism unfolds progressively across prefixes; only a subset of evidence families appears in any given case; some carry partial or delayed legitimate corroboration; some carry challenges that must be overcome
MATURE_FRAUD	30	Mechanism present early and complete; a subset carries corroborating documentation that does not survive scrutiny
HARD_BENIGN	50	Legitimate activity carrying genuine fraud-semantic evidence with real explanations: escrow splits that conserve value, account-sale control transfers, outage-driven control failures with retrospective authorisation, supplier-onboarding dormancy breaks; some explanations arrive late; some documents are incomplete or superseded
BENIGN_DRIFT	40	Ordinary activity with sporadic coincidental satisfaction of the evidence families

Table 4. Composition of the 160-case all-comers corpus.

7.2. All-comers design and leakage control

- Every evidence family occurs in both classes at non-trivial rates. Fraud and legitimate cases share suspicious observables by construction.
- Scenario truth is generated from process mechanics. No architecture quantity, threshold or score is consulted during generation.
- The evaluator receives a stripped evidence view; the family label is absent from its input by construction and the absence is asserted programmatically.

- Timing, volumes, entity counts and evidence availability are stochastic. Document counts are drawn from an identical distribution for every family, so exculpatory availability is not a class proxy; only content differs.
- The evaluation boundary is strict: at prefix p the evaluator reads only evidence occurring before the boundary. Re-evaluating any prefix in isolation reproduces the output exactly.
- Success criteria were frozen and hashed before the evaluator was written. No threshold was revised after results were visible and no case was removed.

7.3. Longitudinal evaluation

Each case is evaluated at ten nested prefixes. Prefix p admits observation evidence with $\text{day} < 30p$, producing cutoffs at 30-day increments. The four p_{opp} values are label-free corpus calibration quantities estimated at the final boundary and treated as fixed constants at every prefix; the evaluator itself does not read later case observations when constructing F , M , B or C at an earlier prefix. This design establishes prefix-isolated state recomputation, but it is not a prospective re-estimation of the opportunity model at each prefix.

7.4. Evaluation metrics

AUC is computed at each prefix over 70 fraud cases (40 EMERGING_FRAUD and 30 MATURE_FRAUD) versus 90 legitimate cases (50 HARD_BENIGN and 40 BENIGN_DRIFT) using rank-based AUC with average ranks for ties. Escalation is an analysis-layer action defined as $LK \geq 70\%$; it is not latched or persisted by the evaluator. Dangerous under-detection is the proportion of fraud cases with $LK < 10\%$ at the final prefix. To quantify uncertainty without altering the frozen run, we additionally report post-hoc stratified percentile bootstrap intervals for AUC using 20,000 within-class resamples (seed 20260825), and exact one-sided 95% binomial upper bounds for observed zero-event escalation rates.

8. Results

Metric	Result
Combined-system AUC at final prefix	0.904 (95% bootstrap CI [0.847, 0.952])
MATURE_FRAUD detection	100%
HARD_BENIGN escalation	0/50 (0%; one-sided 95% upper bound 5.8%)
BENIGN_DRIFT escalation	0/40 (0%; one-sided 95% upper bound 7.2%)
Dangerous under-detection (fraud below prior at final prefix)	4.3%
Future-information leakage	0 of 1,600 prefix-isolation checks
Mature firings on non-positive exemplar evidence	0 of 547
Mature firings outside the selected mechanism	0 of 547
Mature firings inside a net-negative mechanism	0 of 547
F+M-only ablation AUC at final prefix	0.785
F+M-only HARD_BENIGN escalation	21/50 (42%)
F+M-only BENIGN_DRIFT escalation	3/40 (7.5%)

Table 5. Principal results, frozen M8 configuration.

Prefix	AUC	95% bootstrap CI
P1	0.568	[0.525, 0.613]
P2	0.703	[0.644, 0.762]
P3	0.817	[0.761, 0.868]
P4	0.849	[0.792, 0.901]
P5	0.811	[0.738, 0.878]
P6	0.796	[0.713, 0.873]
P7	0.841	[0.763, 0.910]
P8	0.901	[0.842, 0.950]
P9	0.908	[0.852, 0.955]
P10	0.904	[0.847, 0.952]

Table 6. Discrimination trajectory across nested prefixes with post-hoc stratified 95% bootstrap intervals. The trajectory contains mid-course dips but rises substantially from P1 to the final prefixes.

Family	n	F	M	B	C	A	Median LK
EMERGING_FRAUD	40	+0.312	2.325	0.075	0.208	+0.158	63.2%
MATURE_FRAUD	30	+0.569	3.000	0.000	0.000	+1.372	79.4%
HARD_BENIGN	50	+0.237	1.380	0.694	0.000	-1.274	10.0%
BENIGN_DRIFT	40	+0.087	0.225	0.277	0.000	-2.162	10.0%

Table 7. Mean term values at the final prefix, by family.

8.1. Attribution

Three properties of Table 7 carry the result. First, F does not separate the classes: +0.312 for emerging fraud against +0.237 for fraud-shaped legitimate activity. Second, M is substantial in both fraud and

hard-benign families, because hard-benign cases genuinely complete exemplars. Third, B is the differentiating term, at 0.694 for hard-benign against 0.075 for emerging fraud. A direct retained-term ablation confirms the attribution: recomposing the final-prefix score as $L_0 + F + M$, with B and C removed and no evaluator rerun, reduces AUC from 0.904 to 0.785 and raises HARD_BENIGN escalation from 0/50 to 21/50 (42%); BENIGN_DRIFT escalation rises from 0/40 to 3/40 (7.5%). The architecture therefore does not reject fraud-shaped legitimate cases because their fraud evidence is weak; it rejects them because a competing legitimate explanation is positively evidenced and scoped to the same mechanism.

Discrimination is not monotonic at every prefix, but it rises from 0.568 at P1 to 0.904 at P10 and exceeds 0.90 from P8 onward. This differs from earlier formulations in which the comparator strengthened alongside the observation and collapsed discrimination toward chance.

The gating results are direct counts from the frozen run: of 547 mature firings, none rests on non-positive exemplar evidence, none lies outside the selected mechanism, and none occurs inside a mechanism with non-positive net weight. The final-prefix AUC is 0.904 with a post-hoc stratified 95% bootstrap interval of [0.847, 0.952]. For the two observed zero-escalation proportions, exact one-sided 95% binomial upper bounds are 5.8% for HARD_BENIGN (0/50) and 7.2% for BENIGN_DRIFT (0/40).

9. Failure analysis

The architecture reported here is the surviving configuration of a longer investigation. We report the failures because they constrain the claim and because several are, in our view, more transferable than the positive result.

9.1. *Graph-coherence formulations failed*

Three successive relational formulations were evaluated and rejected. A permutation null over anchor labels produced resolution but inverted discrimination: fraud in the corpus distributes evidence across entities while a single busy legitimate account concentrates it, so a statistic measuring anchor concentration is signed backwards. A degree-preserving rewiring null removed that inversion by construction but reproduced the observed structure almost exactly in every family, so the ratio of observed to expected coherent structure sat near unity for fraud and legitimate cases alike. A timestamp-permutation null performed worse still.

9.2. Generic relational structure cannot separate fraud from fraud-shaped legitimate activity

Under the rewiring formulation, discrimination between emerging fraud and unstructured benign activity reached 0.797, while discrimination between emerging fraud and structurally suspicious legitimate activity was 0.305 – that is, inverted. Structurally suspicious legitimate cases carried more relational coherence than fraud by every measure examined. This is the finding that motivated the shift from structural to semantic evidence primitives.

9.3. A semantically inadequate corpus cannot test the architecture

An earlier confirmatory corpus of 1,000 cases represented documentary context with two fields, a case identifier and a timestamp, with no assertion content, no entity reference and no linkage. Thirteen of the fourteen fields the semantic evidence families require were absent from its schema, including transaction amounts. In that setting the exculpatory term was observed to take two values only and to be a deterministic function of whether any mechanism existed – carrying no independent information. A paired control study confirmed the diagnosis: a corpus identical except for the removal of corroborating content reproduced the earlier result to within 0.03 AUC.

9.4. The emerging-before-mature hypothesis did not survive

The architecture was originally motivated by the hypothesis that affirmative fraud evidence would become materially positive before mature recognition, permitting earlier intervention. On a corpus in which we staged the mechanism ourselves, a lead was observed in 87% of cases, but the median lead was one prefix and four cases led in the wrong direction. On the difficult all-comers corpus, no emerging-fraud case reached the pre-declared threshold of affirmative evidence before mature recognition. We do not claim emerging detection.

9.5. Strengthening F damaged benign control

One frame-faithful correction to the affirmative channel was specified, frozen and executed: conditioning opportunity on the degree of exemplar completion rather than on its presence. The prediction that it might damage benign control was recorded before execution. It did. Affirmative evidence rose – 52% of emerging-fraud cases reached the threshold – but escalation rose to 30% in hard-benign and 25% in benign-drift cases and AUC fell from 0.904 to 0.822. The cause is structural: fraud-

shaped legitimate cases complete exemplars legitimately, so raising the weight of completion raises both classes equally. The change was reverted and no further alternatives were trialled.

10. Discussion

The results support a narrow but, we think, useful set of propositions.

Suspicious structure alone is insufficient, and can be worse than uninformative

The failure we report is not that structural methods were weakly accurate; it is that the ordering inverted. When legitimate behaviour is deliberately structured, the observations that a structural method treats as evidence occur more often in legitimate cases than in fraud. A system whose evidence is purely topological is exposed to this, and threshold adjustment cannot repair a sign. The mechanism is specific to structured legitimate behaviour and we do not claim it generalises to every relational detector; the graph-fraud literature has largely studied the opposite direction, in which fraud acquires benign local appearance^{[6][9]}.

Fraud inference benefits from competing explanatory evidence

The differentiating term in the final configuration is exculpatory, not inculpatory. Fraud-shaped legitimate cases carry genuine fraud-semantic evidence and many complete mature exemplars, yet they remain below escalation when a positively evidenced legitimate explanation covers the same mechanism. The architecture therefore does not solve the hard-benign problem by suppressing suspicious evidence; it adjudicates that evidence against a competing explanation of the same activity. Complete mature-fraud detection was retained alongside zero observed escalation in both legitimate families. This is a different remedy from the one the applied literature usually reaches for. Where false-positive load is the presenting problem, the standard response is a downstream suppression or triage layer that reprioritises alerts a detector has already raised^{[23][4]}. Such a layer operates on the detector's output; AFIE instead changes what the detector is entitled to conclude, by requiring the competing legitimate explanation to be evidenced and scoped to the same mechanism before the conclusion is reached.

Exculpatory evidence should be first-class, not post-hoc suppression

Two design constraints proved necessary rather than stylistic. Exculpatory evidence must be admitted from positive assertions only: defining it as the negation of a fraud feature makes it collapse precisely when fraud support rises. And its coverage must be content-scoped: without content matching, the term degenerates into an indicator of whether any mechanism exists and carries no independent information. Both failures were observed empirically before the constraints were adopted.

Mature recognition should be evidentially admissible rather than triggered by completion

Pattern completion is a weak trigger. Prior alert-verification work already shows that a signature match can be correct without establishing the nominal adverse outcome. AFIE extends that caution inside its own architecture: a completed exemplar may be ordinary, may belong to a structure other than the one the affirmative channel selected, or may sit inside a mechanism that is net evidence against fraud. M6, M7 and M8 make those distinctions explicit, and each was introduced only after its corresponding failure was observed.

Mechanism coherence supports interpretability and audit

Because evidence is aggregated only within an entity-connected mechanism whose constituent atoms have already satisfied their family-specific semantic and temporal constraints, an explanation names a specific structure and the observations within it rather than a set of unrelated contributing features. Combined with per-derivation provenance and prefix-isolated recomputation, a decision is reconstructible without operator judgement. This differs in kind from post-hoc attribution over a learned model, where the explanation is estimated from the model's behaviour rather than read from the derivation that produced the decision^[24], and it is the distinction Rudin^[25] draws when arguing that consequential decisions should rest on models that are interpretable by construction. We regard this as a practical benefit of the constraint rather than a claim about explanation quality, which we have not evaluated.

What the results do not support

They do not support operational deployment readiness, real-world efficacy, or any comparison with learned fraud detectors. They do not support the emerging-detection hypothesis that originally

motivated the work. And they establish that the architecture behaves as specified when evidence is semantically adequate – not that the evidence primitives will be adequate in a setting we did not author.

11. Limitations

1. The corpus is authored and diagnostic. It was constructed by the same parties who specified the architecture, so the difficulty of the discrimination task was set by the same hand that built the detector. A result on such a corpus establishes capability, not performance.
2. There is no independent external validation and no real-world production data. Every reported figure is a local mechanism-proof result.
3. Emerging-before-mature detection is not demonstrated and is not claimed. No emerging-fraud case reached the pre-declared affirmative-evidence threshold before mature recognition.
4. The four fraud-semantic evidence primitives are derived from one domain frame and may not transfer. Their opportunity rates in this corpus are high, so their individual evidential weights are small.
5. Three of the mature-recognition gates (M6, M7, M8) are post-diagnostic corrections, added in response to observed failures. The final configuration was therefore not evaluated as an untouched confirmatory test, although the corpus, criteria and all other constants were held fixed throughout.
6. The declared quantities π_f , the prior L_o , the channel scales and the coverage threshold are declared rather than fitted. Sensitivity to them has not been characterised.
7. The corpus contains 160 cases, so the reported rates are not performance estimates for a deployment population. Post-hoc bootstrap intervals quantify uncertainty in this authored corpus but do not substitute for independent validation.
8. The opportunity rates p_{opp} were estimated label-free at the final corpus boundary and then held fixed across prefixes. The prefix evaluator does not inspect later case observations, but the calibration itself is corpus-level rather than prospectively re-estimated; a future longitudinal validation should estimate comparator rates from training or historical data only.

12. Future work

- Independent all-comers validation on a corpus or real dataset not authored by AFIE's designers is the necessary next step; no efficacy claim should precede it.

- Transferability should be tested across domain frames in which the semantic discriminators and ordinary-opportunity structure differ.
- Calibration and robustness should be evaluated under distribution shift, including sensitivity to the declared prior, channel scales, coverage threshold, LLR cap and redundancy factor.
- Emergence detection remains open. These experiments do not show that affirmative evidence precedes mature recognition on difficult all-comers cases, and AFIE should not be described as an early-fraud detector on the basis of the present primitives.

Independent validation should repeat the retained-term ablation reported here and add mechanism-coherence and channel-specific ablations on data not authored by AFIE's designers.

Independent evaluation of explanation quality and analyst usability is also required before the architecture's reconstructibility can be treated as an operational interpretability result.

13. Conclusion

AFIE demonstrates locally that fraud inference can be organised around competing evidentiary explanations rather than suspicion accumulation alone. On a 160-case authored all-comers corpus in which legitimate activity carries genuine fraud-semantic evidence, the final architecture reaches an AUC of 0.904 at the final prefix, with complete mature-fraud detection, zero observed escalation in both legitimate families, 4.3% dangerous under-detection, and 0/1,600 prefix-isolation leakage violations. The decisive behaviour on fraud-shaped legitimate activity is not absence of fraud evidence: those cases often contain affirmative fraud evidence and completed exemplars. They resolve downward because positively evidenced legitimate explanation is scoped to the same mechanism and because mature recognition is admitted only when the completing exemplar, selected mechanism and net evidential direction are mutually consistent.

We do not claim early or emerging fraud detection, real-world efficacy, or independent validation. The contribution is architectural: evidence cannot accumulate across unrelated structures; absence of exoneration is not inculpatory; contradiction can invalidate targeted support; and a consequential fraud conclusion must remain coherent with the mechanism and evidence that justify it. The negative results are part of that contribution because they show why the surviving constraints are present.

Statements and Declarations

Funding

No specific funding was received for this work.

Potential Competing Interests

No potential competing interests to declare.

Data Availability

The frozen run package retains the 160-case corpus (seed AC20260825), AFIE_FRAME_V1.0, the active M1–M8 evaluator, frozen success criteria and gate-rule statements, the 1,600 case-prefix result records, the rejected κ -conditional experiment, manifests, and SHA-256 content digests. The active evaluator uses the constants and opportunity rates reported in Sections 3–4. The exported CSV is a projection of the retained JSON; its evidence-cutoff and escalation columns are derived from the fixed prefix rule and 70% analysis threshold. Escalation/disposition was not persisted by the evaluator itself, and this provenance gap is disclosed rather than repaired by rerunning. The post-revert evaluator contains unreachable helper code from the rejected κ -conditional experiment, but re-execution was byte-identical to the retained M8 result artefact. The data supporting the findings of this study, together with the retained reproducibility artefacts necessary to reconstruct the reported mechanism-proof results, are available from the corresponding author upon reasonable request. They are not deposited in a public repository.

References

1. ^aNgai EWT, Hu Y, Wong YH, Chen Y, Sun X (2011). "The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature." *Decis Support Syst.* 50(3):559–569. doi:[10.1016/j.dss.2010.08.006](https://doi.org/10.1016/j.dss.2010.08.006).
2. ^aAl-Hashedi KG, Magalingam P (2021). "Financial Fraud Detection Applying Data Mining Techniques: A Comprehensive Review from 2009 to 2019." *Comput Sci Rev.* 40:100402. doi:[10.1016/j.cosrev.2021.100402](https://doi.org/10.1016/j.cosrev.2021.100402).
3. ^ADal Pozzolo A, Caelen O, Le Borgne YA, Waterschoot S, Bontempi G (2014). "Learned Lessons in Credit Card Fraud Detection from a Practitioner Perspective." *Expert Syst Appl.* 41(10):4915–4928. doi:[10.1016/j.eswa.2014.02.026](https://doi.org/10.1016/j.eswa.2014.02.026).

4. ^a ^bSeera M, Lim CP, Kumar A, Dhamotharan L, Tan KH (2024). "An Intelligent Payment Card Fraud Detection System." *Ann Oper Res.* **334**(1-3):445–467. doi:[10.1007/s10479-021-04149-2](https://doi.org/10.1007/s10479-021-04149-2).
5. ^ΔOztas B, Cetinkaya D, Adedoyin F, Budka M, Aksu G, Dogan H (2024). "Transaction Monitoring in Anti-Money Laundering: A Qualitative Analysis and Points of View from Industry." *Future Gener Comput Syst.* **159**:161–171. doi:[10.1016/j.future.2024.05.027](https://doi.org/10.1016/j.future.2024.05.027).
6. ^a ^b ^cDou Y, Liu Z, Sun L, Deng Y, Peng H, Yu PS (2020). "Enhancing Graph Neural Network-Based Fraud Detectors Against Camouflaged Fraudsters." *Proceedings of the 29th ACM International Conference on Information & Knowledge Management.* pp. 315–324. doi:[10.1145/3340531.3411903](https://doi.org/10.1145/3340531.3411903).
7. ^ΔBaesens B, Höppner S, Verdonck T (2021). "Data Engineering for Fraud Detection." *Decis Support Syst.* **150**:113492. doi:[10.1016/j.dss.2021.113492](https://doi.org/10.1016/j.dss.2021.113492).
8. ^ΔChandola V, Banerjee A, Kumar V (2009). "Anomaly Detection: A Survey." *ACM Comput. Surv.* **41**(3):1–58. doi:[10.1145/1541880.1541882](https://doi.org/10.1145/1541880.1541882).
9. ^a ^bMotie S, Raahemi B (2024). "Financial Fraud Detection Using Graph Neural Networks: A Systematic Review." *Expert Syst Appl.* **240**:122156. doi:[10.1016/j.eswa.2023.122156](https://doi.org/10.1016/j.eswa.2023.122156).
10. ^ΔVan Belle R, Van Damme C, Tytgat H, De Weerd J (2022). "Inductive Graph Representation Learning for Fraud Detection." *Expert Syst Appl.* **193**:116463. doi:[10.1016/j.eswa.2021.116463](https://doi.org/10.1016/j.eswa.2021.116463).
11. ^ΔVan Belle R, Baesens B, De Weerd J (2023). "CATCHM: A Novel Network-Based Credit Card Fraud Detection Method Using Node Representation Learning." *Decis Support Syst.* **164**:113866. doi:[10.1016/j.dss.2022.113866](https://doi.org/10.1016/j.dss.2022.113866).
12. ^ΔChen J, Chen Q, Jiang F, Guo X, Sha K, Wang Y (2024). "SCN_GNN: A GNN-Based Fraud Detection Algorithm Combining Strong Node and Graph Topology Information." *Expert Syst Appl.* **237**:121643. doi:[10.1016/j.eswa.2023.121643](https://doi.org/10.1016/j.eswa.2023.121643).
13. ^ΔLou C, Wang Y, Li J, Qian Y, Li X (2025). "Graph Neural Network for Fraud Detection via Context Encoding and Adaptive Aggregation." *Expert Syst Appl.* **261**:125473. doi:[10.1016/j.eswa.2024.125473](https://doi.org/10.1016/j.eswa.2024.125473).
14. ^ΔTang Y, Liang Y (2024). "Credit Card Fraud Detection Based on Federated Graph Learning." *Expert Syst Appl.* **256**:124979. doi:[10.1016/j.eswa.2024.124979](https://doi.org/10.1016/j.eswa.2024.124979).
15. ^ΔSuárez-Cetrulo AL, Quintana D, Cervantes A (2023). "A Survey on Machine Learning for Recurring Concept Drifting Data Streams." *Expert Syst Appl.* **213**:118934. doi:[10.1016/j.eswa.2022.118934](https://doi.org/10.1016/j.eswa.2022.118934).
16. ^a ^bAitken CG, Taroni F (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*. 2nd ed. John Wiley & Sons. doi:[10.1002/0470011238](https://doi.org/10.1002/0470011238).
17. ^ΔShafer G (1976). *A Mathematical Theory of Evidence*. Princeton University Press. doi:[10.2307/j.ctv10vm1qb](https://doi.org/10.2307/j.ctv10vm1qb).

18. ^ΔPollock JL (1987). "Defeasible Reasoning." *Cogn Sci.* **11**(4):481–518. doi:[10.1207/s15516709cog11044](https://doi.org/10.1207/s15516709cog11044).
19. ^ΔModgil S, Prakken H (2013). "A General Account of Argumentation with Preferences." *Artif Intell.* **195**:361–397. doi:[10.1016/j.artint.2012.10.008](https://doi.org/10.1016/j.artint.2012.10.008).
20. ^ΔPandzic S (2022). "A Logic of Defeasible Argumentation: Constructing Arguments in Justification Logic." *Argument Comput.* **13**(1):3–47. doi:[10.3233/aac-200536](https://doi.org/10.3233/aac-200536).
21. ^Δ^bKruegel C, Robertson W (2004). "Alert Verification: Determining the Success of Intrusion Attempts." In: Fliegel U, Meier M, editors. *Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA 2004), Lecture Notes in Informatics P-46*. Koellen Verlag. pp. 25–38. ISBN [3-88579-365-X](https://doi.org/10.1007/s3-88579-365-X).
22. ^Δ^bKruegel C, Robertson W, Vigna G (2004). "Using Alert Verification to Identify Successful Intrusion Attempts." *Prax Infverarb Kommun.* **27**(4):220–228.
23. ^ΔBakry AN, Alsharkawy AS, Farag MS, Raslan KR (2024). "Automatic Suppression of False Positive Alerts in Anti-Money Laundering Systems Using Machine Learning." *J Supercomput.* **80**(5):6264–6284. doi:[10.1007/s11227-023-05708-z](https://doi.org/10.1007/s11227-023-05708-z).
24. ^ΔLin K, Gao Y (2022). "Model Interpretability of Financial Fraud Detection by Group SHAP." *Expert Syst Appl.* **210**:118354. doi:[10.1016/j.eswa.2022.118354](https://doi.org/10.1016/j.eswa.2022.118354).
25. ^ΔRudin C (2019). "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nat Mach Intell.* **1**(5):206–215. doi:[10.1038/s42256-019-0048-x](https://doi.org/10.1038/s42256-019-0048-x).

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.