Research Article

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

Ankur Garg[1], **Jeffrey Esposito**[1]

1. Rockland Trust Company, United States

Fraud detection systems commonly accumulate suspicious signals and act when a score crosses a threshold. That approach becomes unreliable when legitimate activity exhibits the same fraud-shaped structure. In our diagnostic corpus, structurally suspicious legitimate cases carried more relational coherence than genuine fraud, and purely structural formulations repeatedly ranked them in the wrong direction. We present AFIE, a knowledge-based evidentiary architecture that treats fraud recognition as adjudication between competing explanations rather than accumulation of suspicion. AFIE combines affirmative fraud evidence, mature recognition, positively evidenced legitimate explanation, and contradiction evidence within a reversible, mechanism-scoped inference process. We evaluate AFIE as a mechanism-level diagnostic architecture on a 160-case authored all-comers corpus designed to exercise genuine fraud, fraud-shaped legitimate activity, competing legitimate explanations, contradiction evidence, and benign drift. In the frozen M8 configuration, AFIE achieves a final-prefix AUC of 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 no prefix-isolation leakage. The principal ablation provides the clearest attribution: removing the competing-explanation channels while retaining affirmative fraud evidence and mature recognition reduces AUC to 0.785 and raises HARD_BENIGN escalation from 0/50 to 21/50 (42%). In this diagnostic environment, fraud-shaped legitimate activity is therefore resolved not because suspicious evidence disappears, but because positively evidenced legitimate explanations are evaluated against the same mechanism carrying that evidence.

These results constitute a local mechanism proof, not evidence of real-world or generalisable fraud-detection performance. The final M8 architecture emerged through failure-driven refinement and

 doi.org/10.32388/NAMQIA.2

now requires frozen confirmation, followed by independent validation, parameter robustness, investigator evaluation and, only then, deployment evaluation.

Corresponding author: Jeffrey Esposito, Jeffrey.Esposito@RocklandTrust.com

1. Introduction

Most fraud detection systems accumulate suspicious signals and act once an aggregate score crosses a threshold. That framing works poorly when legitimate activity produces the same observable structure as fraud. Escrow settlement, account transfers on sale or probate, control overrides during operational incidents, and supplier onboarding after dormancy can all look suspicious for legitimate reasons. Reviews of the field describe scoring-and-classification approaches as dominant across statistical, data-mining, and machine-learning methods[1][2].

The practical consequence is familiar: false-positive volume can overwhelm investigative capacity[3][4], and anti-money-laundering practitioners report substantial alert burden from legitimate customer behaviour[5]. The problem examined here is sharper than alert volume alone. In our diagnostic corpus, structurally suspicious legitimate cases carry more relational coherence than emerging-fraud cases at every observation prefix, so purely structural formulations can rank the wrong cases highest rather than merely separate them weakly.

That observation motivates a different inference question: can a fraud architecture distinguish genuine fraud from legitimate activity exhibiting the same suspicious structure by adjudicating positively evidenced competing explanations rather than by accumulating suspiciousness alone? We study that question at the mechanism level, not as a claim of real-world fraud-detection performance.

AFIE is built around that distinction. Fraud-semantic observations are grouped into coherent mechanisms and weighted against locally estimated ordinary opportunity. Legitimate explanations contribute only when positively evidenced and content-scoped to the mechanism they explain; contradiction evidence invalidates the specific support it targets. Mature recognition is admitted only when the completed exemplar and the affirmative channel agree on mechanism identity and net evidential direction.

The contribution is therefore architectural and diagnostic: AFIE tests whether competing-explanation reasoning can resolve fraud-shaped legitimate activity that suspicion accumulation does not distinguish

in the authored environment. The final M8 configuration is treated as frozen after failure-driven development; the present evidence establishes local mechanism behaviour and defines the architecture to be tested next on untouched and independently sourced data.

Contributions

The contributions are stated at the level the evidence and 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.

$$\Lambda(\mathrm{p}) = \mathrm{L}_0 + \mathrm{F}(\mathrm{p}) + \mathrm{M}(\mathrm{p}) - \mathrm{B}(\mathrm{p}) - \mathrm{C}(\mathrm{p})$$

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

L_0 is a declared standing prior. In the frozen M8 configuration, L_0 = −2.1972 (10% prior probability); λ_B = 3.00; λ_C = 1.00; λ_M = 1.25 with mature scale s_cap = 2.40, so an admitted mature contribution is M = 3.000; the per-observation LLR cap is 4.0; κmin = 0.60; within-family redundancy discount ρ = 0.5; and exculpatory blocking threshold ζ = 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 2 summarises the architectural commitments that distinguish this composition from an accumulated score.

3.1. Parameter provenance and status

The frozen M8 configuration contains three distinct kinds of quantities: declared architectural constants and thresholds, empirical calibration quantities, and evaluation-design quantities. The inventory in Table 1 separates them so that corpus-derived calibration is not conflated with parameter fitting and analysis thresholds are not mistaken for architectural components.

Quantity	Frozen value	Role	Provenance / status
L_0	−2.1972	Standing prior (10% prior probability)	Declared; not fitted
λ_B	3.00	Exculpatory-evidence scale	Declared; not fitted
λ_C	1.00	Contradiction-evidence scale	Declared; not fitted
λ_M; s_cap	1.25; 2.40	Mature-recognition scale; M = 3.000	Declared; derived M; not optimised
LLR cap	4.0	Per-observation weight bound	Declared; not fitted
κmin	0.60	Minimum completion for F	Declared threshold; not fitted
ρ	0.5	Within-family redundancy discount	Declared; not fitted
ζ	0.70	Exculpatory blocking threshold	Declared threshold; not fitted
π_f	0.75; 0.85; 0.90; 0.80	EF-1 through EF-4 fraud-mechanism probabilities	Declared by family; not fitted
p_opp	0.612; 0.680; 0.581; 0.997	EF-1 through EF-4 ordinary-opportunity rates	Empirical, label-free corpus calibration; fixed across prefixes
Log-odds clip	±6	Bound on composed Λ	Declared; not fitted
M3 persistence	2 prefixes	Mature-recognition persistence	Definitional gate; not optimised
Escalation threshold	LK ≥ 70%	Analysis-layer action threshold	Pre-declared evaluation quantity
Under-detection threshold	LK < 10%	Dangerous under-detection metric	Pre-declared evaluation quantity
Prefix interval	30 days	Longitudinal observation boundary	Evaluation-design quantity; not optimised

Table 1. Parameter inventory for the frozen M8 configuration.

None of the declared architectural constants or thresholds above was optimised against the reported diagnostic outcomes. The ordinary-opportunity rates p_opp are different in kind: they are empirical, label-free corpus calibration quantities estimated once over applicable cases at the final corpus boundary and then held fixed across all ten prefixes. Declaring rather than fitting the remaining quantities limits one source of corpus adaptation, but it does not establish parameter robustness. Sensitivity to these choices has not yet been systematically characterised and remains a required validation step (Section 11.3).

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. $B = \lambda\_B \cdot \log(1+0) = 0$.	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 2. Architectural commitments of the AFIE composition.

doi.org/10.32388/NAMQIA.2

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 (Table 3).

	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 3. 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:

$$\mathrm{LLR}(a) = \min(\mathrm{LLR}_{\mathrm{cap}}, \log(\pi_{\mathrm{f}} / \mathrm{p}_{\mathrm{opp}}(\mathrm{f}))) \cdot \kappa(a)$$

doi.org/10.32388/NAMQIA.2

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 = {EF-1: 0.612, EF-2: 0.680, EF-3: 0.581, 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. κ(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:

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

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 8). 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 (Table 4).

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	A completed exemplar that is ordinary under the local opportunity comparator 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 nine cases held a completed exemplar outside the selected mechanism. After M7: zero such cases; AUC unchanged at 0.896
M8	The selected mechanism carries positive net evidential	A mechanism containing one informative exemplar	25 firings across nine cases occurred inside mechanisms with net weight ≤ 0. After M8:

Gate	Condition	Failure mode prevented	Evidence
	weight	while being, on balance, evidence against fraud	zero; AUC rose to 0.904; no valid mature case was lost

Table 4. 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

doi.org/10.32388/NAMQIA.2

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.

$$B(p) = \lambda_{\mathrm{B}} \cdot \log(1 + \Sigma_{\mathrm{g}} \mathrm{b}_{\mathrm{g}} \cdot \mathrm{cov}(\mathrm{g}, p))$$

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 = λ_C·log(1+n_invalidated), where n_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 present evaluation is a mechanism-level diagnostic evaluation. Its purpose is to determine whether AFIE exhibits the intended evidentiary behaviour when affirmative fraud evidence, fraud-shaped legitimate activity, competing legitimate explanations, and contradiction evidence are all present and semantically expressible. The evaluation is designed to expose specific architectural failure modes and to determine whether the competing-explanation mechanism resolves those failures under controlled conditions.

The 160-case corpus is therefore a diagnostic development environment, not an external benchmark or representative deployment sample. It was authored to exercise the evidence families and reasoning conditions AFIE is intended to adjudicate, including genuine fraud mechanisms, legitimate activity exhibiting the same fraud-semantic structure, delayed or incomplete exculpatory evidence, contradictory evidence, and ordinary coincidental activity. Scenario truth is determined from the underlying process mechanics rather than from AFIE scores, thresholds, or outputs.

Because the corpus was authored by the same parties who specified and refined the architecture, results obtained on it establish whether the mechanism can behave as intended under the conditions represented here. They do not establish real-world performance, population-level error rates, external validity, or generalisability to independently sourced fraud data. Those questions require evaluation of the frozen architecture on untouched and independently constructed or observed data.

7.1. Diagnostic corpus

The corpus contains four deliberately contrasting case families chosen to exercise the architecture's intended reasoning boundary: progressively developing fraud, mature fraud, fraud-shaped legitimate activity with genuine competing explanations, and ordinary benign drift. Table 5 summarises their construction. These families define diagnostic conditions for mechanism evaluation; their proportions should not be interpreted as estimates of fraud prevalence or operational case mix.

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 5. 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. Architecture development and freeze

The final M8 configuration was reached through failure-driven architectural refinement rather than through an untouched confirmatory sequence. M6, M7 and M8 were introduced after reproducible failures were observed in this diagnostic environment. The evidence associated with those gates is therefore developmental evidence: it establishes that each refinement eliminated its targeted failure within the authored corpus, but it does not establish independent confirmation of the final architecture.

M6 was added after five legitimate cases fired mature recognition with F = 0, B = 0 and C = 0 because a completed EF-4 exemplar was ordinary under the local opportunity comparator and carried negative evidential weight. Requiring the completing exemplar to have LLR > 0 eliminated those firings; benign-drift escalation fell from 2.5% to 0%, and AUC rose from 0.884 to 0.896. M7 was then added after 33 case-prefixes across nine cases contained a completed exemplar outside the mechanism selected by the affirmative channel. Requiring mechanism identity alignment reduced those cases to zero. M8 followed after 25 firings across nine cases occurred inside mechanisms whose net evidential weight was non-positive. Requiring positive net support reduced those firings to zero, raised AUC to 0.904, and removed no valid mature case.

The resulting M8 configuration is the frozen architecture reported in the final analysis. For purposes of this paper, the architecture-freeze boundary is the completion of M8: no further architectural change is claimed or evaluated after that point in the reported final analysis. Its declared constants, success criteria and thresholds were held fixed through that evaluation, and no case was removed after results were visible. Nevertheless, because M6–M8 were specified in response to failures observed on the same authored corpus, the final M8 results remain part of the development and diagnostic evidence base rather than an untouched confirmatory test. The next confirmatory step must therefore apply this frozen M8 architecture unchanged to untouched data not authored in response to AFIE's observed behaviour.

7.4. Longitudinal evaluation

Each case is evaluated at ten nested prefixes. Prefix p admits observation evidence with 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.

7.5. 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 ≥ 70%; it is not latched or persisted by the evaluator. MATURE_FRAUD detection is the proportion of MATURE_FRAUD cases escalating at the final prefix, that is, with LK ≥ 70%. 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 (bootstrap seed 20260825), and exact one-sided 95% binomial upper bounds for observed zero-event escalation rates.

8. Results of the frozen M8 diagnostic evaluation

The results in this section are the final diagnostic results of the frozen M8 configuration on the authored 160-case corpus (Table 6). They are not an untouched confirmatory evaluation and should be interpreted as mechanism-level evidence within the development environment described in Section 7.

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 6. 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 7. 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	Λ	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 8. Mean component term values and median case likelihood at the final prefix, by family.

8.1. Principal ablation: the effect of competing explanatory evidence

The principal frozen-architecture diagnostic result is the retained-term ablation that removes the competing-explanation channels while leaving the frozen M8 evaluator and retained final-prefix terms

otherwise unchanged. When the score is recomposed as L_0+F+M, excluding B and C and without rerunning or refitting the evaluator, final-prefix AUC falls from 0.904 to 0.785. More importantly for the stress condition that motivated the architecture, HARD_BENIGN escalation rises from 0/50 (0%) under the full mechanism to 21/50 (42%) without B and C; BENIGN_DRIFT escalation rises from 0/40 to 3/40 (7.5%). Thus the central result is not merely an increase in aggregate discrimination. In this diagnostic corpus, affirmative fraud evidence plus mature recognition is insufficient to control fraud-shaped legitimate activity; the competing-explanation channels are the components associated with eliminating the observed HARD_BENIGN escalations while preserving the full-system AUC of 0.904.

This comparison is intentionally bounded. Because the corpus is authored and the B/C terms were developed within the same diagnostic programme, the ablation establishes mechanism attribution within this environment; it is not an estimate of the effect size that would occur on independent or operational data. The next confirmatory test is therefore to repeat the same frozen comparison on untouched data without changing M8 or its declared parameters.

8.2. Attribution within the full mechanism

Table 8 explains why the ablation moves the hard-benign cases so sharply. F does not by itself separate the relevant classes: mean F is +0.312 for emerging fraud and +0.237 for fraud-shaped legitimate activity. M is also substantial in both fraud and HARD_BENIGN because hard-benign cases genuinely complete exemplars. B is the differentiating term, with mean B = 0.694 for HARD_BENIGN versus 0.075 for emerging fraud, approximately a nine-fold difference. The full architecture therefore does not reject fraud-shaped legitimate cases because fraud evidence is absent or weak; it resolves them downward because a positively evidenced legitimate explanation is admitted 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 (Table 7). 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 M8 diagnostic 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. The failures in this section are developmental evidence: they document formulations that were evaluated, rejected, or corrected before the M8 freeze. We report them 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 1,000-case diagnostic corpus 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 was consistent with 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.

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

The failure observed here arises because suspicious structure is not uniquely diagnostic of fraud. In this corpus, the HARD_BENIGN cases deliberately contain the same fraud-shaped structures as fraud cases, and in the rejected relational formulations they often express those structures more strongly. A structural or suspicion-accumulating system therefore has no evidentiary basis for distinguishing a fraud mechanism from a legitimate process that produces the same observable pattern; adding more weight to the shared pattern can strengthen both classes together rather than separate them. This is the mechanism-level reason threshold adjustment cannot repair the inversion observed here. The claim is bounded to the diagnostic conditions represented in this corpus and is not a general claim about every relational fraud detector.

10.2. Competing explanatory evidence is the principal discriminator in this diagnostic corpus

The retained-term ablation provides the clearest evidence for the corresponding corrective mechanism. Removing B and C while retaining F and M lowers final-prefix AUC from 0.904 to 0.785 and changes HARD_BENIGN escalation from 0/50 under the full mechanism to 21/50 (42%); BENIGN_DRIFT escalation changes from 0/40 to 3/40 (7.5%). The mechanism is not that AFIE makes fraud-shaped legitimate activity less suspicious: those cases retain genuine fraud-semantic evidence and often complete mature exemplars. Instead, the architecture requires that the suspicious mechanism be adjudicated against affirmative evidence bearing on the same activity. B contributes only when a positively evidenced legitimate explanation covers that mechanism, while C can invalidate specifically challenged affirmative observations. Within this diagnostic environment, those competing-evidence channels are therefore the components associated with resolving shared suspicious structure downward when a legitimate explanation is actually evidenced, rather than allowing the shared structure to accumulate unchecked. Complete mature-fraud detection is retained in the full configuration alongside zero observed escalation in both legitimate families. This differs from downstream suppression or triage, which reprioritises alerts after a detector has already raised them[23][4]. AFIE instead changes what the detector is entitled to conclude. Because the corpus is authored and the B/C terms were developed within the same diagnostic programme, this is a local mechanism interpretation of the observed ablation, not an independently established causal effect size.

10.3. 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.

10.4. 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.

10.5. Mechanism coherence supports decision reconstructibility 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, the derivation of a decision can be reconstructed from a specific structure and the observations within it rather than from a set of unrelated contributing features. Combined with per-derivation provenance and prefix-isolated recomputation, this provides mechanistic traceability and decision reconstructibility. This differs from post-hoc attribution over a learned model, where an explanation is estimated from model behaviour rather than read directly from the derivation that produced the decision[24], while remaining consistent with Rudin's[25] distinction between intrinsically interpretable decision processes and post-hoc explanation. These properties establish traceability of how AFIE reached a result; they do not establish that the resulting explanation is understandable, useful, trustworthy, or actionable to a human investigator. Those are separate human-centred properties that have not been evaluated here.

Reasoning provenance and investigator utility are therefore distinct claims. AFIE records the evidentiary path by which a result was produced, which makes the decision inspectable and reproducible at the mechanism level, but that provenance does not show that an investigator will comprehend the explanation, find it useful, place appropriate trust in it, or act differently because of it. Section 11.4 states the resulting limitation.

10.6. Declared parameters do not establish robustness

The frozen M8 configuration deliberately uses declared rather than outcome-fitted architectural constants and thresholds, but that design choice should not be read as evidence that the reported behaviour is insensitive to them. The present study does not characterise how the principal results change under plausible variation in the prior, channel scales, coverage threshold, LLR cap, redundancy discount, or related declared quantities. Nor does holding those values fixed through the final diagnostic evaluation establish robustness to alternative settings or to distribution shift. Parameter robustness therefore remains an open empirical question: it requires a systematic sensitivity analysis of the frozen architecture, evaluated without retuning to preserve the reported outcomes. Until that analysis is

performed, the results support the mechanism-level claim reported here, not a claim of parameter stability or robustness.

10.7. What the results do not support

They do not support operational deployment readiness, real-world efficacy, parameter robustness, 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 or declared parameter settings will be adequate in a setting we did not author.

10.8. Evidence status and validation hierarchy

The evidence reported here occupies two completed stages of a broader validation sequence: failure-driven mechanism development and diagnostic evaluation of the resulting architecture in the authored environment. The remaining stages are deliberately prospective. Table 9 makes that boundary explicit and separates what this paper establishes from the confirmation and validation still required.

Evidence stage	What it establishes or tests	Status
Mechanism development	Identifies reproducible failure modes and refines the architecture to prevent them; includes the M6–M8 progression reported here.	Completed; developmental evidence
Diagnostic evaluation	Tests whether the final M8 architecture exhibits the intended competing-explanation behaviour in the 160-case authored diagnostic environment.	Completed; evidence in this paper
Frozen confirmation	Applies the unchanged M8 architecture, with no further architectural revision, to untouched data not used in the M6–M8 refinement process.	Required next step
Independent validation	Tests transfer and performance on independently sourced or independently constructed data not authored by AFIE's designers.	Future evidence
Parameter robustness	Pre-specified sensitivity study of the frozen M8 architecture over declared ranges for the prior, channel scales and thresholds, with predeclared stability metrics (Section 12.3).	Future evidence
Investigator evaluation	Human-centred test of whether AFIE's reconstructible reasoning is understandable, useful, and actionable for fraud investigators.	Future evidence
Deployment evaluation	Operational assessment of prospective error rates, alert volume, workflow effects, drift and audit burden, undertaken only after the preceding stages (Section 12.5).	Future evidence

Table 9. Evidence hierarchy for the present study, distinguishing completed evidence from required confirmation and future validation.

11. Limitations

11.1. Diagnostic corpus and external validity

The reported evidence comes from a 160-case authored diagnostic corpus constructed by the same parties who specified the architecture. Its purpose is to exercise fraud, fraud-shaped legitimate activity, competing explanations, contradiction, and temporal evidence under controlled conditions; it is not an

external benchmark, a representative deployment sample, or an estimate of operational prevalence. Accordingly, the reported discrimination, escalation, and under-detection rates establish local mechanism behaviour rather than real-world efficacy or generalisability. The four fraud-semantic evidence primitives are derived from one domain frame and may not transfer unchanged, and the emerging-before-mature hypothesis remains unsupported on the difficult all-comers corpus. Post-hoc bootstrap intervals quantify uncertainty within this authored corpus but do not convert it into independent validation.

Next test: evaluate the frozen M8 architecture on untouched all-comers data not authored by AFIE's designers, then repeat the principal retained-term ablation and assess transfer across at least one independent domain frame.

11.2. Developmental reuse and untouched confirmation

M6, M7, and M8 are failure-driven developmental refinements. Each was introduced after a reproducible failure was observed in the diagnostic environment, so the final M8 configuration cannot be treated as an untouched confirmatory result merely because the architecture was frozen afterward. The reported M8 results therefore show what the surviving architecture does on the development corpus; they do not show that the same behaviour will recur when no architectural change is permitted after outcomes are visible.

Next test: execute the frozen M8 configuration unchanged on an untouched confirmatory corpus with predeclared success criteria, with no gate, threshold, evidence-family, or aggregation changes permitted after inspection of confirmatory outcomes.

11.3. Parameter robustness and calibration

The prior L_0, channel scales, coverage threshold, LLR cap, redundancy factor, and related architectural quantities were declared rather than outcome-fitted, but fixed declaration does not establish robustness. The ordinary-opportunity rates p_opp are empirical calibration quantities estimated label-free at the final corpus boundary and then held fixed across prefixes; this avoids label fitting but remains corpus-level calibration rather than prospective estimation from historical data. The present study therefore does not establish that the reported behaviour is stable to plausible parameter perturbation, parameter interaction, or comparator shift.

Next test: run the pre-specified sensitivity study in Section 12.3 using one-at-a-time sweeps, selected interaction grids, and predefined stability metrics, while evaluating p_opp separately under historical or training-only calibration and distribution shift.

11.4. Human-centred explanation utility

AFIE provides decision reconstructibility and mechanistic traceability: the derivation can identify which evidence was admitted, which mechanism it supported, which competing explanation applied, and which contradiction displaced support. Those properties do not establish that a fraud investigator will understand the explanation, judge it useful, calibrate trust appropriately, or take better action because of it. No human-centred explanation study has been conducted, so interpretability, explanation quality, investigator usefulness, trust, and actionability remain untested claims.

Next test: conduct a blinded investigator-facing evaluation that measures comprehension, perceived and task-based usefulness, trust calibration, decision quality, and actionability against an appropriate baseline explanation or workflow.

12. Future work

The next research programme is staged deliberately so that each new claim is licensed by a corresponding increase in evidentiary independence. The sequence is frozen confirmation, independent validation, parameter robustness, investigator evaluation, and only then deployment evaluation. Later stages should not be used to compensate for evidence missing at an earlier stage.

12.1. Frozen confirmation

The immediate next test is an untouched confirmatory evaluation of the frozen M8 architecture. The evaluator, gate logic, evidence-family definitions, declared constants, thresholds, and success criteria should be fixed before the confirmatory corpus is inspected, and no architectural modification should be permitted after outcomes are visible. The purpose of this stage is narrow: to determine whether the mechanism-level findings reported here, including the principal F+M−B−C versus F+M-only ablation direction, survive a dataset that played no role in the M6–M8 refinement sequence. A failure at this stage should be reported as a failure of confirmation rather than repaired inside the confirmatory run.

12.2. Independent validation

If frozen confirmation succeeds, the next stage is independent all-comers validation on data not authored by AFIE's designers. That evaluation should include independently sourced or independently constructed fraud and legitimate cases, repeat the retained-term ablation, and add mechanism-coherence and channel-specific ablations. Transferability should also be tested across at least one domain frame in which the semantic discriminators and ordinary-opportunity structure differ. This stage is the first one capable of supporting claims about generalisation beyond the present diagnostic environment; no real-world efficacy claim should precede it.

12.3. Parameter robustness

Parameter robustness will be evaluated as a separate, pre-specified study of the frozen M8 architecture rather than as a retuning exercise. Before outcomes are inspected, plausible ranges will be fixed for L_0, λ_B, λ_C, λ_M, s_cap, the per-observation LLR cap, κmin, the within-family redundancy discount ρ, and the exculpatory blocking threshold ζ. Each quantity will first be varied one at a time while all others remain at their frozen values. Bounded interaction grids will then examine only pairs with a direct mechanistic reason to interact: $\lambda\_B \times \zeta$, $\lambda\_M \times s\_cap$, $\kappa min \times$ LLR cap, and $L_0 \times \lambda\_B$. The empirical p_opp quantities will be analysed separately through label-free perturbation or re-estimation from historical or partitioned data.

Stability will be assessed using predeclared metrics tied to the paper's central findings: final-prefix AUC; HARD_BENIGN and BENIGN_DRIFT escalation; dangerous under-detection; MATURE_FRAUD detection; M6–M8 violation counts; and preservation of the principal ablation direction on both AUC and HARD_BENIGN escalation. Results will be classified as stable, sensitive but bounded, or fragile according to whether the qualitative conclusions persist across the predefined region. Robustness will not be inferred from finding a favourable setting, and no range will be altered after outcome inspection merely to preserve the reported conclusions.

12.4. Investigator evaluation

Human-centred evaluation follows only after the mechanism has survived independent validation. A blinded investigator-facing study should measure comprehension of the evidentiary basis, usefulness for case assessment, calibrated trust, and actionability directly. AFIE's mechanistic traceability and decision reconstructibility should be treated as properties of the derivation, not as proxies for human explanation

quality. This stage is therefore required before claims about operational interpretability or investigator benefit are made.

12.5. Deployment evaluation

Deployment evaluation is the final stage, not a substitute for the preceding evidence. Only after frozen confirmation, independent validation, parameter robustness, and investigator evaluation should AFIE be assessed in an operational setting for prospective error rates, alert volume, workflow effects, latency, drift, calibration maintenance, audit burden, and consequential investigator outcomes. A deployment study should preserve the same evidence-state and provenance controls used in the research evaluator so that operational adaptation can be distinguished from changes to the underlying inference mechanism.

Emergence detection remains a separate open question. The present 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 unless a future study designed specifically for that hypothesis supports the claim.

13. Conclusion

AFIE shows, within this diagnostic environment, 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

Both authors are employees of Rockland Trust Company. The authors declare no other competing interests.

Data Availability

The frozen run package retains the 160-case corpus (corpus generation 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 artifact. The data supporting the findings of this study, together with the retained reproducibility artifacts 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. [a], [b] *Ngai 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.*
2. [a], [b] *Al-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.*

3. ^Dal 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.
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.
5. ^Oztas 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.
6. a, bDou Y, Liu Z, Sun L, Deng Y, Peng H, Yu PS (2020). "Enhancing Graph Neural Network-Based Fraud Detectors Against Camouflaged Fraudsters." In Proceedings of the 29th ACM International Conference on Information & Knowledge Management. pp. 315–324. doi:10.1145/3340531.3411903.
7. ^Baesens B, Höppner S, Verdonck T (2021). "Data Engineering for Fraud Detection." Decis Support Syst. **150**:113492. doi:10.1016/j.dss.2021.113492.
8. ^Chandola V, Banerjee A, Kumar V (2009). "Anomaly Detection: A Survey." ACM Comput. Surv.. **41**(3):1–58. doi:10.1145/1541880.1541882.
9. ^Motie 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.
10. ^Van Belle R, Van Damme C, Tytgat H, De Weerdt J (2022). "Inductive Graph Representation Learning for Fraud Detection." Expert Syst Appl. **193**:116463. doi:10.1016/j.eswa.2021.116463.
11. ^Van Belle R, Baesens B, De Weerdt 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.
12. ^Chen 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.
13. ^Lou 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.
14. ^Tang Y, Li S, Fang Z, Sun J (2024). "Credit Card Fraud Detection Based on Federated Graph Learning." Expert Syst Appl. **256**:124979. doi:10.1016/j.eswa.2024.124979.
15. ^Suárez-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.

16. ^a, ^b*Aitken CG, Taroni F (2004). Statistics and the Evaluation of Evidence for Forensic Scientists. 2nd ed. John Wiley & Sons. doi:10.1002/0470011238.*

17. ^*Shafer G (1976). A Mathematical Theory of Evidence. Princeton University Press. doi:10.2307/j.ctv10vm1qb.*

18. ^*Pollock JL (1987). "Defeasible Reasoning." Cogn Sci. **11**(4):481–518. doi:10.1207/s15516709cog1104_4.*

19. ^*Modgil S, Prakken H (2013). "A General Account of Argumentation with Preferences." Artif Intell. **195**:361–397. doi:10.1016/j.artint.2012.10.008.*

20. ^*Pandzic S (2022). "A Logic of Defeasible Argumentation: Constructing Arguments in Justification Logic." Argument & Computation. **13**(1):3–47. doi:10.3233/aac-200536.*

21. ^a, ^b*Kruegel C, Robertson W (2004). "Alert Verification: Determining the Success of Intrusion Attempts." In U. Flegel & M. Meier (Eds.), 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.*

22. ^a, ^b*Kruegel C, Robertson W, Vigna G (2004). "Using Alert Verification to Identify Successful Intrusion Attempts." Prax Infverarb Kommun [Practice of Information Processing and Communication]. **27**(4):220–228.*

23. ^*Bakry 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.*

24. ^*Lin 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.*

25. ^*Rudin 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.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.