WHITE PAPER

The Lineage Audit Reference Architecture

Detecting Recursive Contamination in Production AI

Five Detection Tiers · Reference Deployment · Procurement Language

Anna R. Dudley

SIGNAL · STRATEGY · SYSTEMS

May 2026

§ 1

Purpose

This paper specifies a reference architecture for detecting recursive contamination in production AI systems, with cross-vendor lineage attestation, divergence monitoring, and audit artifacts compatible with existing model-risk governance frameworks.

SCOPE STATEMENT

This is detection infrastructure, not a vendor product. The architecture complements existing model risk management programs governed by SR 11-7, the NIST AI Risk Management Framework, and ISO/IEC 42001:2023. It assumes contamination is real and operationally consequential, consistent with the distinction drawn by the Berryville Institute of Machine Learning: pollution is not collapse, but pollution is sufficient to produce measurable degradation in operational systems long before collapse.

The architecture is vendor-agnostic, runs on open substrates (OpenLineage, OWASP AIBOM, C2PA v2.3), and emits SR 11-7-compatible audit artifacts natively. Adoption does not require a parallel compliance project.

§ 2

The Contamination Gap

Existing model risk frameworks require lineage in principle but were written for parametric statistical models, not for AI systems whose inputs are continuously refreshed from third-party data products. SR 11-7 assumes a model whose training set is bounded and whose validation can be reperformed on demand. OCC Bulletin 2026-13, which rescinded the long-standing OCC Bulletin 2011-12, explicitly excluded generative and agentic systems from its scope and deferred their treatment to forthcoming guidance. The NIST Generative AI Profile (AI 600-1) identifies provenance as a control objective without prescribing how to verify it across vendor boundaries. EU AI Act Article 10, enforceable from August 2, 2026, requires data governance and traceability for high-risk systems but leaves the cross-vendor attestation question to implementing standards still in draft.

The academic literature confirms the failure mode. Shumailov et al. (Nature 631:755, 2024) demonstrated model collapse as a generational phenomenon: models trained on their own outputs lose tail distributions and converge toward degenerate means. Dohmatob et al. (ICLR 2025, "Strong Model Collapse") proved analytically that contamination as low as 0.1 percent of training data is sufficient to trigger collapse dynamics in scaling-law regimes. McGovern et al. (BIML, January 2026) distinguished collapse (a training-time phenomenon) from pollution (an inference-time phenomenon affecting deployed systems whose retrieval pipelines or fine-tuning loops ingest model-generated content). Pollution does not require collapse to produce material harm.

Provenance metadata is largely absent at the source. The MIT Data Provenance Initiative documents that roughly 70 percent of widely used training datasets have ambiguous or missing provenance metadata, with

license terms frequently fabricated or misattributed downstream. This means that even when a deployer wants to verify the lineage of an input, the upstream record often does not exist.

The commercial tooling gap is the third pillar of the problem. Databricks Unity Catalog traces lineage within the Databricks boundary; its external-lineage feature entered Public Preview in 2026 and remains scoped to declared connectors. IBM watsonx.governance integrates with OpenLineage but does not yet support cross-vendor attribution of recursive reentry. Credo AI, Robust Intelligence, and Patronus address evaluation and policy mapping, not lineage. Snowflake, Collibra, and Atlan stop at the catalog boundary. No commercial tool traces cross-vendor recursive reentry as an operational concern. That is the gap this architecture addresses.

§ 3

The Five-Tier Detection Architecture

The architecture organizes detection into five tiers, ordered from cheap surface checks to expensive attribution analysis. Each tier produces a distinct class of evidence and routes to the next. Deployers should run all five in production; lower tiers gate higher tiers to control cost.

TIER 1

Presence

Manifest check. Every input to the production model must carry an AIBOM or OpenLineage record. Inputs without a manifest are flagged opaque and surfaced to the model risk officer.

EVIDENCE Boolean coverage map

LATENCY Real-time **COST** Low

TIER 2

Lineage

Trace each input back through the vendor chain to a signed original source. Resolves multi-hop redistribution where vendor B repackages vendor A's product. Produces a lineage DAG with per-vendor attribution and signature verification status.

EVIDENCE Lineage DAG **LATENCY** Seconds

COST Medium

TIER 3

Divergence

Kernel Divergence Scoring (KDS) on embeddings of the live input distribution against a pre-deployment baseline. Runs continuously without requiring vendor cooperation. Surfaces distributional drift independent of label availability.

EVIDENCE KDS time series **LATENCY** Minutes

COST Medium-high

TIER 4

Attribution

Membership inference attacks (Min-K%, ReCaLL) applied to each vendor feed to identify whether incoming inputs derive from this model's own prior outputs. Confirms recursive reentry suspected at Tier 3.

EVIDENCE Attribution scores per feed

LATENCY Hours **COST** High

TIER 5

Alerting

Threshold-based escalation. Routes incidents to the model risk officer, emits SR 11-7-compatible audit artifacts, and, for asset classes where it applies, exposes a trade-halt or inference-halt hook. Operates on the outputs of Tiers 1 through 4.

EVIDENCE Incident record + audit artifact

LATENCY Real-time **COST** Low

Tier 3 is the central innovation. Tiers 1 and 2 depend on upstream vendor cooperation; Tier 4 is expensive and runs in batch. Tier 3 runs continuously on input embeddings alone, requires no vendor cooperation, and produces actionable signal at minutes-level latency. KDS is the appropriate kernel because it is non-parametric, sensitive to multi-modal distributional shifts, and well-characterized for embedding spaces under the formulation of Kim et al. (arXiv:2502.00678, 2025).

§ 4

Architectural Principles

Four principles govern the design. Each addresses a constraint that has caused prior lineage tooling to stall at the single-vendor boundary.

1 Vendor cooperation is not assumed

The architecture degrades gracefully. Tiers 1, 2, and 4 require some level of upstream emission or accessible API surface; Tier 3 does not. KDS divergence detection runs on input embeddings alone when upstream vendors emit no provenance metadata. A deployer with zero cooperating vendors can still operate Tier 3 plus Tier 5 and obtain useful signal. As more vendors emit OpenLineage and AIBOM records, the higher tiers light up and the system gains precision.

2 OpenLineage as substrate; AIBOM and C2PA-AI as assertion formats

OpenLineage is the transport and event model; the architecture consumes OpenLineage events as the canonical interchange format. AIBOM records the AI system's bill of materials at the artifact level. C2PA v2.3 provides the cryptographic assertion format for signed content provenance. Vendors emitting none of these are flagged as opaque at Tier 1; they are not blocked, but their inputs carry a degraded confidence rating that propagates into the audit artifact.

3 Two-tier thresholds

Alerting uses two thresholds rather than a single binary cutoff. A "watch" alert fires when Tier 3 KDS divergence from baseline exceeds 2 sigma; a "halt" alert fires when divergence exceeds 3 sigma sustained across N inference windows, with N tuned per asset class. The watch threshold routes to the model risk officer for review; the halt threshold can trigger automated inference suspension where the deployer has authorized that hook. Two-tier thresholds avoid the well-documented failure mode of single-threshold alerting systems: either too noisy to be actioned or too coarse to detect early degradation.

4 The model risk officer owns the audit artifact

The architecture emits SR 11-7-compatible documentation natively, including the lineage DAG, KDS time series, attribution scores, and threshold history at the moment of any incident. Adoption does not require a separate compliance project or a parallel governance workstream. The model risk officer can present the audit artifact to examiners or to the EU AI Act notified body without additional reconciliation work.

§ 5

Reference Deployment

The reference deployment is an air-gapped on-prem variant suitable for intelligence-community use cases and bank-secret deployments. Cloud-resident variants are supported but not the focus of this section, because the constraints that govern air-gapped deployment are strictly more demanding and serve as the architectural worst case.

Per the 2026 DoD generative AI procurement directive, the air-gapped variant requires three properties: signed offline manifest updates with cryptographic provenance, an immutable audit log suitable for after-action review, and identity passthrough to existing access-control infrastructure. The deployment runs entirely within the air gap; updates to the AIBOM registry, the C2PA trust list, and the KDS baseline are delivered via signed offline bundles and verified before ingestion.

The cryptographic baseline follows the NSA, CISA, and FBI joint Cybersecurity Information Sheet on AI Data Security (May 22, 2025): ECDSA P-256 for signatures, SHA-256 for content hashing, and Merkle trees for lineage attestation. Merkle structure permits efficient verification of any single input's lineage path without exposing the full provenance graph, which is operationally important when the lineage graph itself is classified or contains commercially sensitive vendor relationships. A Framework for Cryptographic Verifiability of End-to-End AI Pipelines (arXiv:2503.22573, 2025) provides the formal verification model that the air-gapped variant implements.

Multimedia inputs are handled per CISA's January 2025 guidance on multimedia integrity: C2PA manifests are verified at ingest, content credentials are preserved through the lineage DAG, and any input lacking a content credential is flagged at Tier 1.

§ 6

Worked Example: ATLAS-FX Walkthrough

The ATLAS-FX scenario, documented in the "$14 Billion Hallucination" red team briefing, replays a financial AI signal that triggered a $14B portfolio reallocation based on recursively contaminated inputs. The scenario is constructed but the failure mode is empirically documented across the academic literature. Replayed tier by tier through the reference architecture, the contamination is caught.

T1

Presence: two opaque sources

Of the eleven principal sources feeding the ATLAS-FX signal, two lack AIBOM records. Both are flagged opaque at ingest and surfaced to the model risk officer's morning queue with a degraded confidence rating attached.

T2

Lineage: convergent vendor path

The lineage DAG reveals that four of the eleven sources route through a single upstream vendor data product. That product, on inspection of its OpenLineage emission, retrained on syndicated ATLAS-FX commentary during a quarterly refresh. The DAG visualization makes the convergence immediately visible to the reviewer.

T3

Divergence: 2.4 sigma KDS spike

Kernel Divergence Scoring on the cross-asset correlation embedding spikes to 2.4 sigma above the pre-deployment baseline in the 72 hours preceding the signal. This crosses the "watch" threshold and fires an alert to the MRO queue, with the lineage DAG from Tier 2 automatically attached.

T4

Attribution: confirmed recursive reentry

Membership inference using Min-K% and ReCaLL confirms that three of the eleven inputs contain content the model itself produced and that was syndicated, ingested by a third-party aggregator, and routed back as an independent signal. The attribution score for these three feeds exceeds 0.85.

T5

Alerting: routed to MRO with halt option

The Tier 5 incident record fires to the model risk officer at confidence 0.94 with the full lineage DAG, KDS time series, and attribution scores attached. The trade-halt hook is presented as an option; the MRO can suspend reliance on the ATLAS-FX signal pending review without disrupting other model outputs. The SR 11-7-compatible audit artifact is generated automatically.

The scenario also exposes what the architecture would not have caught absent Tier 3. Tiers 1 and 2 alone would have flagged the opaque sources and the convergent path, but neither would have triggered escalation in the absence of an explicit policy threshold; the convergent path is an architectural smell, not a violation. Tier 4 confirms what Tier 3 detects, but it runs in batch with hours of latency. Without continuous KDS monitoring, the signal would have fired before the contamination was detected.

§ 7

Implementation Checklist

The architecture splits cleanly into two operational concerns: what the deployer builds and operates, and what the deployer requires from upstream vendors. The procurement-language column is the lever for adoption. Model risk officers can paste these clauses into RFPs.

What the deployer builds

- **OpenLineage ingestion service.** Receives lineage events from upstream vendors and internal pipelines, normalizes to the canonical event schema.
- **AIBOM and C2PA-AI parser.** Extracts the bill-of-materials and content-credential assertions; verifies signatures against the trust list.
- **Embedding-baseline service.** Captures the pre-deployment input distribution as a reference embedding manifold; refreshes on documented model updates.
- **KDS divergence monitor.** Computes Kernel Divergence Score continuously against the baseline; emits time series at minute granularity.
- **Membership inference batch service.** Runs Min-K% and ReCaLL against each vendor feed on the documented cadence; produces attribution scores.
- **Alert router.** Applies two-tier thresholds; routes "watch" alerts to the MRO queue, "halt" alerts to the inference-suspension hook.
- **Audit artifact emitter.** Generates SR 11-7-compatible documentation on every incident; archives to the immutable audit log.

What to require of vendors

- **OpenLineage emission.** "Vendor shall emit OpenLineage events for all data products supplied under this agreement, conforming to the OpenLineage v1.x event schema, with the per-event source attribution field populated."
- **AIBOM compliance.** "Vendor shall furnish, with each data product release, an AI Bill of Materials conforming to the OWASP AIBOM v1.0 schema, including model provenance, training data sources, and known fine-tuning history."
- **Training-data provenance attestation.** "Vendor shall attest, in writing, to the absence of recursive ingestion of customer-derived outputs in any training or fine-tuning operation conducted on the vendor's models supplied under this agreement."
- **Signed manifest delivery.** "Vendor shall sign all delivered manifests using ECDSA P-256 against a certificate chain rooted in a CA acceptable to customer's trust list, with SHA-256 content hashing."
- **Breach-notification windows.** "Vendor shall notify customer within 24 hours of any discovered training-data contamination affecting products supplied under this agreement, and shall furnish a remediation plan within 5 business days."
- **C2PA-AI content credentials.** "For any multimedia input supplied under this agreement, vendor shall preserve and propagate C2PA v2.3 content credentials end-to-end."

Procurement language alone will not produce coverage on day one. Deployers should plan for staged rollout: cooperating vendors lit up first via Tiers 1, 2, and 4; non-cooperating vendors monitored via Tier 3 from the outset. The architecture's value compounds as vendor coverage grows, but it produces actionable signal from day one.

§ 8

Limits and Future Work

This architecture is a detection layer. It does not address every failure mode in the model lifecycle and should not be marketed as if it did.

DOES NOT REPLACE ADVERSARIAL RED-TEAMING

Adversarial probing of model behavior under crafted prompts, jailbreaks, and prompt-injection attacks is a distinct discipline. The lineage audit architecture detects contamination in the input pipeline; it does not test model resilience to adversarial inputs at inference time.

DOES NOT SUBSTITUTE FOR GROUND-TRUTH VALIDATION

Tier 3 divergence is a distributional signal, not a correctness signal. A model that produces accurate predictions on a drifted input distribution is still drifted. Ground-truth validation against held-out labels remains necessary for any prediction whose accuracy can be measured.

SINGLE-VENDOR SELF-DEPLOYMENT IS OUT OF SCOPE

Where the vendor is also the deployer (e.g., a foundation model lab serving its own model on its own infrastructure), the architecture's cross-vendor assumptions collapse. Internal-only contamination requires the deployer's own data-governance program. The architecture can run, but Tiers 1 and 2 will reduce to single-source attestation.

PRE-DEPLOYMENT TRAINING-DATA CONTAMINATION IS OUT OF SCOPE

Contamination introduced during pre-training or fine-tuning, before the model is deployed to production, is the domain of upstream data-provenance frameworks (MIT DPI, C2PA training-data provisions, the EU AI Act Article 10 obligations). The architecture detects contamination introduced during deployment, not contamination baked into the model at training time.

MEMBERSHIP INFERENCE IS PROBABILISTIC

Min-K% and ReCaLL produce calibrated probability scores, not deterministic membership labels. False positive and false negative rates are documented in the membership-inference literature and propagate into Tier 4 outputs. Threshold calibration per deployment is required.

Future work centers on three problems. First, joint optimization of Tier 3 and Tier 4 to reduce membership-inference cost without sacrificing detection sensitivity. Second, formal verifiability of the lineage DAG itself using the cryptographic framework of arXiv:2503.22573, extended to multi-hop vendor

chains. Third, integration with the implementing standards under development for EU AI Act Article 10 enforcement, expected in 2027.

§ 9

References

References are organized by category and numbered consecutively. Hyperlinks point to the canonical source where available.

REGULATORY AND FRAMEWORK

1. Federal Reserve Board. *SR 11-7: Guidance on Model Risk Management*. April 2011. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

2. Office of the Comptroller of the Currency. *OCC Bulletin 2026-13: Model Risk Management Guidance Update (rescinds OCC Bulletin 2011-12)*. 2026. https://www.occ.treas.gov/news-issuances/bulletins/2026/bulletin-2026-13.html

3. National Institute of Standards and Technology. *AI Risk Management Framework (AI 100-1)*. January 2023. https://www.nist.gov/itl/ai-risk-management-framework

4. National Institute of Standards and Technology. *AI 600-1: Artificial Intelligence Risk Management Framework — Generative Artificial Intelligence Profile*. July 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

5. European Union. *EU AI Act, Article 10: Data and Data Governance*. Enforceable August 2, 2026. https://artificialintelligenceact.eu/article/10/

6. International Organization for Standardization. *ISO/IEC 42001:2023, Information Technology — Artificial Intelligence — Management System*. December 2023. https://www.iso.org/standard/42001

7. NSA, CISA, and FBI. *AI Data Security: Best Practices for Securing Data Used to Train and Operate AI Systems*. Joint Cybersecurity Information, May 22, 2025. https://media.defense.gov/2025/May/22/2003722410/-1/-1/0/CSI_AI_DATA_SECURITY.PDF

8. Cybersecurity and Infrastructure Security Agency. *Strengthening Multimedia Integrity in the Generative AI Era*. January 2025. https://www.cisa.gov/sites/default/files/2025-01/strengthening-multimedia-integrity-genai-era.pdf

9. U.S. Department of Defense. *2026 Generative AI Procurement Directive*. 2026. https://www.acquisition.gov/dod-genai-2026

ACADEMIC

10. Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., and Anderson, R. *The Curse of Recursion: Training on Generated Data Makes Models Forget*. arXiv:2305.17493. 2023. https://arxiv.org/abs/2305.17493

11. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. *AI Models Collapse When Trained on Recursively Generated Data*. Nature 631:755, July 2024. https://www.nature.com/articles/s41586-024-07566-y

12. Alemohammad, S., et al. *Self-Consuming Generative Models Go MAD*. ICLR 2024 (arXiv:2307.01850). https://arxiv.org/abs/2307.01850

13. Dohmatob, E., Feng, Y., Subramonian, A., and Kempe, J. *Strong Model Collapse*. ICLR 2025. https://openreview.net/forum?id=et5l9qPUhm

14. Gerstgrasser, M., et al. *Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data*. arXiv:2404.01413. 2024. https://arxiv.org/abs/2404.01413

15. Kim, S., et al. *How Contaminated Is Your Benchmark? Quantifying Dataset Leakage in Large Language Models with Kernel Divergence Scores*. arXiv:2502.00678. 2025. https://arxiv.org/abs/2502.00678

16. Longpre, S., et al. *Data Provenance Initiative: A Large-Scale Audit of Dataset Licensing & Attribution in AI (multimodal)*. MIT Data Provenance Initiative. 2024. https://www.dataprovenance.org

17. McGovern, A. *Recursive Pollution and Model Collapse Are Not the Same Thing*. Berryville Institute of Machine Learning, January 2026. https://berryvilleiml.com/2026/01/recursive-pollution-vs-model-collapse

18. Shi, W., et al. *Detecting Pretraining Data from Large Language Models (Min-K%)*. arXiv:2310.16789. 2023. https://arxiv.org/abs/2310.16789

19. Xie, R., et al. *ReCaLL: Membership Inference via Relative Conditional Log-Likelihoods*. arXiv:2404.11262. 2024. https://arxiv.org/abs/2404.11262

20. *A Framework for Cryptographic Verifiability of End-to-End AI Pipelines*. arXiv:2503.22573. 2025. https://arxiv.org/abs/2503.22573

COMMERCIAL AND STANDARDS

21. Coalition for Content Provenance and Authenticity. *C2PA Specification, Version 2.3*. December 2025. https://spec.c2pa.org/specifications/specifications/2.3/specs/C2PA_Specification.html

22. LF AI & Data Foundation. *OpenLineage Project*. https://openlineage.io

23. OWASP. *AI Bill of Materials (AIBOM) Initiative*. https://owasp.org/www-project-aibom

24. Databricks. *Unity Catalog External Lineage (Public Preview)*. 2026. https://docs.databricks.com/en/data-governance/unity-catalog/data-lineage.html

25. IBM. *watsonx.governance with OpenLineage Integration*. 2026. https://www.ibm.com/products/watsonx-governance

SCENARIO SOURCE

26. Dudley, A. R. *The $14 Billion Hallucination: A Red Team Briefing on Recursive Contamination in Financial AI*. annardudley.com, 2026. https://annardudley.com/briefing-fourteen-billion-hallucination.html

§10

Download & Cite

This white paper is available as a PDF for offline reading and citation. The reference architecture, procurement language, and audit artifact templates are released for adoption without restriction. Deployers, regulators, and vendors are encouraged to extend the specification.

SUGGESTED CITATION

Dudley, Anna R. *The Lineage Audit Reference Architecture: Detecting Recursive Contamination in Production AI*. annardudley.com, May 2026. https://annardudley.com/lineage-audit-white-paper.html