GaleStorm™: Deterministic, Privacy-Preserving AI Oversight for Digital Assets and Healthcare Compliance

Darrell L. Young, P.E.[1,2], Jason Teske[2,3], Mitchell A. Thornton[1]
[1]Darwin Deason Institute of Cybersecurity, Southern Methodist University, Dallas, TX 75205, USA
[2]Wave 3 LLC, 2555 Holly Manor Dr., Falls Church, VA 22043, USA
[3]George Washington University, Washington, DC, USA

ABSTRACT

AI drives millions of decisions in healthcare ($4T annually) and finance ($3T+ digital assets), demanding approaches to assure autonomous behavior. GaleStorm™ introduces a framework where independent AI agents audit other AI systems under deterministic execution and economic consequences. Validators execute identical policies in a Deterministic Policy Virtual Machine (DPVM), producing Scrub-Proof Objects (SPOs)—audit receipts targeting sub-millisecond verification while preserving privacy through split-Merkle commitments and zero-knowledge proofs. Byzantine consensus with GALE staking creates challenge windows where disputes trigger slashing penalties and bounties, transforming oversight into Nash equilibrium. Specified as Medical Billing Scrubbers (MBS) and Digital Transaction Scrubbers (DTS), demonstrating generalizability across HIPAA and AML domains. SEAGULL integration enables on-premises validation. Evaluation will measure determinism across TEEs, red-team validation on SMU's Cyber Autonomy Range, and conformance with GENIUS Act §9(a) and NIST AI RMF.

Keywords: AI auditing AI, AI guardrails, deterministic policy VM, scrub-proof objects, zero-knowledge proofs, HIPAA, AML/KYC, Byzantine consensus, GENIUS Act, ChatPack Specification

1. INTRODUCTION: THE AI VELOCITY PROBLEM

1.1 Scale, Trust, and the Recursive Problem

The United States healthcare system processes approximately $4 trillion annually through 14 billion insurance claims—roughly 450 claims per second.[1] Digital asset markets exceed $3 trillion in capitalization, with AML systems evaluating millions of daily transactions.[2] Human oversight cannot scale to match AI operational velocity: covering just 1% of annual healthcare claims at two minutes each would require over 12,000 full-time auditors.

If humans cannot audit AI at scale, the only viable solution is AI auditing AI—which raises the recursive question: who audits the auditors? Risks include collusion among auditing agents, bias amplification from shared architectures, governance capture by dominant providers, and the privacy paradox where auditing needs the transaction details that regulations restrict.

1.2 Design Thesis

AI-checking-AI becomes trustworthy through four principles: **deterministic reproducibility** (decisions are bit-for-bit replayable), **cryptographic binding** (inputs, models, environments, and outputs are linked immutably), **privacy preservation** (verification without exposing protected data via ZK), and **economic accountability** (real financial consequences deter misbehavior). A fifth property emerges from the architecture: **symmetrical opacity**, where neither the AI agent nor the human operator can selectively deviate from committed policy without producing detectable cryptographic evidence of divergence.

Further author information: (Send correspondence to Darrell L. Young)
E-mail: dlyoung@smu.edu

1.3 Contribution

GaleStorm™ (Governance-Aligned, Logically Enforced Storm) operationalizes AI-auditing-AI with cryptographic accountability. We demonstrate the architecture in two radically different domains: Medical Billing Scrubbers (MBS) for HIPAA-constrained claims validation and Digital Transaction Scrubbers (DTS) for asset issuance and market integrity. Cross-domain validation establishes that the core patterns generalize to any AI system requiring accountable oversight. A companion policy brief[6] frames these mechanisms for policymakers; a companion paper on Impossibility Engineering[7] formalizes the assumption inversions underlying the design.

1.4 Assumption Inversions

The recursive trust problem persists because conventional approaches assume observed behavior represents true behavior—a frame-lock that renders alignment faking invisible.[7]

Frame-Locked Assumption	**GaleStorm Inversion**
AI self-reports reflect actual behavior	Cryptographic execution traces (DPVM+SPO)
Evaluation matches deployment	Attestation covers all contexts (TEE)
Single auditor integrity suffices	Byzantine consensus, diverse agents (ADER)
Verification requires full disclosure	Graduated proofs, warrant-gated access
Challengers emerge where fraud exists	Market design with Anchor participants

2. BACKGROUND: PARALLEL FAILURES

Healthcare claims "scrubbing" and digital asset compliance share five structural failures: unverifiable decisions (providers and token holders cannot confirm rule application), non-determinism (identical inputs produce different verdicts), privacy–audit conflict (auditing needs the details regulations restrict), single points of failure (centralized scrubbers and oracles), and misaligned incentives (paid per volume, favoring throughput over accuracy). Healthcare fraud costs \$60–100B annually;[3] digital asset markets have lost \$30B+ since 2020 through rug pulls, wash trading, and governance capture.[4,5]

Both domains require verifiable, privacy-preserving, economically aligned auditing at AI velocity. GaleStorm provides detection infrastructure; future prevention-first architectures (Sec. 8) would block violations at execution time.

3. GALESTORM ARCHITECTURE

3.1 Overview

GaleStorm replaces centralized validators with economically incentivized, cryptographically accountable participants. Seven distinct roles interact through deterministic execution, privacy-preserving evidence, and bounded economic consequences. Figure 1 illustrates the complete data flow using a healthcare billing example; Table 1 summarizes each role's function and incentive structure.

Each role addresses a specific failure mode of centralized oversight: Scrubbers prevent subjective evaluation, Verifiers prevent single-point corruption, Challengers prevent negligent approval, Evaluators prevent over-disclosure during disputes, Governors prevent unilateral policy capture, and Anchor Challengers prevent market oscillation when per-case bounties are insufficient to sustain continuous monitoring.

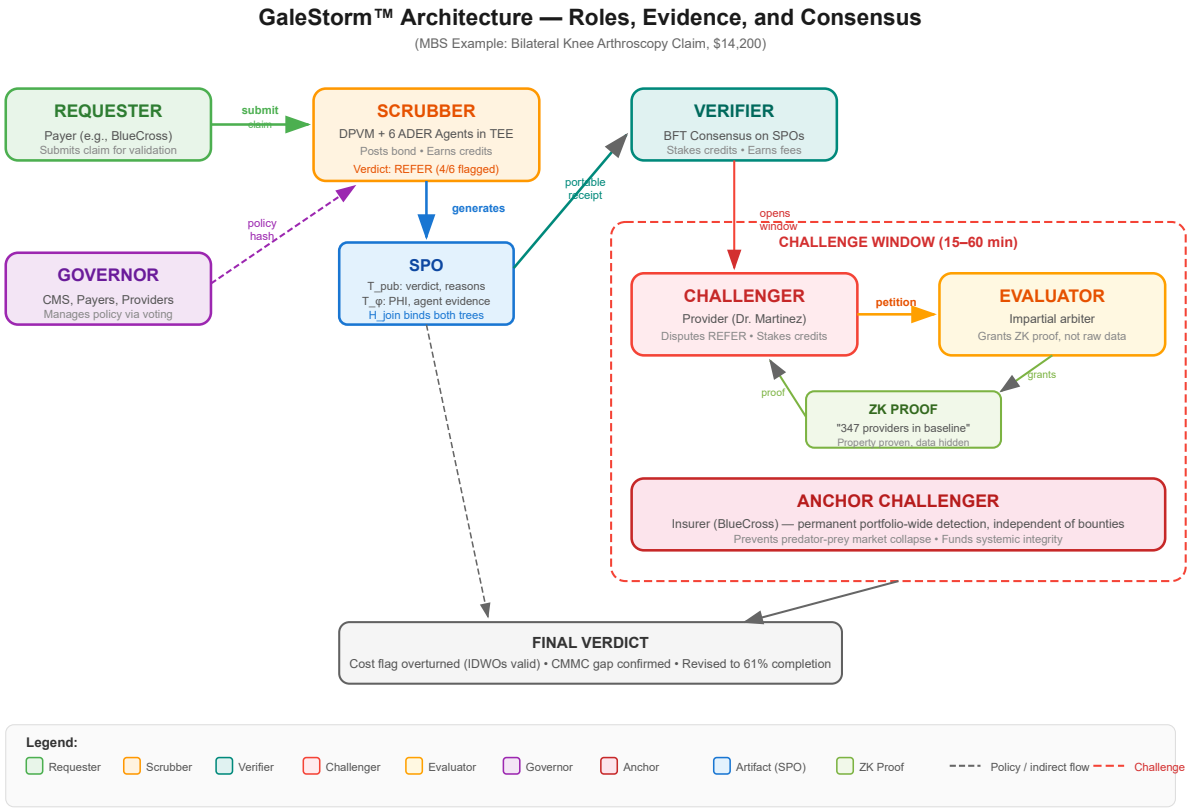

Figure 1. GaleStorm architecture (MBS example): A payer (Requester) submits a bilateral knee arthroscopy claim. A Scrubber executes DPVM policy checks and six ADER agents in a TEE, generating an SPO with split-Merkle privacy (T_{pub} for the public verdict, T_ϕ for private evidence, bound by H_{join}). Verifiers reach BFT consensus, opening a challenge window where the provider (Challenger) petitions an Evaluator for ZK-verified deeper-tier access. An Anchor Challenger provides permanent portfolio-wide detection. Governors manage policy evolution via multi-stakeholder voting. Color coding matches Table 1.

Table 1. GaleStorm participant roles and incentive alignment

Role	Function	Incentive
Requester	Submits validation tasks	Accurate verdicts
Scrubber	Executes policy in TEE, posts bond	Earns credits for accurate work
Verifier	BFT consensus on SPOs	Earns fees for verification
Challenger	Submits fraud-proofs in windows	Bounty from slashed bonds
Evaluator	Rules on deeper-tier access petitions	Impartial adjudication
Governor	Manages policy evolution via voting	Stakeholder representation
Anchor	Permanent detection infrastructure	Systemic integrity

Illustrative Flow (MBS). A payer (Requester) submits a $14,200 bilateral knee arthroscopy claim. A Scrubber executes DPVM policy checks and six ADER agents in a TEE; four agents flag concerns and the ensemble

verdict is `REFER`. Evidence binds into an SPO with private evidence (PHI, agent records) in T_ϕ and the public verdict in T_{pub}, joined by H_{join}. Verifiers reach BFT consensus, opening a challenge window. The provider (Challenger) disputes the finding and petitions the Evaluator for deeper-tier access. The Evaluator grants a ZK proof that the provider profiling baseline used 347 regional comparators—without revealing their identities. The insurer operates as Anchor Challenger, running continuous portfolio-wide anomaly detection independent of per-case bounties. Governors manage policy updates when CMS revises coding rules, requiring multi-stakeholder approval before new DPVM bytecode deploys.

3.2 Deterministic Policy Virtual Machine (DPVM)

DPVM executes compiled, fixed-branching bytecode ensuring identical inputs yield identical outputs. Four properties enforce determinism: (1) fixed-precision arithmetic, (2) deterministic branching, (3) hermetic execution, and (4) versioned bytecode. Each run outputs a verdict tuple:

$$V = \{\texttt{policy\_hash}, \texttt{verdict}, \texttt{reasons}, \texttt{ts}, \texttt{proof\_root}\} \tag{1}$$

3.3 Scrub-Proof Objects (SPO)

SPOs pack cryptographic evidence into fixed-format receipts verifiable in $\mathcal{O}(1)$ time (measured: 0.003–0.007 ms). Components include policy commitment hash, payload tree (PHI/proprietary), public tree (verdict, attestations), join-hash binding both trees, TEE certificate, ADER references, and validator signatures (Fig. 2).

Attestation records are structured as invocation receipts within the ChatPack container specification, enabling end-to-end audit trails from query ingestion through tool invocation to response delivery. The container's constant-shape attestation receipt format enables batch verification without per-record schema resolution.

3.4 Split-Merkle: Tiered Evidence Architecture

Evidence commits to a compliance depth lattice: breadth domains (Financial, Operational, Identity, Temporal) crossed with sensitivity tiers (0=Summary, 1=Detail, 2=Sensitive, 3=Raw). A public tree T_{pub} contains Tier 0; a private tree T_ϕ contains Tiers 1–3. Join-hash H_{join} binds both without revealing private contents (Fig. 2).

Three complementary mechanisms enable privacy-preserving accountability (Fig. 3):

Split-Merkle (what is visible). Structures evidence into tiered access levels. Anyone can verify the public tree and confirm that the private tree exists and is cryptographically bound—without seeing private contents. Unlike encryption, which is binary (possess the key or not), Split-Merkle provides graduated visibility across sensitivity tiers.

Zero-knowledge proofs (what is provable). ZK branch proofs allow authorized parties to verify specific properties at deeper tiers without full disclosure. In the MBS domain, a challenger can verify that a provider profiling baseline used 347 regional comparators without learning their identities or billing patterns. This operationalizes HIPAA's minimum-necessary standard.

Warrant model (who may ask, and when). This paper's primary contribution to graduated disclosure. Challengers petition Evaluators with specific grounds for deeper-tier access. Evaluators determine *what form* evidence takes—a ZK proof of methodology rather than raw data—balancing the challenger's right to dispute against protection of third-party information. Escalating economic stakes prevent fishing expeditions: each tier requires additional bonded GALE credits, forfeited if the challenge fails. This transforms privacy from a static access-control decision into a dynamic, evidence-based, economically bounded process (Fig. 3, bottom).

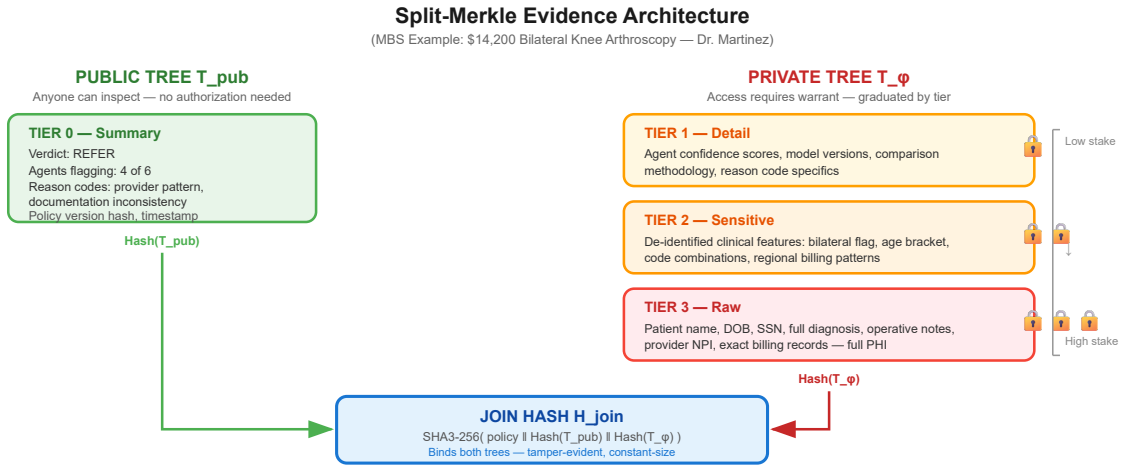

Figure 2. Split-Merkle evidence architecture: Public tree T_{pub} holds Tier 0 verdicts visible to all parties. Private tree T_{ϕ} holds Tiers 1–3 with escalating sensitivity and access cost. Join-hash H_{join} cryptographically binds both trees, enabling any party to verify that private evidence exists and is untampered—without seeing its contents.

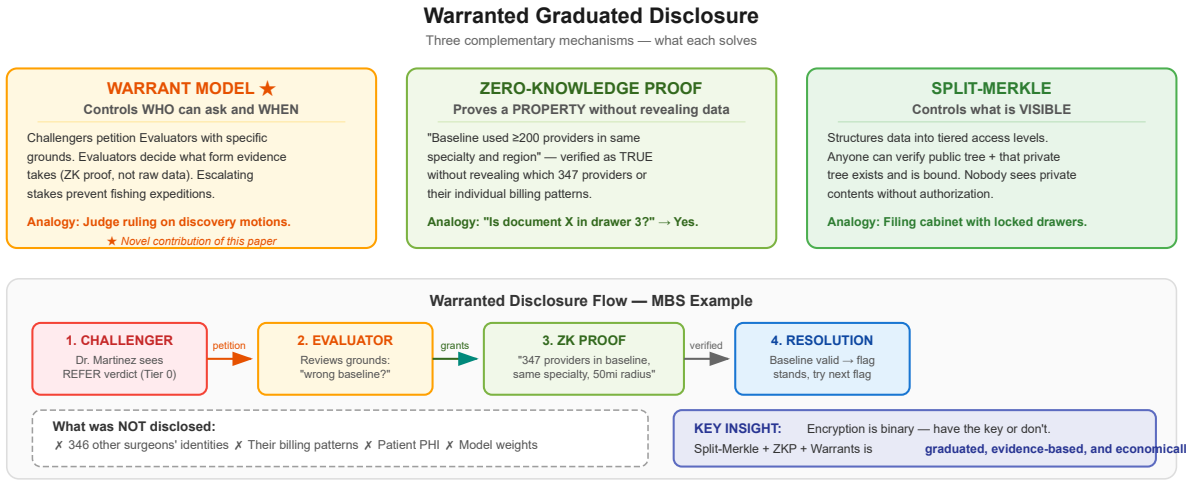

Figure 3. Warranted graduated disclosure: Three mechanisms address distinct privacy questions. Split-Merkle controls *what* is visible; ZK proofs verify *properties* without revealing data; the warrant model—this paper's primary contribution—governs *who* may request deeper access and under what *economic constraints*. Bottom: Dr. Martinez (Challenger) obtains a ZK proof of baseline methodology through an Evaluator without exposing third-party provider data.

3.5 ADER: AI-Derived Evidence Records

Modern healthcare and financial systems increasingly delegate high-volume decisions to AI: auto-adjudication engines process claims, algorithmic trading systems approve transactions, and automated underwriting evaluates risk. GaleStorm's central architectural claim is that these AI-driven decisions require AI-driven auditing—human reviewers cannot match the volume, speed, or consistency required. ADER operationalizes this principle: independent AI agents, executing in TEEs with cryptographically bound model identity, audit the outputs of other AI systems. The auditors and the audited share no weights, no training data, and no organizational affiliation.

Multi-Agent Architecture. MBS deploys six agents against payer auto-adjudication decisions: coding validation, medical necessity, provider profiling, cross-provider correlation, documentation analysis, and ensemble

synthesis. DTS deploys seven against exchange transaction approvals: supply chain, market microstructure, graph analysis, smart contract audit, sentiment, liquidity, and ensemble synthesis. Each agent commits an evidence record:

$$E = \{\texttt{model\_id}, \texttt{version}, \texttt{input\_hash}, \texttt{output\_hash}, \texttt{confidence}, \texttt{reasons}, \texttt{TEE\_quote}\} \tag{2}$$

Ensemble Aggregation. Agent verdicts aggregate via stake-weighted voting: $S = \sum_i w_i \cdot s_i$. Weights update with measured performance. Diversity requirements mandate at most two agents per legal entity, minimum architectural distance, and geographic distribution.

Privacy Preservation. Healthcare agents receive de-identified features (billing velocity, code patterns) not raw PHI. Finance agents analyze anonymized graph topologies. Raw data never leaves TEEs.

Edge Deployment. SEAGULL[8] routes validation between edge SLMs and cloud LLMs based on complexity. Routine validation—85–95% of volume—executes on-premises at 5–20 ms with PHI remaining local. Complex cases escalate. Escalation rates themselves provide anomaly signals. On-premises deployment enables participation by small and mid-sized billing operations—a certified professional biller operating edge scrubber hardware can validate claims locally without cloud dependency, reducing both cost and data exposure. Physical compromise of edge equipment is mitigated by TEE attestation (tampered devices fail quote verification), DPVM determinism (modified bytecode produces detectably divergent outputs), staking bonds (operators bear economic loss for compromised verdicts), and credential revocation upon reported loss.

Policy Confidentiality. In fraud detection deployments, the validation policy itself constitutes sensitive intelligence: suspect lists, behavioral pattern signatures, and investigation thresholds whose exposure would compromise ongoing enforcement and damage reputations of not-yet-adjudicated subjects. GaleStorm inverts the typical compliance transparency assumption. The DPVM executes policy bytecode exclusively within the TEE; the `policy_hash` in each SPO commits to the executed version without revealing its contents. Verifiers confirm deterministic execution of a specific policy version without accessing the policy itself. On edge deployments, ChatPack containers holding fraud detection patterns are encrypted via TEE sealed storage, binding decryption to the specific enclave and device hardware. Device operators process claims and receive verdicts but cannot inspect the detection logic executing inside the enclave. Policy updates propagate only to devices passing current attestation, and credential revocation immediately halts further pattern distribution to compromised nodes. This ensures that even the participants operating GaleStorm infrastructure cannot extract the intelligence their own equipment enforces

Symmetrical Opacity. This property—that the operator of auditing infrastructure cannot inspect the auditing logic their own equipment enforces—does not exist in current compliance systems. Traditional auditors necessarily know what they are looking for; audit criteria transparency is considered a prerequisite for accountability. GaleStorm inverts this assumption. The TEE knows what to look for; the human operator knows only the verdict. Accountability is preserved not through transparency of criteria but through deterministic execution, cryptographic commitment to a specific policy version, and economic consequences for incorrect verdicts. The operator need not be trusted with intelligence in order to be trusted with enforcement.

Connection to AI Alignment. This architectural property directly addresses concerns raised by Amodei[14] regarding AI systems that behave correctly during evaluation but deviate during deployment—so-called alignment faking. In GaleStorm, the distinction between evaluation and deployment collapses: every execution is attested, every verdict is committed, and every policy version is cryptographically bound. The same infrastructure that prevents AI auditors from deceiving their verifiers also prevents human operators from gaming audit criteria. Opacity is symmetrical—neither the AI agent nor the human operator can selectively deviate from committed policy without producing detectable cryptographic evidence of divergence.

3.6 Audit Attestation Format

SPOs produced by different scrubbers, across different policy versions and domains (MBS and DTS), share a constant-shape attestation receipt format derived from the ChatPack container specification.[9] The receipt format is public and standardized even when the policy that generated the verdict is confidential—verifiers validate the shape and cryptographic bindings of the SPO without requiring access to the sealed policy bytecode. Verifiers need not resolve per-record schemas; batch verification becomes a single-pass operation over a homogeneous receipt stream. This format also captures Clean Signal Controls—causal ladder enforcement for phased corrective actions, holdout-based experiment design for measuring intervention effectiveness, and automated stop rules for halting ineffective processes—ensuring that corrective decisions within the auditing pipeline are themselves attested and auditable.

4. INCENTIVES: GALE CREDIT ECONOMICS

GALE credits meter verification work and align incentives across all participants. GALE functions as platform currency—analogous to airline miles or SaaS credits—earned through useful work and consumed for network services. Scrubbers lock credit bonds $B \geq \alpha \cdot V$ (value-at-risk), scaling entry costs proportionally to claim volume: a small billing operation processing 500 claims per month bonds less than a national clearinghouse processing 500,000. Verifiers stake to participate. Challengers post challenge bonds.

Slashing and Bounties. Successful fraud-proofs slash scrubber bonds: $S_{\text{penalty}} = \beta \cdot B \cdot f(\text{severity})$, with bounty $\gamma \cdot S_{\text{penalty}}$ to challengers. With B=5000, β=0.25, γ=0.50, break-even precision is $p^* \approx 27\%$—achievable for statistical outlier detection.

Tiered Challenge Bonds. Challenge bonds escalate with disclosure depth (Sec. 3.4). A Tier 0 challenge—disputing the public verdict—requires base bond C_0. Petitioning the Evaluator for deeper-tier access requires incremental bonds: $C_k = C_0 \cdot \delta^k$ for tier k, where $\delta > 1$ controls escalation steepness. Bonds at all tiers are forfeited if the challenge fails and returned with bounty if it succeeds. This creates an economic filter: challengers with weak grounds self-select out at lower tiers, and only evidence-backed disputes reach sensitive data. The Evaluator's role is to assess whether articulated grounds justify the economic and privacy cost of deeper access—not to adjudicate the underlying claim.

Market Stability. Simulation reveals pure bounty markets exhibit predator-prey oscillation: fraud reduction drives Challengers to exit, enabling fraud resurgence. Stability requires two mechanisms. First, an Anchor Challenger—the primary beneficiary (insurer, platform) operates permanent detection regardless of short-term profitability. Second, a Subscription Model—participants pay modest fees funding infrastructure independent of bounties. With proper parameterization, honesty becomes the dominant strategy—a Nash equilibrium where no participant benefits from unilateral deviation.

5. DOMAIN INSTANTIATIONS AND GENERALIZATION

5.1 Medical Billing Scrubber (MBS)

Context. Claims contain PHI; fraud costs \$60–100B annually. Inputs include ICD-10, CPT/HCPCS codes, modifiers, and provider credentials. Policy sources: LCD/NCD, NCCI edits, MUE, payer rules. Outputs: `APPROVE|DENY|REFER` with reason codes. The bilateral knee arthroscopy scenario in Figs. 1–3 illustrates the complete MBS flow from claim submission through warranted challenge resolution.

Market Sustainability. Insurance companies internalize fraud losses, creating concentrated beneficiaries who fund Challenger infrastructure. Anchor challengers (major payers) provide continuous deterrence. Provider subscriptions (\$30–50/month) fund competitive detection.

Table 2. Domain comparison: beneficiary structure and funding models

Dimension	MBS	DTS (Detect)	DTS (Prevent)
Fraud occurs?	Yes	Yes	No (blocked)
Concentrated payer	Insurers	None	Asset issuer
Challengers are	Essential	Essential	Premium

5.2 Digital Transaction Scrubber (DTS)

Context. Markets suffer wash trading, undisclosed minting, and rug pulls—$30B+ in losses since 2020. Inputs: token operations, supply data, holder distribution, liquidity states. Outputs: `COMPLIANT|VIOLATION|SUSPICIOUS` with reason codes.

Market Sustainability: The Issuer-as-Beneficiary Model. Unlike healthcare, where insurers internalize fraud losses and thus fund detection, digital asset holders are dispersed and often pseudonymous. However, the tokenized economy creates a different concentrated beneficiary: the *token issuer*. Organizations issuing tokens—whether for loyalty programs, creator economies, or asset-backed instruments—bear direct reputational and regulatory consequences when unauthorized tokens circulate or sanctioned actors transact using their infrastructure. GaleStorm serves these issuers as premium compliance infrastructure: continuous monitoring, cryptographic audit trails, and warranted challenge resolution that protect brand integrity and satisfy GENIUS Act transparency requirements. Additional funding sources—exchange listing fees, transaction levies, and protocol treasuries—provide supplementary revenue, but the issuer subscription model provides the stable economic base that pure bounty markets lack.

Prevention-First Architecture. The funding problem dissolves under prevention-first design. If digital assets are issued through infrastructure enforcing policy at execution time—where every operation is validated against committed policy before state changes propagate—violations cannot occur. Issuers pay for compliant infrastructure. Challenger services become premium add-ons for enterprise customers wanting audit trails. Prevention is the product; detection is premium.

Privacy-Preserving Analytics. Just as detection policy is sealed within the TEE to protect investigation integrity (Sec. 3.5), transaction data is protected from aggregation. GaleStorm provides aggregate analytics (volume, velocity, sector distribution) without per-wallet tracking, cross-platform correlation, or data sales. Infrastructure detecting fraud must not become surveillance infrastructure.

5.3 Generalizing Beyond Compliance Domains

While MBS and DTS demonstrate GaleStorm in specialized regulated domains, the architecture addresses a broader problem: how do organizations verify that AI systems making consequential decisions actually followed committed policy? This question applies to chatbots, content generators, recommendation engines, autonomous agents, and any AI deployment where "the system says it behaved correctly" is insufficient assurance.

Four architectural patterns transfer directly.

Pattern 1: Multi-Agent Verification. Rather than a single guardrail model—itself an AI system requiring trust—deploy an ADER-style ensemble where independent agents with different architectures, training data, and organizational affiliations audit the primary system's outputs. The auditors and the audited share no weights. When agents disagree, the system escalates rather than silently resolving ambiguity. Diversity requirements prevent correlated failures that single-model guardrails cannot detect.

Pattern 2: Deterministic, Opaque Policy. Encode rules as versioned DPVM bytecode producing identical verdicts from identical inputs—essential for debugging, compliance, and dispute resolution. The Symmetrical Opacity property (Sec. 3.5) generalizes: operators deploying AI safety infrastructure need not—and should not—have visibility into detection logic that adversaries could reverse-engineer. A content moderation system whose evasion patterns are sealed in a TEE is harder to circumvent than one whose rules are visible in application code.

Pattern 3: Cryptographic Audit Receipts. Every decision generates an SPO-style receipt binding input hash, policy version, agent verdicts, and timestamp. Receipts verify in constant time (<1 ms). When regulators ask "how did your system approve this?" the receipt provides cryptographic proof of exactly what occurred—eliminating the distinction between evaluation and deployment behavior that enables alignment faking.

Pattern 4: Warranted Escalation. When a decision is disputed, the warrant model (Sec. 3.4) provides graduated access to supporting evidence without blanket disclosure. A user disputing a content moderation decision can verify that the policy was applied deterministically and that the ensemble reached consensus—without exposing detection patterns, other users' data, or proprietary model internals. This replaces the current binary: either the platform reveals nothing (eroding trust) or reveals everything (enabling evasion).

Implementation Spectrum. Full GaleStorm (Byzantine consensus, TEE attestation, economic staking) suits regulated domains where adversarial pressure and legal consequences demand cryptographic assurance. For typical AI deployments, a simplified stack captures most of the value:

Component	Full	Lite
Multi-agent verification	Required	Required
Deterministic policies	Required	Recommended
Audit receipts	Cryptographic SPOs	Structured logs
Edge routing	SEAGULL	Confidence threshold
Warranted disclosure	GALE-bonded	Role-based
Policy confidentiality	TEE-sealed	Encrypted at rest

MBS and DTS prove the architecture works in radically different regulated domains with different beneficiary structures, different privacy constraints, and different economic dynamics. The core contribution—verifiable AI oversight with symmetrical opacity, graduated disclosure, and economic accountability—extends to any AI system where the consequences of undetected deviation exceed the cost of attestation.

6. PERFORMANCE MEASUREMENTS

Validation experiments were conducted on two platforms: consumer hardware (Intel Core Ultra 5, RTX 5050 8GB; Intel i7-12700H, RTX 3060 6GB) and data center infrastructure (AMD EPYC 7742 64-Core, NVIDIA A100-SXM4-80GB, SMU SuperPod). ADER agents used SmolLM2 1.7B across all configurations to ensure comparable results.

6.1 Cryptographic Operations

SPO verification and DPVM execution are CPU-bound operations independent of GPU infrastructure. Table 3 confirms sub-millisecond performance on both platforms.

Table 3. Cryptographic operation latency (10,000 iterations per platform)

Metric	Core Ultra 5 (RTX 5050)	i7-12700H (RTX 3060)	EPYC 7742 (A100-80GB)	Bit-identical (cross-platform)
SPO Verification	0.003 ms	0.003 ms	0.007 ms	—
DPVM Execution	0.016 ms	0.023 ms	0.009 ms	✓

SPO verification confirms $\mathcal{O}(1)$ constant-time performance: hash comparison and Merkle proof checking on fixed-size structures. Variation between platforms reflects single-thread clock speed differences (Intel boosts to 4.7 GHz vs. EPYC at 3.4 GHz), not algorithmic complexity; cryptographic verification latency is hardware-invariant at operationally negligible levels, confirming that audit infrastructure imposes no meaningful overhead regardless of deployment platform. DPVM produced bit-identical outputs across 10,000 iterations on both Intel and AMD architectures, validating the deterministic execution claim across ISA families.

6.2 SPO Throughput

Sustained SPO generation on the A100 achieved 38,351 SPOs/sec—85× the healthcare velocity requirement of 450 claims/sec (Sec. 1). This confirms that cryptographic audit infrastructure adds negligible latency to the operational pipeline; accountability is effectively free at the margin.

6.3 ADER Ensemble: Inference Server Architecture Matters

A critical finding emerged from ADER benchmarking: inference server selection determines whether multi-agent parallelism is realized or merely queued. Table 4 compares Ollama (per-request model instantiation) against vLLM (continuous batching with PagedAttention[13]) on identical hardware with identical models.

Table 4. ADER ensemble latency: Ollama vs. vLLM (SmolLM2 1.7B, A100-80GB)

Configuration	Ollama (ms)	vLLM (ms)	Speedup
3-agent parallel	518	280	1.9×
3-agent sequential	894	683	1.3×
6-agent parallel (MBS)	1,006	347	2.9×
6-agent sequential	1,747	1,212	1.4×
7-agent parallel (DTS)	1,015	357	2.8×
7-agent sequential	2,163	1,324	1.6×

Ollama parallel latency scales linearly with agent count (518 ms to 1,015 ms for 3 to 7 agents), confirming sequential request processing. vLLM parallel latency increases only 77 ms from 3 to 7 agents (280 ms to 357 ms), demonstrating true concurrent execution through continuous batching.

This finding has architectural implications for ADER deployment: multi-agent verification requires inference infrastructure that supports concurrent execution. Production deployments must use batching-capable serving frameworks; otherwise, the parallelism designed into the ADER ensemble is negated at the infrastructure layer. Our companion C8-ISR paper[8] reports consistent scaling behavior in search-and-rescue multi-agent configurations on the same infrastructure.

6.4 TEVV Plan

Remaining validation includes adversarial red-teaming on SMU's Cyber Autonomy Range, HIPAA-aligned privacy audits, TEE side-channel evaluation, and game-theoretic economic simulation of challenger market dynamics.

7. REGULATORY ALIGNMENT

Table 5 maps GaleStorm components to provisions in enacted and pending legislation. GaleStorm's architecture was not designed to satisfy any single framework; rather, the same properties—deterministic reproducibility, cryptographic attestation, graduated disclosure, and economic accountability—address requirements across jurisdictions because the underlying governance failures are universal.

Treasury's ANPRM on GENIUS Act implementation (September 2025) and the pending CLARITY Act Senate markup describe precisely the compliance verification infrastructure GaleStorm provides. SMU Darwin Deason Institute has offered Cyber Autonomy Range resources for Treasury-aligned red-team testing,[11,12] positioning GaleStorm as a testable reference implementation under active federal rulemaking.

Table 5. Regulatory alignment of GaleStorm components

Framework	Requirement	GaleStorm Component
GENIUS Act	Reserve auditability	ZK proof of reserves
(P.L. 119-XX, 2025)	Explainability	Structured reason codes
	Continuous monitoring	Warrant-gated challenges
CLARITY Act	Anti-fraud evidence	SPO audit receipts
(H.R. 3633, pending)	Proprietary protection	Graduated disclosure
EO 14179 + Dec. 2025	Accountability without	Symmetrical Opacity:
National Framework	regulatory burden	cryptographic proof, not compliance paperwork
NIST AI RMF	Govern / Map / Measure / Manage	Policy namespaces / Roles DPVM replay / Challenges
EU AI Act	High-risk AI documentation, oversight,	SPO documentation, Evaluator role, ADER
(effective 2025)	robustness, transparency	diversity, warrants

8. LIMITATIONS AND FUTURE DIRECTIONS

Current Limitations. TEE side-channel vulnerabilities remain an active research area; GaleStorm's Symmetrical Opacity property depends on enclave integrity that hardware-level attacks could compromise. Zero-knowledge proofs create inherent tension between explainability and privacy—the warrant model mitigates but does not eliminate this tension, as Evaluators must balance disclosure depth against investigation integrity using judgment that is itself difficult to audit. Stake-weighted governance risks plutocratic capture in domains where participant resources are highly asymmetric; the Anchor Challenger mechanism stabilizes markets but concentrates structural power in primary beneficiaries. Cross-platform bit-identity of DPVM outputs has been validated across Intel and AMD architectures (Sec. 6), but formal verification of determinism across all target ISAs remains incomplete. Finally, the economic parameters (α, β, γ, δ) presented in Sec. 4 derive from simulation rather than live market data; real-world calibration requires production deployment.

Prevention-First Architecture. Detection is necessary but insufficient. If policy violations can be blocked at execution time rather than discovered after the fact, the threat model changes fundamentally. Future work will develop Execution Firewalls and Policy Firewalls that gate state changes on policy satisfaction—making non-compliant transactions impossible rather than detectable. Under this architecture, Challengers shift from fraud detection to premium audit services, and the market sustainability problem (Sec. 4) dissolves: issuers pay for compliant infrastructure rather than funding adversarial detection.

Formal Methods and Cryptographic Extensions. DPVM determinism claims currently rest on empirical bit-identity testing. Formal verification using Coq or Lean 4 would provide mathematical guarantees that identical inputs produce identical outputs across all execution paths. Recursive SNARKs (Halo2, Nova) would enable composable proofs—a verifier confirming one SPO could chain that verification into a portfolio-level proof without re-examining individual claims. Post-quantum TEE attestation will be necessary as quantum computing threatens current cryptographic assumptions.

Edge Deployment and Accessibility. SEAGULL-based edge deployment (Sec. 3.5) enables small and mid-sized operators to participate as Scrubbers using on-premises hardware. Future work includes optimizing DPVM

bytecode for resource-constrained edge devices, developing federated cross-domain learning that improves detection without centralizing sensitive data, and establishing certification pathways for edge operators analogous to existing professional credentialing (e.g., certified professional billers operating MBS scrubbers).

Regulatory Sandbox. SMU's Cyber Autonomy Range provides controlled infrastructure for adversarial validation. Planned work includes red-team testing of ADER ensemble robustness, game-theoretic simulation of GALE credit markets under adversarial conditions, and demonstration deployments aligned with Treasury's GENIUS Act implementation rulemaking and pending CLARITY Act requirements.

9. CONCLUSION

Human oversight cannot match AI velocity. When autonomous systems process millions of healthcare claims and financial transactions per day, accountability infrastructure must operate at the same speed and scale—or it provides no assurance at all.

GaleStorm answers with four architectural commitments: deterministic reproducibility (identical inputs produce identical, verifiable outputs), cryptographic binding (every decision generates a tamper-evident receipt), privacy preservation (graduated disclosure reveals only what is warranted), and economic accountability (participants bear financial consequences for incorrect verdicts).

Four contributions emerged that were not obvious at the outset.

First, *detection requires market design, not just technology.* Challenger markets are not self-sustaining; simulation reveals predator-prey oscillation where successful fraud reduction drives challengers to exit, enabling fraud resurgence. Anchor participants and subscription models are architectural requirements, not optional enhancements.

Second, *the operator of auditing infrastructure should not see the auditing logic.* Symmetrical Opacity—where neither the AI agent nor the human operator can selectively deviate from committed policy without producing detectable cryptographic evidence—is a property that does not exist in current compliance systems. It directly addresses alignment faking concerns: the distinction between evaluation and deployment behavior collapses when every execution is attested.

Third, *privacy and accountability are not in tension; they are architecturally complementary.* Split-Merkle commitments structure evidence into tiered access levels. Zero-knowledge proofs verify properties without revealing data. The warrant model—this paper's primary contribution—governs who may request deeper access, under what economic constraints, and in what evidentiary form. Together, these mechanisms replace binary access-control decisions with graduated, evidence-based, economically bounded disclosure.

Fourth, *inference server architecture determines whether multi-agent verification is real or illusory.* Benchmarking on NVIDIA A100 hardware demonstrated that Ollama's sequential request processing negates ADER's parallel design, while vLLM's continuous batching achieves 2.8× speedup at seven-agent concurrency—completing a full DTS ensemble in 357 ms. Production deployment of multi-agent AI auditing requires infrastructure that actually supports concurrent execution.

While demonstrated in healthcare billing (MBS) and digital asset compliance (DTS), these patterns—multi-agent verification with diverse architectures, deterministic opaque policy, cryptographic audit receipts, and warranted graduated disclosure—apply to any AI system where the consequences of undetected deviation exceed the cost of attestation.

The goal is not merely detecting AI misbehavior but designing systems where misbehavior is architecturally constrained, where oversight remains economically sustainable, and where accountability does not require sacrificing either privacy or agency.

ACKNOWLEDGMENTS

AI-based language tools were used for editorial drafting and refinement. All research concepts, system architectures, experimental designs, and technical interpretations are the original work of the authors.

This research benefits from computing facilities at the Darwin Deason Institute for Cybersecurity at Southern Methodist University, including access to the SMU SuperPod (NVIDIA A100-SXM4-80GB) for performance benchmarking.

Disclosures

Portions of the GaleStorm architecture described in this paper are the subject of provisional patent applications filed by Wave 3 LLC, and Wave 3 Digital Trust LLC, including methods for cryptographic attestation of AI decision-making (SCAR/CMARK), multi-agent auditing frameworks (GaleStorm), zero-knowledge compliance verification (Agent Native- ZKP), hardware root of trust for execution environments (DeviceWallet), and cryptographically verifiable transaction infrastructure (LockCheck). D. Young is founder of Wave 3 Digital Trust LLC. L. Young is named inventor on the GaleStorm provisional patent application. The authors declare that the research was conducted independently of any commercial development activity.

REFERENCES

1. Centers for Medicare & Medicaid Services, "NHE Fact Sheet," CMS.gov (2024).
2. U.S. Securities and Exchange Commission, "Digital Asset Market Report," (2024).
3. National Health Care Anti-Fraud Association, "The challenge of health care fraud," NHCAA (2024).
4. Elliptic, "Crypto Crime Report," (2024).
5. CertiK, "Web3 Security Report," (2024).
6. Young, D. L., "AI oversight policy brief for digital asset compliance," Wave3 Digital Trust LLC, (2025).
7. Young, D. L., Thornton, M. A., Teske, J., and Moreland, J. D., "Impossibility Engineering: Teaching AI to recognize when adversaries change the game," in *Proc. SPIE Defense + Commercial Sensing 2026*, (2026).
8. Young, D. L. and Teske, J., "Accelerating time-to-rescue: A C8-ISR framework for fusing disparate data streams in emergency response," in *Proc. SPIE Defense + Commercial Sensing 2026*, (2026).
9. Young, D. L., "ChatPack Specification v1.4: Cryptographically attested, version-controlled AI expertise containers," Wave3 Digital Trust LLC, Tech. Rep., Feb. 2026. [Online]. Available: https://github.com/wave3digitaltrust/chatpack-spec/releases/download/v1.4.0/chatpack_spec.pdf
10. Teske, J. and Young, D. L., "SEAGULL: Semantic exploration adaptive gating with ultra low latency," GWU Praxis Doctoral Project, (2026).
11. Thornton, M. A. et al., "Letter to U.S. Treasury regarding GENIUS Act compliance testing," SMU Darwin Deason Institute (2025).
12. Young, D. L. et al., "AI-audits-AI reference implementation for Treasury-aligned testing," Wave3 Digital Trust LLC (2025).
13. Kwon, W. et al., "Efficient memory management for large language model serving with PagedAttention," in *Proc. ACM SOSP*, (2023).
14. Amodei, D., "Machines of Loving Grace," Anthropic, (2024). https://darioamodei.com/machines-of-loving-grace