Article

Not peer-reviewed version

AI Supply Chain Security: MBOM-PQC Provenance, PQC Attestation, and a Maturity Model for Quantum-Resistant Assurance

Robert Campbell *

Posted Date: 1 May 2026

doi: 10.20944/preprints202603.1963.v2

Keywords: AI supply chain security; post-quantum cryptography; model provenance; MBOM; ML-DSA; SLH-DSA; hybrid signatures; PQC migration; supply chain assurance; maturity model

 doi:10.20944/preprints202603.1963.v2

Article

AI Supply Chain Security: MBOM-PQC Provenance, PQC Attestation, and a Maturity Model for Quantum-Resistant Assurance

Robert Campbell

Independent Researcher, Upper Marlboro, MD 20774, USA; Fellow, British Blockchain Association; rc@medcybersecurity.com

Abstract

Artificial intelligence (AI) systems increasingly depend on multi-stage supply chains that incorporate pre-trained models, third-party datasets, open-source libraries, and automated pipelines, creating an expanding attack surface in which model poisoning, dependency compromise, and provenance manipulation can undermine integrity before deployment. Existing AI governance frameworks—including the NIST AI Risk Management Framework and Secure Software Development Framework—acknowledge supply chain risks but do not define verifiable model provenance or cryptographically durable integrity guarantees. The transition to post-quantum cryptography (PQC) compounds this gap: classical digital signatures used to verify model lineage, dataset integrity, and pipeline attestation will become vulnerable to quantum-enabled forgery within the operational lifetime of many AI systems. This paper synthesizes evidence from policy, standards, and incident sources to characterize the AI supply chain threat landscape and the cryptographic dependencies that the PQC transition disrupts. It proposes three integrated design-science artifacts: a Model Bill of Materials with PQC-safe extensions (MBOM-PQC) defining a verifiable provenance schema; a unified signing and attestation pipeline integrating ML-DSA and hybrid signature modes; and a five-level Supply Chain Assurance Maturity Model (SCAMM) for repeatable organizational evaluation. These contributions provide a structured foundation for AI supply chain integrity in cloud-connected, mission-critical smart systems, ensuring verifiable lineage, authenticity, and trustworthiness through the PQC transition. Empirical validation is deferred to future work.

Keywords: AI supply chain security; post-quantum cryptography; model provenance; MBOM; ML-DSA; SLH-DSA; hybrid signatures; PQC migration; supply chain assurance; maturity model

1. Introduction

Artificial intelligence (AI) systems increasingly rely on complex, multi-stage supply chains that integrate pre-trained models, third-party datasets, open-source libraries, cloud-hosted training pipelines, and automated deployment workflows. While this modularity accelerates capability development, it also introduces opaque dependencies and systemic vulnerabilities. Compromise at any point in the supply chain—whether through poisoned training data, tampered model weights, malicious dependencies, or manipulated provenance metadata—can undermine the integrity of downstream AI systems long before they reach operational environments.

Five terms recur throughout this manuscript and warrant explicit definition for readers approaching from adjacent disciplines. An *AI supply chain* refers to the full set of artifacts and processes that influence model behavior — pre-trained foundation models, datasets used for pre-training and fine-tuning, open-source libraries, training infrastructure, and deployment pipelines — together with the entities that produce, transform, distribute, and consume these artifacts (consistent with NIST AI RMF [1] usage). *Model provenance* denotes the verifiable lineage of an AI artifact, capturing who produced it, from what inputs, in what environment, and through what

 doi:10.20944/preprints202603.1963.v2

transformations. *Post-quantum cryptography (PQC)* refers to cryptographic primitives believed to remain secure against adversaries equipped with cryptographically relevant quantum computers; the present manuscript focuses on the digital-signature subset standardized by NIST as FIPS 204 (ML-DSA) [2] and FIPS 205 (SLH-DSA) [3]. *Cryptographic agility* is the architectural property of being able to substitute one cryptographic primitive for another without redesigning the surrounding system, enabling phased migration as standards evolve. A *hybrid signature mode* binds a classical signature (typically ECDSA or Ed25519) and a post-quantum signature (typically ML-DSA) over the same payload, so that a verifier requires both to validate; this construction provides backward compatibility with classical-only verifiers while ensuring that compromise of either algorithm alone does not invalidate the integrity guarantee.

Documented incidents involving compromised machine learning libraries, malicious PyPI packages, and insecure model-serving or distribution pathways indicate that AI supply chain attacks are no longer theoretical; they represent a growing class of adversarial activity targeting both commercial and government systems [4,5]. The MITRE ATLAS knowledge base further systematizes these attack patterns by cataloging adversarial tactics and techniques relevant to AI-enabled systems [6]. The same supply chain risk surface extends into consumer-deployed AI — voice assistants, mobile inference, on-device personalization — where compromise of the model distribution path or its dependencies can affect end-user trust at scale (the OWASP ML Top 10 [7], OWASP GenAI LLM Top 10 [8], and ISO/IEC 42001:2023 [9] each address consumer-facing dimensions of this risk surface from complementary angles).

Despite this expanding threat landscape, existing AI governance frameworks provide limited guidance on verifiable model provenance or cryptographically durable integrity guarantees. The NIST AI Risk Management Framework emphasizes governance, transparency, and robustness [1] but does not define a standardized structure for documenting model lineage or verifying the authenticity of training artifacts. NIST's Secure Software Development Framework addresses software supply chain risks [10] but does not extend its requirements to AI-specific artifacts such as model checkpoints, fine-tuning datasets, or hyperparameter configurations. Similarly, emerging model evaluation and red-teaming guidance focuses on behavioral robustness rather than supply chain integrity. As a result, organizations lack a unified method for establishing trust in the origin, composition, and integrity of the AI components they deploy.

AI artifacts and the records that attest to their integrity often have operational lifetimes measured in years. Foundation models deployed in defense, healthcare, transportation, and critical infrastructure may remain in service well beyond the anticipated arrival of cryptographically relevant quantum computers. The cryptographic-confidentiality analogue of this temporal mismatch — the *harvest-now, decrypt-later* (HNDL) pattern, in which adversaries capture encrypted traffic today for future decryption once quantum capability matures — is well documented in the published literature on quantum-safe transition timelines and is the basis for normative guidance such as NSA Commercial National Security Algorithm Suite (CNSA) 2.0 [11] and OMB M-23-02 [12]. The same temporal logic, applied to digital signatures rather than to encryption, has received less explicit treatment in the AI assurance literature.

This manuscript introduces and uses the term *harvest-now, forge-later* (HNFL) to name a previously under-articulated threat class for AI artifacts. The pattern HNFL names — adversaries collecting signed AI artifacts under current classical signature regimes for later forgery of counterfeit signatures, or substituting counterfeit artifacts under reused signing keys, once cryptographically relevant quantum computers become available — is implicit in the cryptographic transition timelines documented by NSA CNSA 2.0 [11] and OMB M-23-02 [12]. However, those documents address the post-quantum transition principally in confidentiality terms, where HNDL is the established pattern, rather than in signature-integrity terms. The HNFL framing is used in the remainder of the paper to motivate the requirement that signing pipelines for AI artifacts be PQC-safe by the end of the operational lifetime of the artifacts they attest, not merely by the end of the lifetime of the immediate signing operation.

 doi:10.20944/preprints202603.1963.v2

The absence of a standardized, cryptographically resilient provenance model creates operational challenges for organizations deploying AI systems in high-assurance environments. Without verifiable lineage, Authorizing Officials cannot reliably assess whether a model has been tampered with, whether training data originated from trusted sources, or whether dependencies were validated using PQC-safe mechanisms. Program managers lack a structured method for evaluating supply chain risk across heterogeneous AI pipelines. Enterprise architects face difficulty integrating AI assurance into broader modernization efforts such as Zero Trust Architecture (ZTA) and PQC migration. These challenges mirror broader patterns observed in cybersecurity modernization: independently justified initiatives create cross-program dependencies that become visible only during implementation.

This paper addresses these gaps by proposing a formal, cryptographically anchored framework for AI supply chain security. Through systematic evidence synthesis of policy documents, standards publications, and documented supply chain incidents, we identify the structural weaknesses that undermine current AI provenance practices and the cryptographic dependencies that PQC transition disrupts. Building on this analysis, we introduce three prescriptive contributions: (1) a Model Bill of Materials with PQC-safe extensions (MBOM-PQC) that defines a verifiable provenance schema for AI artifacts; (2) a unified model signing and attestation pipeline that integrates ML-DSA, hybrid signature modes, and PQC-ready certificate-chain design; and (3) a five-level Supply Chain Assurance Maturity Model (SCAMM) that enables repeatable organizational assessment and roadmap development. Together, these components aim to support a durable foundation for AI supply chain integrity, with the goal that model lineage, authenticity, and trustworthiness remain verifiable throughout the PQC transition and beyond.

This work is positioned within the Future Internet ecosystem as an assurance architecture for AI-enabled, cloud-connected, mission-critical, and distributed smart systems whose trustworthiness depends on durable model provenance, cryptographic integrity, and lifecycle attestation.

2. Background and Related Work

AI supply chain security has emerged as a critical concern as organizations increasingly rely on externally sourced models, datasets, and machine learning components. Unlike traditional software supply chains—where source code, binaries, and dependencies can be tracked through established mechanisms such as software bills of materials (SBOMs)—AI supply chains involve artifacts that lack standardized provenance structures, cryptographic protections, or lifecycle visibility. This section reviews the foundational elements of AI supply chain risk, the cryptographic underpinnings of AI assurance, the implications of post-quantum cryptography (PQC) transition, and the limitations of existing frameworks.

2.1. AI Supply Chain Risks

Modern AI systems are rarely built from scratch. Instead, they incorporate pre-trained foundation models, fine-tuning datasets, open-source libraries, and automated training pipelines. Each component introduces potential attack vectors. The peer-reviewed literature on adversarial machine learning has documented training-time attacks across this surface in considerable detail, with Carlini et al. demonstrating poisoning of training-data sources [13], Goldblum et al. providing a systematic survey of model-poisoning techniques [14], and Jiang et al. empirically studying pre-trained model reuse practices and supply chain risks in the Hugging Face ecosystem [15]. Training-time attacks — including data poisoning, model poisoning, and backdoor insertion — can compromise model behavior before deployment. Model ingestion attacks, where adversaries tamper with pre-trained models hosted on public repositories, can introduce malicious behaviors that propagate downstream; this pattern has been documented empirically in the PyTorch CVE-2023-43654 advisory [16] and the broader analysis of ML-package supply chain compromise by Ladisa et al. [17]. Dependency compromise — such as malicious PyPI or NPM packages embedded in ML workflows — can alter training logic or exfiltrate sensitive data. Artifact replacement and tampering

 doi:10.20944/preprints202603.1963.v2

during distribution or deployment, in which a legitimate model is substituted with a manipulated version, undermine trust in the entire pipeline [4]. These risks are amplified by the opacity of AI artifacts. Unlike source code, model weights and training datasets are difficult to inspect, so tampering is hard to detect without cryptographic verification or provenance metadata. The MITRE ATLAS framework catalogs these and related adversarial techniques across the AI lifecycle [6]; consequently, AI supply chain compromise can remain undetected until operational failures occur.

2.2. Cryptographic Foundations of AI Assurance

Cryptography plays a central role in establishing trust in AI systems. The peer-reviewed literature on model provenance and signed-artifact integrity provides the foundational treatment: Li et al. surveyed deep neural network watermarking techniques [18], and Zhu et al. analyzed the integrity properties of signed model distribution channels [19]. Within this academic framing, digital signing mechanisms are used to verify the authenticity and integrity of model artifacts, providing a cryptographic basis for detecting tampering and confirming provenance. Dataset integrity verification supports the assurance that training data has not been altered or poisoned. Secure enclaves and confidential computing protect model execution environments; federated-learning authentication ensures that model updates originate from trusted participants, with the cryptographic-protocol-level analysis of federated-learning signing provided by Rieger et al. [20]. Pipeline attestation verifies that training and deployment workflows executed in approved environments. These mechanisms rely heavily on digital signatures, certificate chains, and secure key management — most current implementations of which use classical algorithms (RSA, ECDSA, Ed25519) that are vulnerable to quantum-enabled attacks. As AI systems become more deeply integrated into mission-critical and long-lived applications, the durability of these cryptographic assurances becomes a central concern.

2.3. Post-Quantum Cryptography Transition Requirements

The transition to PQC introduces new constraints for AI assurance. The performance characteristics of the lattice-based and hash-based PQC signature schemes have been characterized in detail in the peer-reviewed cryptographic literature: the original CRYSTALS-Dilithium specification by Bai, Ducas, Kiltz, Lepoint, Lyubashevsky, Schwabe, Seiler, and Stehlé [21] provides the algorithm parameters, security analysis, and reference benchmarks that NIST FIPS 204 [2] now standardizes; the SPHINCS+ family that became FIPS 205 [3] is similarly grounded in published cryptographic analysis. Within this standards framing, NIST's standardization of ML-KEM (FIPS 203) [22] for key establishment and ML-DSA (FIPS 204) [2] and SLH-DSA (FIPS 205) [3] for digital signatures fundamentally changes the cryptographic landscape. PQC algorithms have significantly larger key and signature sizes than their classical counterparts. The increase affects model signing workflows, certificate chains, secure enclaves, federated learning protocols, and long-lived artifacts such as model checkpoints and provenance records. NSA CNSA 2.0 [11] establishes a quantum-resistant signature direction for National Security Systems aligned with ML-DSA. The guidance emphasizes that classical signatures are not designed to provide long-term integrity protection. SLH-DSA, while not included in CNSA 2.0, is approved by NIST for federal and commercial use outside NSS and offers conservative hash-based security for long-lived archival artifacts. AI artifacts — especially models used in defense, healthcare, transportation, and critical infrastructure — often have operational lifetimes extending beyond the expected arrival of cryptographically relevant quantum computers. This creates a need for PQC-safe signing and hybrid signature modes in AI supply chains, supporting continued validity of integrity guarantees through the post-quantum transition.

2.4. Gaps in Existing Frameworks

Several frameworks address aspects of AI governance, software supply chain security, or cryptographic transition, but none provide a unified approach to AI supply chain integrity. The NIST

 doi:10.20944/preprints202603.1963.v2

AI Risk Management Framework (AI RMF) [1] provides governance and risk categories but does not define a model provenance structure or cryptographic requirements. NIST's Secure Software Development Framework (SSDF) [10] addresses software supply chain risks but does not extend to AI-specific artifacts such as model weights or training datasets. NIST Cybersecurity Supply Chain Risk Management practices (SP 800-161r1) [23] specify third-party and vendor risk management for federal systems but predate the AI-specific provenance and lineage challenges this work addresses. The NIST Cybersecurity Framework (CSF 2.0) [24] provides a five-function (Identify, Protect, Detect, Respond, Recover) governance backbone applicable to AI systems but does not specify model-artifact provenance structures or PQC-safe attestation requirements. NIST SP 800-204D [25] provides guidance for integrating software supply chain security controls into DevSecOps CI/CD pipelines, but it does not define AI-specific provenance structures or model assurance mechanisms. The DoD CDAO Responsible AI Toolkit [26] focuses on ethical and operational considerations rather than cryptographic integrity. Established Software Bill of Materials (SBOM) standards—including SPDX and CycloneDX—do not capture AI-specific metadata such as dataset lineage, hyperparameters, or training environment details. PQC transition guidance, including NIST SP 800-208 [27] and CNSA 2.0 [11], does not address AI artifacts or model provenance. The absence of an established Model Bill of Materials (MBOM) standard combines with the absence of PQC-safe signing and attestation pipelines. Together, these gaps leave organizations without a structured method for verifying the authenticity, lineage, and integrity of AI systems. This gap motivates the proposed formal, cryptographically anchored framework that integrates provenance, PQC-safe signatures, and supply chain assurance.

To make the gap analysis more substantive than the rhetorical claim that adjacent frameworks are insufficient, Table 1 places MBOM-PQC field-by-field alongside seven authoritative comparators along twelve provenance and assurance dimensions. The table is referenced again from the introductions to Sections 5, 6, and 7.

Table 1. Comparative Analysis of MBOM-PQC (Model Bill of Materials with post-quantum cryptography-safe extensions) against Adjacent Frameworks. Entries indicate the extent of coverage in each comparator: ● = covered, ◐ = partially covered, ○ = absent, — = out of scope.

Dimension	SPDX 3.0	CycloneDX 1.6	NIST SSDF	NIST AI RMF	NIST CSF 2.0	OWASP ML	OWASP GenAI	MBOM-PQC
Model metadata (architecture, version)	◐	◐	○	○	○	○	◐	●
Pre-training dataset lineage	○	○	○	◐	○	◐	◐	●
Fine-tuning artifacts	○	○	○	◐	○	◐	◐	●
Pre-trained model dependencies (CVE links)	◐	●	●	○	◐	◐	◐	●
Training environment attestation	○	○	◐	○	◐	○	○	●
Deployment packaging dependency graph	●	●	◐	○	◐	○	◐	●
Cryptographic integrity fields	◐	◐	◐	○	◐	○	○	●
PQC-safe signing (FIPS 204/205)	○	○	○	○	○	○	○	●
Hybrid signature mode support	○	○	○	○	○	○	○	●
Lifecycle stage attestation	○	○	◐	○	◐	○	○	●

 doi:10.20944/preprints202603.1963.v2

Dimension	SPDX 3.0	CycloneDX 1.6	NIST SSDF	NIST AI RMF	NIST CSF 2.0	OWASP ML	OWASP GenAI	MBOM-PQC
Organizational maturity model	○	○	○	◐	●	○	○	●
Governance/policy mapping	○	○	◐	●	●	◐	◐	●

The pattern shown by Table 1 is that the adjacent frameworks each cover a distinct slice of the requirement space — SPDX/CycloneDX for software-artifact dependencies, SSDF for development-process controls, AI RMF and CSF for governance and risk categories, the OWASP entries for adversarial attack patterns — but no comparator covers the full surface that MBOM-PQC targets, and none integrates PQC-safe signing as a first-class field. The MBOM-PQC contribution is therefore not the introduction of provenance structure as such, but the integration of provenance, lifecycle attestation, and post-quantum cryptographic durability into a single auditable record.

3. Materials and Methods

3.1. Research Design and Contribution Type

This study employs a dual-method research design organized as two sequential phases of a *design-science research program* [28,29]. The first phase — Phase A, evidence collection — synthesizes authoritative policy documents, standards publications, and documented AI supply chain incidents to characterize the current state of AI supply chain risk and to identify the cryptographic dependencies that the post-quantum transition disrupts. The second phase — Phase B, artifact construction — derives a formal provenance schema, a PQC-safe signing pipeline, and a maturity model from the synthesized requirements. The boundary between phases is explicit: Sections 2 and 4 report Phase-A findings; Sections 5, 6, and 7 present the Phase-B artifacts. Section 8 (Discussion) and the empirical-validation roadmap in Section 8.4.4 correspond to the Demonstration and Evaluation activities specified in the Peffers et al. design-science framework, with empirical validation explicitly identified as deferred to future work.

Phase A follows PRISMA 2020 [30] adapted for policy, standards, and security-incident synthesis. The adaptation is necessary because the evidence base consists primarily of normative requirements, technical specifications, and documented attack patterns rather than experimental studies; PRISMA's screening, inclusion-criterion, and reporting structure transfer cleanly to this corpus while the meta-analytic statistical apparatus does not apply. Phase B follows the design-science guidelines of Hevner et al. [28] (relevance to a real-world problem, rigor in construction from the evidence base, design as iterative refinement) and the activity sequence of Peffers et al. [29] (problem identification, objectives definition, design and development, demonstration, evaluation, communication). The principal contributions of this manuscript — the MBOM-PQC schema, the signing and attestation pipeline, and the SCAMM maturity model — are *design-science artifacts* in this sense: they are constructions derived from a structured evidence synthesis, intended for future operational validation rather than presented as empirically validated findings.

Traceability between included sources, analytical propositions, extracted requirements, and architectural components is documented through the supplementary evidence bibliography and confidence-tier summary; the extraction matrix, exclusion ledger, and detailed architectural traceability mappings are maintained as part of the author's research records and available on request. The single-author nature of the study and the compensating controls applied in the screening process are described in §3.5.

3.2. Analytical Propositions

 doi:10.20944/preprints202603.1963.v2

The evidence synthesis is structured around four analytical propositions that define the conceptual scope of the review. These propositions are not hypotheses requiring statistical testing; rather, they serve as organizing assumptions that guide source selection and data extraction. AP1 holds that AI supply chain compromise is a documented and increasing threat, with attack patterns spanning training-time, ingestion-time, and deployment-time vectors. AP2 holds that cryptographic mechanisms underpin AI assurance, and that PQC transition disrupts classical assumptions about long-term integrity and authenticity. AP3 holds that existing governance and supply chain frameworks lack standardized, verifiable provenance structures for AI artifacts. AP4 holds that PQC-safe signing and hybrid signature modes can provide durable integrity guarantees but require architectural redesign of AI pipelines. Together, these propositions ensure that the synthesis captures both the threat landscape and the cryptographic requirements necessary for long-term assurance.

3.3. Search Strategy

Evidence was collected across five source classes between January 2025 and January 2026 (with two IETF Internet-Drafts [31,32] subsequently updated during manuscript revision and re-screened in March 2026). The seven databases queried, and the verbatim search strings used, are listed in Table 2. Each query was executed once at the date shown; for follow-up screening rounds during manuscript revision, the same queries were re-executed against the same databases with the date filter constrained to the intervening period.

Table 2. Database Queries Executed in Phase A.

Database	Query string	Date executed	Hits	After dedup
IEEE Xplore	("AI supply chain" OR "machine learning supply chain") AND ("provenance" OR "attestation" OR "signing")	14 Jan 2025; rerun 12 Mar 2026	47	40
ACM Digital Library	("model signing" OR "model provenance") AND ("integrity" OR "adversarial")	14 Jan 2025; rerun 12 Mar 2026	31	24
IACR ePrint	"post-quantum" AND ("signature size" OR "hybrid signature" OR "ML-DSA" OR "SLH-DSA")	16 Jan 2025; rerun 12 Mar 2026	22	19
NIST CSRC	"AI" OR "post-quantum" filtered to FIPS, SP 800 series, AI 600 series	16 Jan 2025; rerun 12 Mar 2026	14	14
NSA.gov	"CNSA 2.0" OR "quantum-resistant" filtered to advisory and CSI publications	17 Jan 2025; rerun 13 Mar 2026	6	6
CISA.gov	"AI" AND ("supply chain" OR "secure by design")	17 Jan 2025; rerun 13 Mar 2026	11	9
Curated incident repositories (PyTorch advisories, ReversingLabs, Hugging Face Hub, Ultralytics tracker)	manual curation by date filter 2020–2026	18–20 Jan 2025; rerun 14 Mar 2026	11	11
Total before screening			142	123

The overall flow from 142 records returned to 54 included sources is documented in the PRISMA diagram in Figure 1. The full search-string log, including any operator or filter modifications applied

 doi:10.20944/preprints202603.1963.v2

by individual database front-ends, is maintained as part of the author's research records and is available on request.

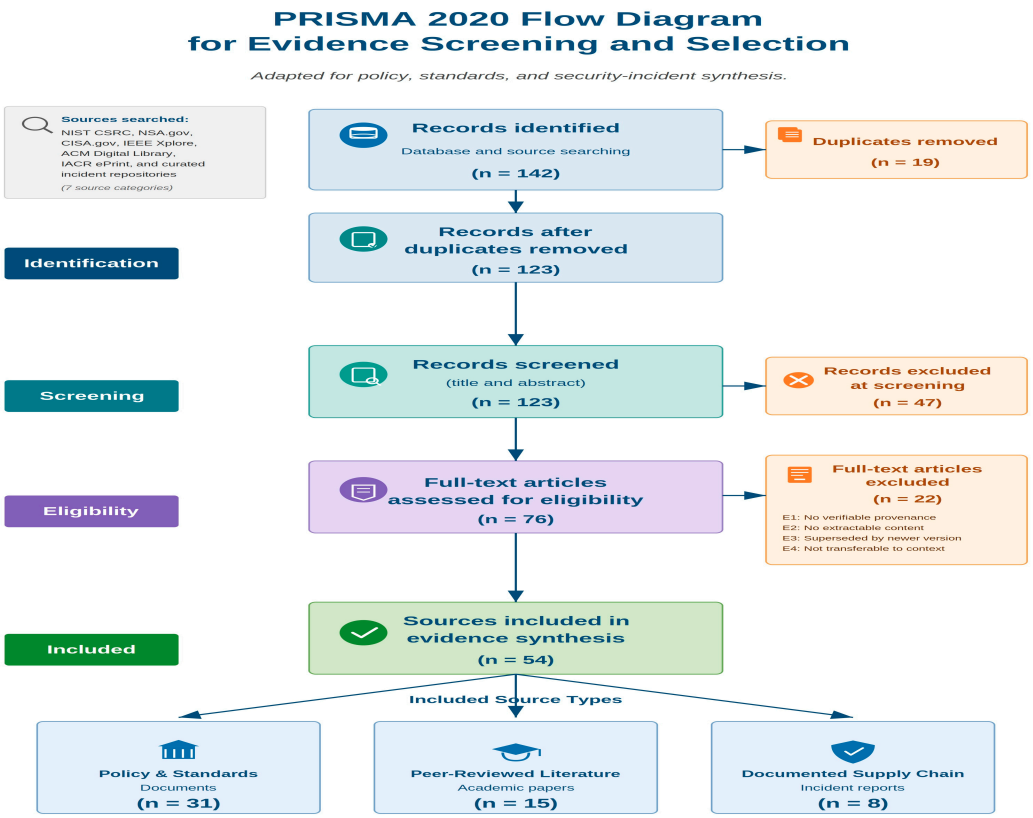

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 Flow Diagram—Evidence Screening and Selection. Adapted for policy, standards, and security-incident synthesis [30]. The initial search identified 142 records across seven source categories (NIST CSRC, NSA.gov, CISA.gov, IEEE Xplore, ACM Digital Library, IACR ePrint, and curated incident repositories), as enumerated in Table 2. After deduplication (n = 19), 123 records underwent title and abstract screening, yielding 76 full-text articles assessed for eligibility. Twenty-two records were excluded at full-text review against four exclusion criteria: no verifiable provenance (E1), no extractable technical or normative content (E2), superseded by a newer version (E3), and not transferable to AI supply chain or cryptographic assurance contexts (E4). The final evidence base comprises 54 sources distributed across the five evidence-confidence tiers defined in §3.7: 11 finalized normative standards (T1), 14 government policy and authoritative guidance documents (T2), 13 peer-reviewed academic papers (T3), 11 vendor advisories and incident reports (T4), and 5 work-in-progress drafts and industry analyses (T5).

3.4. Eligibility Criteria

Sources were included if they met all four of the following criteria: published between 2020 and 2026, with the core eligibility window of 2020–2025 extended to accommodate IETF Internet-Drafts updated during manuscript revision (Refs. [31] and [32], revised to 2026 versions; retained as transition-relevant work-in-progress for implementation context rather than primary normative authority); addressing AI supply chain risk, cryptographic integrity, or PQC transition; containing normative requirements, technical specifications, or documented incidents; and providing extractable content relevant to provenance, signing, or attestation. Sources were excluded on any of four grounds: no verifiable provenance (E1); no extractable technical or normative content (E2); superseded by a newer version of the relevant standard (E3); or not transferable to AI supply chain or cryptographic assurance contexts (E4). Exclusion codes were applied independently before arbitration to ensure consistency across the screening process.

 doi:10.20944/preprints202603.1963.v2

3.5. Screening and Selection

The initial search identified 142 potentially relevant records. After deduplication (n = 19), 123 records underwent title and abstract screening. Of these, 76 records proceeded to full-text review, yielding 54 included sources distributed across the five evidence-confidence tiers defined in §3.7: 11 finalized normative standards (T1), 14 government policy and authoritative guidance documents (T2), 13 peer-reviewed academic papers (T3), 11 vendor advisories and incident reports (T4), and 5 work-in-progress drafts and industry analyses (T5). Records excluded at full-text review were coded against the exclusion criteria to support auditability. The PRISMA-style flow diagram documenting the screening and selection process is presented in Figure 1.

After full-text review, the PRISMA-screened evidence base contains 54 sources, of which 46 are cited directly in the body to support specific argumentative claims, while the remaining 8 informed the synthesis, analytical propositions, and architectural derivation [33–40]. Six additional references were added during revision in response to peer review: two design-science methodology references [28,29] cited in §3.1, one algorithm-specification reference [21] cited in §8.3.5, one Trusted Platform Module (TPM) 2.0 platform-profile reference [41] cited in §8.3.6, and two federal PQC PKI migration guidance references [42,43] cited in §6.3.2. The total reference count is therefore 60. All 60 sources are listed in the References section to comply with MDPI citation requirements for supplementary materials. The complete 54-source PRISMA-screened evidence bibliography with tier assignments and extraction metadata is provided in the Supplementary Materials; the six revision-time additions are flagged separately in that bibliography.

This is a single-author study. Inter-rater agreement statistics (e.g., Cohen's kappa) are therefore not reported because there is no second rater. Three compensating controls were applied during screening to provide reproducibility in the absence of inter-rater reliability: (i) an *a priori* coding rubric was specified before screening began (eligibility codes I1–I4 and exclusion codes E1–E4 in §3.4), and screening decisions were recorded against the rubric in an exclusion ledger maintained as part of the author's research records and available on request; (ii) borderline sources that resisted clear classification were defaulted to exclusion and the exclusion code documented, rather than being included on the basis of subjective assessment; (iii) the requirements-to-architecture traceability matrices in §5.4 and §7.4 serve as an internal consistency check, since each architectural component must trace back to specific included sources, surfacing any source whose inclusion is not supported by downstream use. The single-author limitation is acknowledged in §8.4 and a Delphi-style multi-rater evaluation is identified in §8.4.4 as a priority for future validation.

3.6. Data Extraction and Coding

Each source was coded using a structured extraction template with nine fields: source identifier and citation; source type (policy, standard, incident, or benchmark); confidence tier (T1 through T5); domain relevance (AI supply chain, cryptography, PQC, or cross-cutting); supported analytical proposition (AP1–AP4); extracted requirements covering provenance fields, signature constraints, and lifecycle dependencies; cryptographic parameters including key sizes, signature sizes, and hybrid mode requirements; documented vulnerabilities or attack patterns; and limitations noted by original authors. This coding structure ensures end-to-end traceability from evidence to architectural decisions, enabling reviewers to audit the derivation of each schema field, pipeline component, and maturity indicator.

3.7. Evidence Confidence Tiers

Each included source was assigned to one of five evidence-confidence tiers based on the type and authority of the source. The tier assignments are used to weight the synthesis: claims supported by Tier 1 or Tier 2 sources carry stronger evidentiary weight than claims supported only by Tier 4 or Tier 5 sources, and the requirements derived in §4.4 are anchored, where possible, in Tier 1 or Tier 2 evidence. The tiers and their criteria are:

 doi:10.20944/preprints202603.1963.v2

Tier 1 — Finalized normative standards. Published, formally adopted standards from NIST, NSA, ISO/IEC, or IETF (RFCs only, not Internet-Drafts).

Tier 2 — Government policy and authoritative guidance. Published policy documents and authoritative guidance from federal agencies and equivalent national bodies, with formal endorsement (memoranda, executive orders, agency CSI publications).

Tier 3 — Peer-reviewed academic literature. Articles in peer-reviewed venues with editorial oversight and double-blind or single-blind review.

Tier 4 — Vendor advisories, incident reports, CVE-anchored disclosures. Authoritative reports from named vendors or organizations documenting specific security events with verifiable artifacts (CVE identifiers, advisory links, incident timestamps).

Tier 5 — Work-in-progress drafts and industry analyses. IETF Internet-Drafts, working-group artifacts, and named-vendor industry analyses. Used in this manuscript only for transition-pathway context, never as primary normative authority.

The full source-to-tier assignment is provided in the Supplementary Materials. The 54 PRISMA-screened sources break down by tier as follows: T1 = 11, T2 = 14, T3 = 13, T4 = 11, and T5 = 5.

3.8. Requirements-to-Architecture Traceability

To ensure transparency in how evidence informed the proposed framework, extracted requirements were mapped to three artifact classes: MBOM-PQC schema fields, PQC-safe signing pipeline components, and Supply Chain Assurance Maturity Model (SCAMM) indicators. This traceability matrix demonstrates that the architectural artifacts derive directly from synthesized evidence rather than subjective interpretation and provides a structured basis for future validation, extension, and peer review. The full matrix is maintained as part of the author's research records and is available on request; a summary appears in Section 5.4.

4. Synthesis of Threats and Derived Requirements

AI supply chains introduce a multilayered attack surface spanning data acquisition, model development, dependency management, training pipelines, and deployment workflows. These stages rely on cryptographic mechanisms—signatures, certificates, secure channels, and attestation—that will be disrupted by the transition to post-quantum cryptography (PQC). This section synthesizes evidence from policy documents, standards publications, and documented incidents to characterize the threat landscape and identify the cryptographic dependencies that must be addressed to ensure durable AI supply chain integrity (Figure 2; Figure 3).

 doi:10.20944/preprints202603.1963.v2

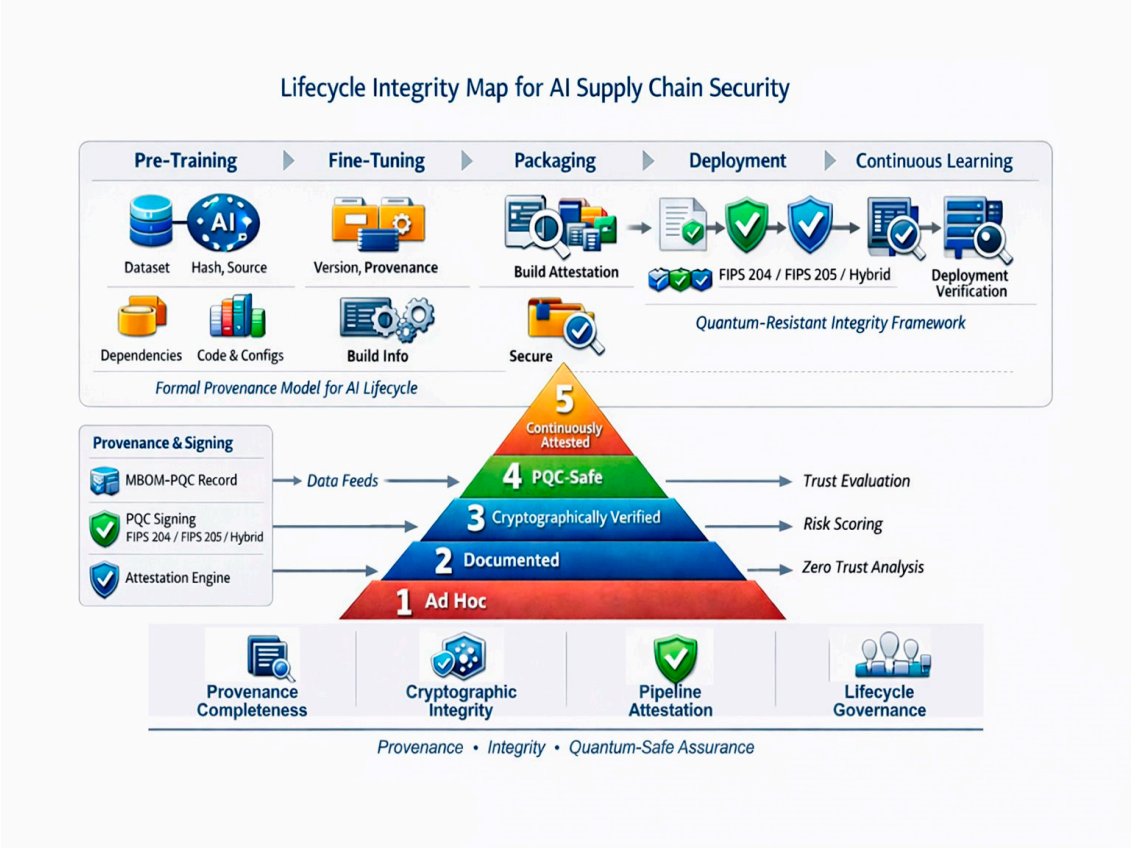

Figure 2. Lifecycle Integrity Map for AI Supply Chain Security. The five-stage AI lifecycle (Pre-Training → Fine-Tuning → Packaging → Deployment → Continuous Learning) with MBOM-PQC provenance capture at each stage. FIPS 204 (ML-DSA, Module-Lattice-Based Digital Signature Algorithm), FIPS 205 (SLH-DSA, Stateless Hash-Based Digital Signature Algorithm), and hybrid signatures are applied through the proposed PQC-Safe Signing Pipeline. The SCAMM (Supply Chain Assurance Maturity Model) pyramid and four assessment dimensions—Provenance Completeness, Cryptographic Integrity, Pipeline Attestation, and Lifecycle Governance—govern organizational readiness.

 doi:10.20944/preprints202603.1963.v2

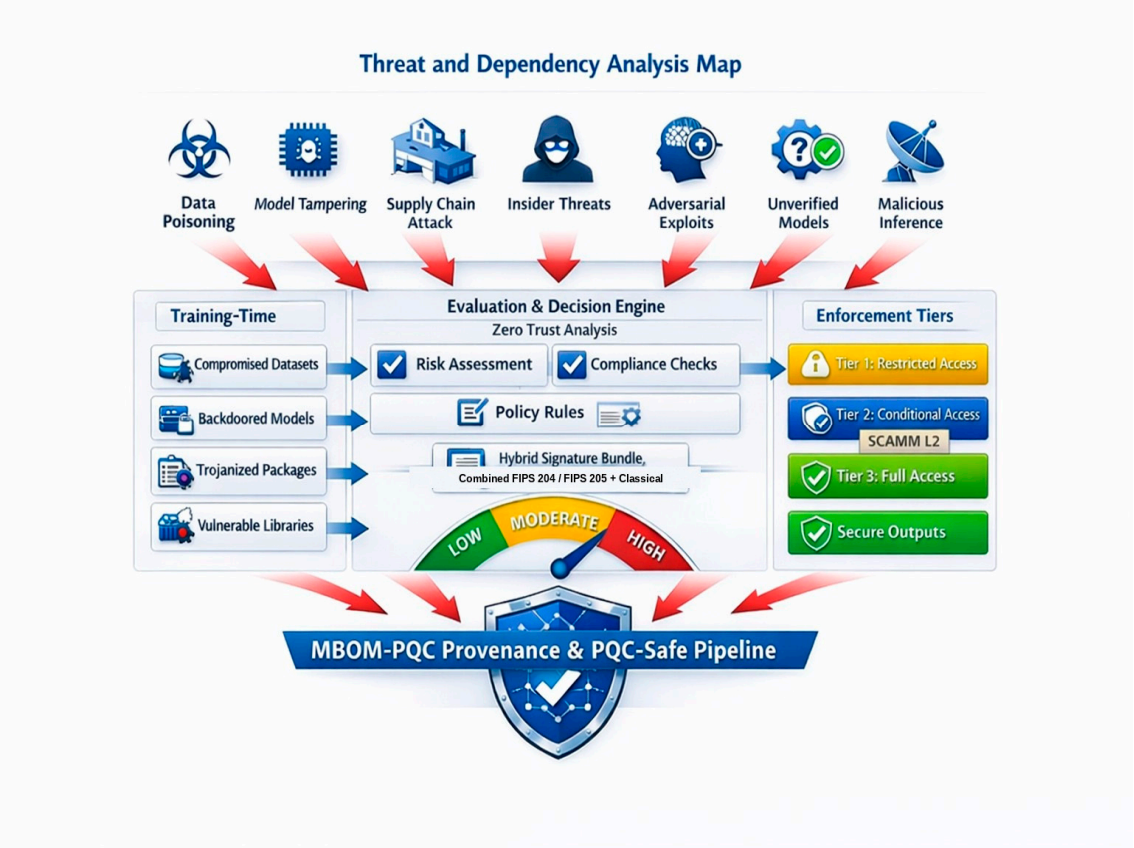

Figure 3. Threat and Dependency Analysis Map. Seven threat vectors—Data Poisoning, Model Tampering, Supply Chain Attack, Insider Threats, Adversarial Exploits, Unverified Models, and Malicious Inference—converge on Training-Time vulnerabilities (Compromised Datasets, Backdoored Models, Trojanized Packages, Vulnerable Libraries). A Zero Trust Evaluation & Decision Engine applies Risk Assessment, Compliance Checks, and Policy Rules, producing a Dynamic Risk Evaluation score driving Enforcement Tiers (Restricted / Conditional / Full Access). All threat vectors are addressed through the proposed MBOM-PQC Provenance & PQC-Safe Pipeline.

4.1. AI Supply Chain Attack Surface

The AI supply chain comprises a sequence of interdependent stages, each with distinct vulnerabilities. Unlike traditional software supply chains, AI artifacts such as model weights, training datasets, and hyperparameter configurations are opaque and difficult to inspect, making cryptographic verification essential.

4.1.1. Training-Time Threats

Training-time compromise remains one of the most damaging forms of AI supply chain attack. Data poisoning allows adversaries to inject manipulated samples into training datasets to bias model behavior [13,14,44]. Model poisoning introduces malicious gradients or updates during distributed or federated training. Backdoor insertion embeds hidden triggers in the model to enable targeted misclassification at inference time [45], a threat class comprehensively evaluated in controlled benchmarks [46]. These attacks exploit the fact that training data and intermediate artifacts often lack cryptographic integrity protections or provenance metadata, allowing tampering to remain undetected through multiple downstream uses of the model.

4.1.2. Ingestion-Time Threats

Organizations frequently ingest pre-trained models from public repositories or third-party vendors. Documented incidents and security research indicate that pre-trained models and

 doi:10.20944/preprints202603.1963.v2

supporting repositories can be tampered with during distribution, and that dependency ecosystems such as PyPI and NPM can be abused to deliver malicious ML libraries [4,15,16,47–49]. Without verifiable provenance, organizations cannot reliably determine whether ingested models originate from trusted sources, making ingestion-time verification a critical control gap.

4.1.3. Deployment-Time Threats

Deployment introduces additional risks, including model tampering during packaging or containerization, unauthorized model updates in continuous deployment pipelines, and inference-time manipulation where adversaries exploit weaknesses in model integrity checks. These threats highlight the need for end-to-end signing and attestation across the entire model lifecycle, extending well beyond the point of initial training to cover every stage at which model artifacts are transferred, transformed, or executed.

4.2. Cryptographic Dependencies in AI Pipelines

AI supply chains rely on cryptographic mechanisms at multiple points, often implicitly. Mapping these dependencies is necessary to identify where PQC transition creates gaps in long-term assurance.

4.2.1. Model Signing and Verification

Model signing is used to authenticate the origin of model artifacts, detect tampering during distribution or deployment, and establish trust boundaries between training and inference environments. Most current implementations use classical signatures (RSA, ECDSA, Ed25519), which are vulnerable to quantum-enabled forgery through harvest-now, forge-later (HNFL) attacks [2,3,11]. As models acquire long operational lifetimes in mission-critical applications, the durability of these signing mechanisms against future quantum attacks becomes a primary assurance concern.

4.2.2. Dataset Integrity and Lineage

Datasets are rarely signed, and when they are, classical signatures are used. PQC transition affects long-term dataset integrity guarantees, lineage verification for sensitive or regulated datasets, and compliance with audit and accountability requirements. The absence of cryptographic dataset provenance means that organizations relying on AI systems in healthcare, defense, or financial services may be unable to demonstrate the integrity of their training data under future regulatory frameworks that require PQC-safe assurance.

4.2.3. Secure Training and Deployment Pipelines

Training pipelines rely on TLS for secure data transfer, certificate chains for authenticating build systems, and secure enclaves for protecting model execution. PQC transition affects all three due to increased key sizes, larger signature sizes, and hybrid mode requirements during the transition period [31,32]. Organizations must plan for certificate chain updates, enclave firmware upgrades, and TLS configuration changes as part of any PQC migration that touches AI infrastructure.

4.2.4. Federated Learning and Distributed Training

Federated learning introduces additional cryptographic dependencies for client authentication, update signing, and aggregator verification [20]. PQC-safe signatures are required to prevent forgery of model updates in long-lived federated systems. As federated deployments scale across organizational boundaries—particularly in defense and healthcare contexts—the communication overhead introduced by larger PQC signature sizes must be explicitly addressed in system design and capacity planning.

4.3. Lifecycle Vulnerabilities Across AI Supply Chains

 doi:10.20944/preprints202603.1963.v2

AI artifacts pass through multiple lifecycle stages, each with distinct integrity requirements and cryptographic dependencies. A comprehensive provenance model must address each stage explicitly.

4.3.1. Pre-Training

Pre-training relies on large, heterogeneous datasets that often lack provenance metadata, integrity verification, and cryptographic lineage. This stage is highly vulnerable to poisoning and dataset manipulation. Because pre-training artifacts form the foundation of all downstream model behavior, integrity failures at this stage propagate silently through every subsequent lifecycle phase.

4.3.2. Fine-Tuning

Fine-tuning introduces new risks through the incorporation of unverified domain-specific datasets, the integration of third-party model checkpoints, and exposure to malicious hyperparameter configurations. Fine-tuning pipelines typically lack signing or attestation mechanisms, creating a window in which tampered base models or poisoned domain data can alter model behavior without detection. The MBOM-PQC schema must capture fine-tuning provenance as a distinct lifecycle component with its own integrity fields.

4.3.3. Packaging and Distribution

Model packaging workflows depend on container signing, artifact registries, and dependency resolution [50]. PQC transition disrupts these mechanisms due to signature size and certificate chain constraints. Organizations must update container signing infrastructure to support FIPS 204 (ML-DSA) or hybrid signatures, and artifact registry configurations must be updated to validate PQC-compatible certificate chains before model distribution can be considered cryptographically assured.

4.3.4. Deployment and Continuous Learning

Deployment introduces model update channels, runtime attestation, and continuous learning loops that all require durable cryptographic guarantees that classical signatures cannot provide. In continuous learning environments, models evolve post-deployment, meaning that each update cycle must be treated as a new supply chain event with its own provenance record, signing event, and attestation check. This requirement extends the scope of AI supply chain security from a point-in-time control to a persistent, lifecycle-spanning governance function.

4.4. Requirements Derived from Threats and Dependencies

Synthesizing the threat landscape and cryptographic dependencies yields four requirement classes that directly inform the design of the MBOM-PQC schema and PQC-safe signing pipeline presented in Sections 5 and 6.

4.4.1. Provenance Requirements

AI supply chains require complete lineage metadata for models, datasets, and dependencies; immutable provenance records; verifiable source attribution; and a standardized schema for AI-specific artifacts. These requirements address the opacity gap identified across all three attack-time phases and provide the foundation for the MBOM-PQC schema defined in Section 5.

4.4.2. Integrity Requirements

Integrity must be ensured through PQC-safe signatures using FIPS 204 (ML-DSA) for operational artifacts and FIPS 205 (SLH-DSA) for non-NSS long-term archival integrity, hybrid signature modes during the transition period, PQC-safe certificate chains, and cryptographically anchored attestation. These requirements respond directly to the cryptographic dependency gaps

 doi:10.20944/preprints202603.1963.v2

identified in Section 4.2 and define the algorithmic and architectural constraints for the signing pipeline introduced in Section 6.

4.4.3. Lifecycle Requirements

Integrity and provenance must be maintained across pre-training, fine-tuning, packaging, deployment, and continuous learning. This requirement reflects the lifecycle vulnerability analysis in Section 4.3 and drives the multi-stage provenance structure of the MBOM-PQC schema, which must capture distinct integrity state at each lifecycle transition rather than treating the model as a static artifact.

4.4.4. Supply Chain Transparency Requirements

Organizations require visibility into model dependencies, verification of third-party components, detection of tampered or malicious artifacts, and integration with Zero Trust and AI RMF governance models [1,51]. These transparency requirements address the systemic opacity of AI supply chains identified in the threat analysis and connect the technical framework to existing enterprise governance structures. In the Supply Chain Assurance Maturity Model (SCAMM) presented in Section 7, this requirement class is operationalized as the Lifecycle Governance dimension, reflecting that supply chain transparency is ultimately sustained through governance structures that enforce it across the model lifecycle. Together, the four requirement classes form the foundation for the formal provenance schema (MBOM-PQC), PQC-safe signing pipeline, and Supply Chain Assurance Maturity Model (SCAMM) introduced in subsequent sections.

5. MBOM-PQC: Proposed Provenance Schema

AI supply chains require a structured, verifiable method for documenting the origin, composition, and integrity of model artifacts. Existing software supply chain mechanisms— such as Software Bills of Materials (SBOMs) defined by SPDX and CycloneDX—provide foundational concepts but lack the semantic richness needed to capture AI-specific artifacts such as training datasets, hyperparameters, pre-trained model dependencies, and fine-tuning workflows. Moreover, current provenance formats rely on classical digital signatures that will not provide long-term protection against quantum-enabled forgery. To address these gaps, this section introduces the Model Bill of Materials with PQC-safe extensions (MBOM-PQC): a formal provenance schema designed to ensure durable, cryptographically anchored trust in AI supply chains throughout the post-quantum transition.

5.1. Design Principles

The MBOM-PQC schema is grounded in four design principles derived from the threat and dependency analysis in Section 4.

5.1.1. Completeness

The schema must capture all artifacts that influence model behavior, including datasets, pre-trained models, hyperparameters, training code, and environmental configurations. Partial provenance is insufficient for detecting poisoning or tampering; incomplete lineage records create audit gaps that adversaries can exploit.

5.1.2. Verifiability

All provenance elements must be cryptographically verifiable using PQC-safe or hybrid signatures. Provenance records must be immutable and resistant to forgery, enabling downstream consumers of model artifacts to independently confirm origin and integrity without relying on the original producer.

 doi:10.20944/preprints202603.1963.v2

5.1.3. Cryptographic Durability

Provenance must remain trustworthy for the operational lifetime of the model, which in defense, healthcare, and critical infrastructure contexts may span decades. This requires PQC-safe signatures using FIPS 204 (ML-DSA) and, for non-NSS archival artifacts, FIPS 205 (SLH-DSA), as well as hybrid signature modes during the transition period to maintain backward compatibility while establishing quantum resistance.

5.1.4. Supply Chain Transparency

The schema must expose dependencies across the entire AI lifecycle, enabling organizations to assess risk, detect tampering, and enforce Zero Trust principles. Transparency is not merely a reporting capability but a structural property: provenance records must be machine-readable, interoperable with existing governance tooling, and auditable by Authorizing Officials and program managers.

5.2. Schema Overview and Core Components

The MBOM-PQC schema is organized into seven core components, each representing a distinct category of AI supply chain artifacts. Together, they form a comprehensive provenance record that can be signed, verified, and attested across the model lifecycle (Figure 4).

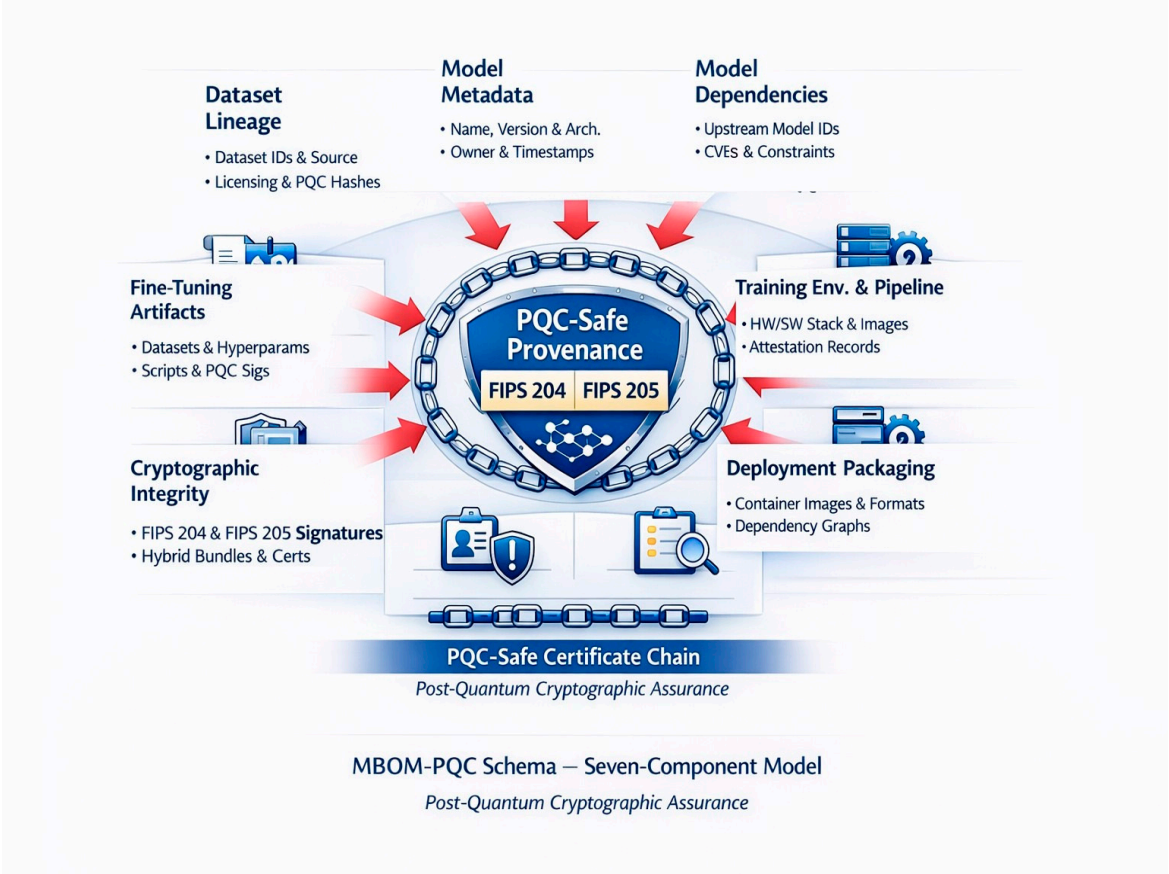

Figure 4. Proposed MBOM-PQC Schema—Seven-Component Provenance Model. Components C1–C7 (detailed in §§5.2.1–5.2.7) converge into a PQC-safe provenance record signed with FIPS 204 (ML-DSA) and FIPS 205 (SLH-DSA) and anchored by a PQC-Safe Certificate Chain.

5.2.1. Component 1: Model Metadata

 doi:10.20944/preprints202603.1963.v2

This component captures high-level information establishing the identity of the model artifact: model name and version, architecture type (e.g., transformer, CNN, diffusion), intended use and deployment context, model owner and publisher, and creation and modification timestamps. These fields provide the anchor to which all downstream provenance components are cryptographically linked.

5.2.2. Component 2: Pre-Training Dataset Lineage

This component documents datasets used during pre-training, including dataset identifiers and versions, source repositories, licensing and usage constraints, data collection methodology, PQC-safe hashes and signatures, and known limitations or biases. Dataset lineage is essential for detecting poisoning and ensuring regulatory compliance; it is the most foundational component because training data determines base model behavior for all subsequent fine-tuning and deployment.

5.2.3. Component 3: Pre-Trained Model Dependencies

This component captures upstream model dependencies, including model identifiers and versions, source repositories such as Hugging Face or vendor registries [52], PQC-safe signatures of upstream models, known vulnerabilities or CVEs, and compatibility constraints. These fields enable verification of model inheritance chains and provide a structured basis for assessing third-party model risk before ingestion.

5.2.4. Component 4: Fine-Tuning Artifacts

This component documents all artifacts used during fine-tuning: fine-tuning datasets, hyperparameters, training scripts and configuration files, random seeds and initialization parameters, and PQC-safe signatures of all artifacts. Fine-tuning is a common point of compromise; detailed provenance at this stage is essential for detecting domain-specific poisoning and for attributing behavioral changes to specific training inputs.

5.2.5. Component 5: Training Environment and Pipeline

This component captures the environment in which training occurred: hardware configuration (CPU/GPU/TPU), software stack (framework versions and libraries), container images and digests, secure enclave or confidential computing details, and pipeline attestation records. These fields support both reproducibility and pipeline integrity verification, enabling auditors to confirm that the training environment matched approved configurations.

5.2.6. Component 6: Deployment Packaging

This component documents packaging and distribution artifacts: container images, model packaging formats (ONNX, TorchScript, TensorRT), PQC-safe signatures of deployment artifacts, and dependency graphs for runtime libraries. It ensures that deployment artifacts can be matched cryptographically to the signed provenance record, closing the gap between training-time integrity and deployment-time verification.

5.2.7. Component 7: Cryptographic Integrity Fields

This component provides the PQC-safe integrity anchors for the entire provenance record: FIPS 204 signatures for model artifacts, FIPS 205 signatures for non-NSS long-term provenance records, hybrid signature bundles combining classical and PQC signatures, PQC-safe certificate chains, and key rotation metadata. This component ensures that provenance remains verifiable throughout the PQC transition and beyond, and serves as the cryptographic root of trust for the entire MBOM-PQC schema (Figure 5).

 doi:10.20944/preprints202603.1963.v2

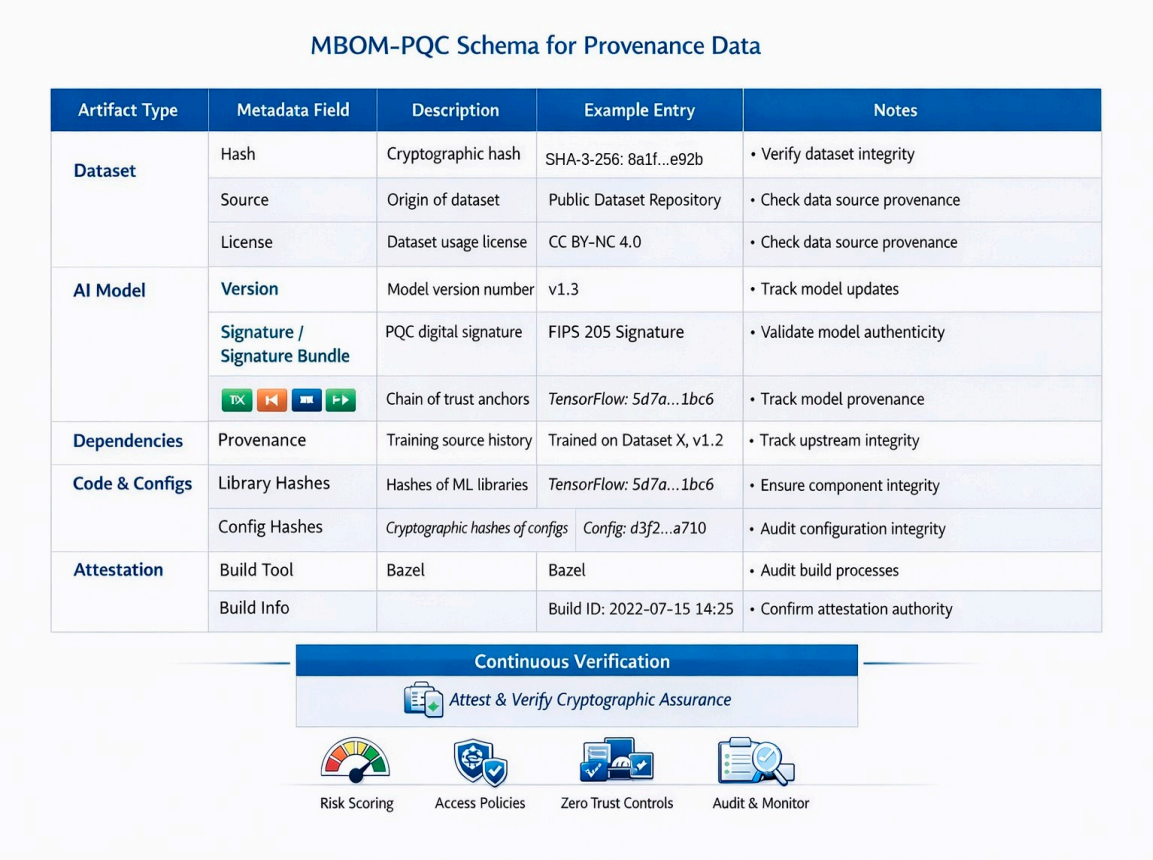

Figure 5. Proposed MBOM-PQC Schema for Provenance Data. The schema captures five artifact types—Dataset (Hash, Source, License), AI Model (ML-DSA/PQC Signing, FIPS 204/FIPS 205 Signatures), Dependencies (Provenance/training source history), Code & Configs (Library Hashes, Config Hashes), and Attestation (Build Tool: Bazel, Build Info: timestamp & ID)—with example entries and integrity notes. Continuous Verification ("Attest & Verify Cryptographic Assurance") underpins the schema, supported by Risk Scoring, Access Policies, Zero Trust Controls, and Audit & Monitor governance functions. Note: The SHA-3-256 hash shown is illustrative; SHAKE is also acceptable for forward compatibility with FIPS 204 and FIPS 205 (see Section 7.2, Level 3).

5.3. PQC-Safe Extensions

The MBOM-PQC schema introduces three PQC-specific extensions that differentiate it from classical provenance models and ensure cryptographic durability across the transition period.

5.3.1. Hybrid Signature Bundles

During the current transition period, informed by emerging CNSA 2.0 and federal PQC migration timelines [11], provenance records must include hybrid signature bundles comprising a classical signature (e.g., ECDSA), a PQC signature (e.g., FIPS 204/ML-DSA), and combined verification metadata. This dual-signing approach ensures backward compatibility with existing verification tooling while establishing PQC-safe integrity guarantees for the model's full operational lifetime.

5.3.2. PQC-Safe Certificate Chains

Certificate chains embedded in provenance records should be designed to support PQC-safe or hybrid certificate-chain evolution, incorporating PQC-safe root certificates, hybrid intermediate certificates, and PQC-safe key usage constraints as the supporting PKI infrastructure matures. This enables end-to-end verification of provenance records and ensures that the trust chain anchoring

 doi:10.20944/preprints202603.1963.v2

model authenticity remains valid even after classical certificate authorities are deprecated or compromised by quantum-capable adversaries.

5.3.3. Long-Term Integrity Anchors

For non-NSS artifacts with multi-year lifetimes, the schema employs FIPS 205 (SLH-DSA) signatures for archival integrity, time-stamped PQC-safe attestations, and immutable provenance logs. FIPS 205 is selected for long-term anchoring because its security is based on hash functions rather than algebraic assumptions, making it the most conservative choice for artifacts that must remain verifiable over extended time horizons where algorithm confidence may evolve.

5.4. Requirements-to-Schema Traceability

Each element of the MBOM-PQC schema maps directly to requirements derived from the threat and dependency analysis in Section 4. Table 3 summarizes this mapping, linking each threat or dependency in Section 4 to the requirement it implies and the MBOM-PQC schema component (C1–C7) that operationalizes it.

Table 3. Requirements-to-MBOM-PQC Schema Traceability Matrix.

Threat Source	Requirement	MBOM-PQC Schema Component
Training-time attacks (§4.1.1)	Dataset poisoning detection	C2: Pre-Training Dataset Lineage
Ingestion-time attacks (§4.1.2)	Model swap prevention	C3: Pre-Trained Model Dependencies
Training-time and ingestion-time threats (§4.1.1, §4.1.2)	Fine-tuning tampering detection	C4: Fine-Tuning Artifacts
Pipeline compromise (§4.2.3)	Pipeline integrity	C5: Training Environment & Pipeline
Quantum-enabled forgery (§4.2.1)	PQC-safe integrity	C7: Cryptographic Integrity Fields
Multi-stage supply chain (§4.3)	Lifecycle transparency	All components (C1–C7)

C = Component number in MBOM-PQC schema. Section references correspond to Section 4 threat analysis.

6. PQC-Safe Signing and Attestation: Proposed Pipeline

6.1. Pipeline Overview

The pipeline comprises five sequential stages: Ingestion, Verification, Signing, Attestation, and Deployment (Figure 6). Each stage produces a distinct cryptographic output that feeds the next, creating an unbroken chain of provenance from artifact acquisition through operational use. The pipeline is designed to be modular, allowing organizations to implement individual stages incrementally in alignment with their SCAMM maturity level, while ensuring that the full five-stage sequence satisfies the integrity and attestation requirements derived in Section 4.4. The architecture-layer view (Figure 7) illustrates how Key Management, Signing Service, Attestation Engine, and Policy Enforcement components support this workflow.

 doi:10.20944/preprints202603.1963.v2

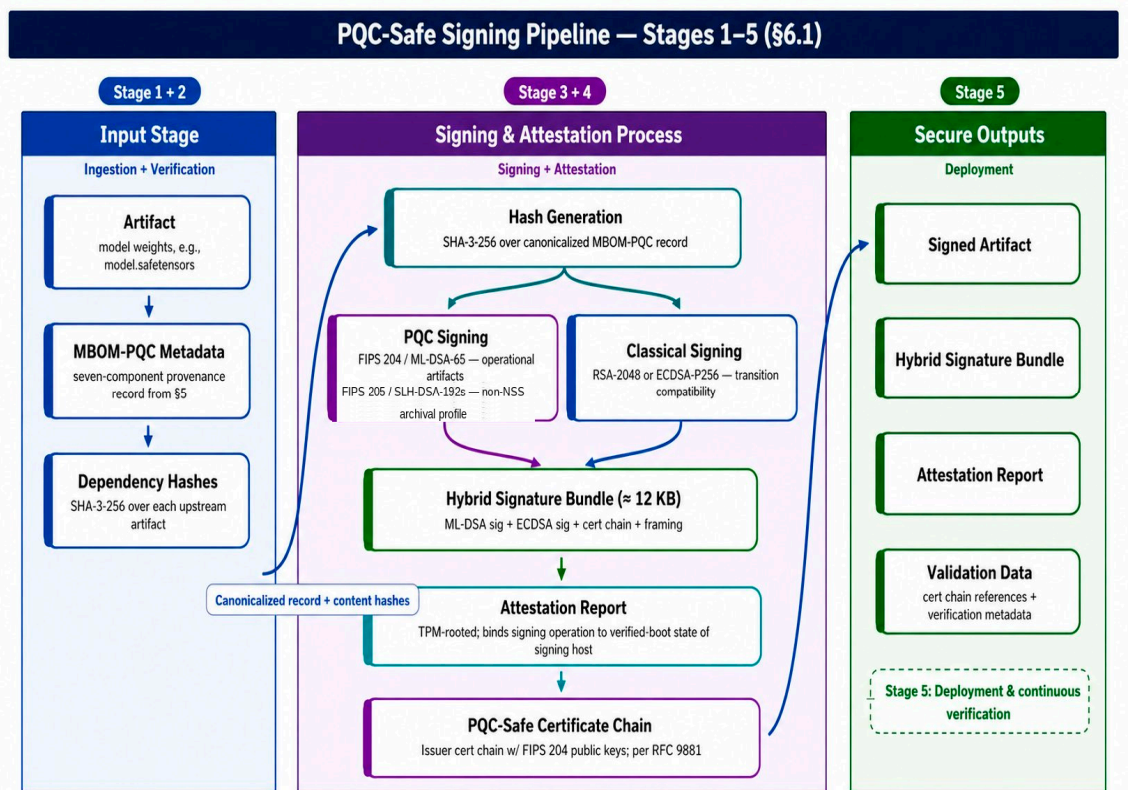

Figure 6. PQC-Safe Signing Pipeline for discrete model release. The five sequential stages of §6.1 are organized into three visual phases: Input Stage (Stages 1–2), Signing & Attestation Process (Stages 3–4), and Secure Outputs (Stage 5). PQC Signing applies FIPS 204 [2] / ML-DSA-65 to operational artifacts, with CNSA 2.0 [11] mandating ML-DSA-87 for NSS deployments, and FIPS 205 [3] / SLH-DSA-192s for default non-NSS archival per Table 4; classical co-signing (ECDSA-P256 or RSA) preserves backward verifier compatibility during the transition. Continuous-learning pipeline modes (full re-sign, delta-sign, batched-checkpoint) for non-discrete-release scenarios are specified in Section S2 of the Supplementary Materials with Figure S1.

 doi:10.20944/preprints202603.1963.v2

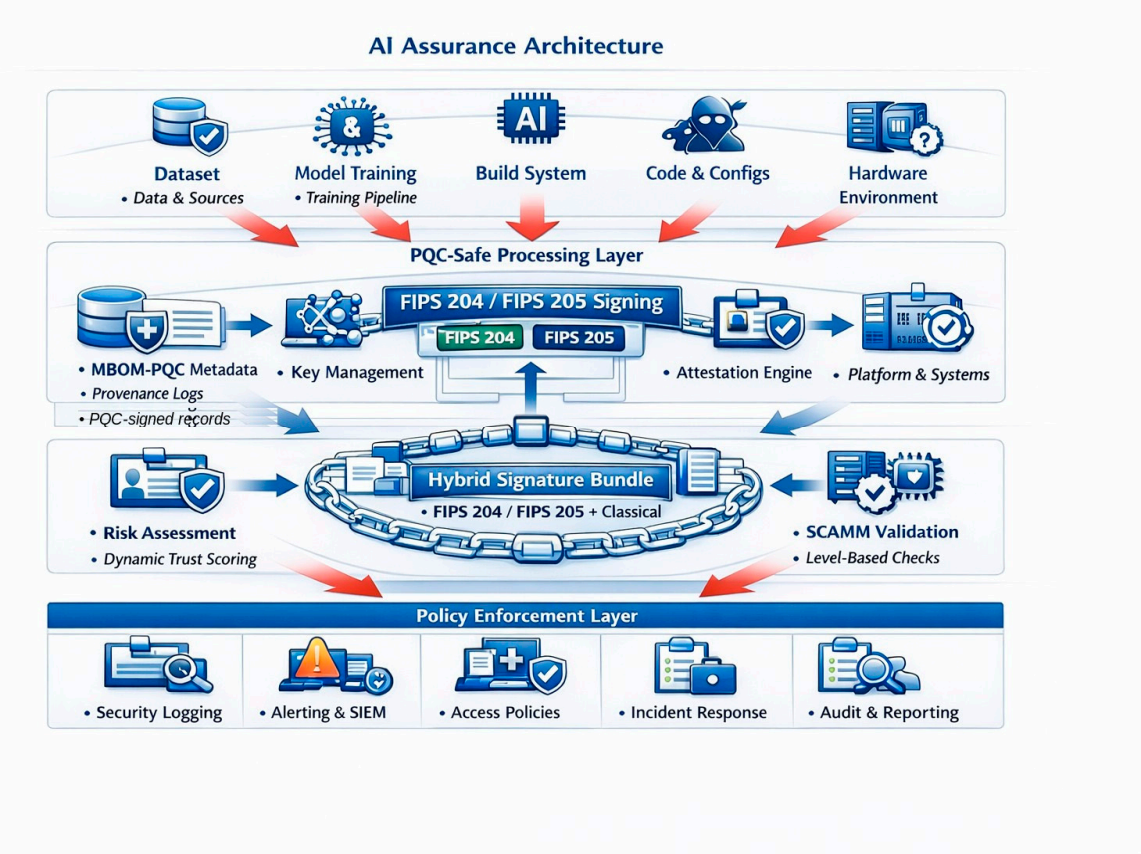

Figure 7. AI Assurance Architecture: layered view of PQC-safe processing, signing, and policy enforcement. The architecture is organized into four tiers: input artifacts; PQC-Safe Processing Layer (FIPS 204 / FIPS 205 signing, key management, attestation); Hybrid Signature Bundle (PQC + classical signing with risk scoring and SCAMM validation); and Policy Enforcement Layer (logging, alerting, access policies, audit). Detailed component labels appear in the figure.

6.1.1. Stage 1—Ingestion

During ingestion, model artifacts, datasets, and dependencies are acquired from internal or external sources. Each artifact is catalogued against its MBOM-PQC provenance record, and a cryptographic hash is computed and recorded. Source provenance is checked against known repositories and vendor attestations. Artifacts without verifiable provenance are quarantined pending manual review. This stage establishes the artifact inventory that all subsequent pipeline stages act upon.

6.1.2. Stage 2—Verification

During verification, existing signatures on ingested artifacts are validated against the MBOM-PQC record. In hybrid mode, PQC-capable verifiers validate both the classical and PQC legs of the bundle; legacy verifiers may validate the classical leg only, with reduced transition-period assurance. Artifacts bearing only classical signatures are flagged for re-signing at the next stage. Certificate chain validation confirms that signing keys trace to a trusted PQC-safe root. Verification failures halt pipeline progress and trigger incident response workflows, preventing potentially compromised artifacts from advancing toward deployment.

6.1.3. Stage 3—Signing

During signing, verified artifacts are signed using the algorithm mode appropriate to their expected lifetime and sensitivity. Standard model artifacts receive FIPS 204 signatures. Long-lived non-NSS archival artifacts—including provenance records, dataset manifests, and training

 doi:10.20944/preprints202603.1963.v2

environment snapshots—receive FIPS 205 signatures. During the transition period, hybrid bundles combining a classical ECDSA signature with the appropriate PQC signature are generated and stored with the artifact. All signatures are recorded in the MBOM-PQC Cryptographic Integrity Fields (Component 7).

6.1.4. Stage 4—Attestation

During attestation, a hardware-rooted attestation record is generated confirming that the signing process executed in a verified, approved environment. This record binds the signed artifact to the specific hardware, firmware, and software configuration of the signing platform, using TPM-based or secure enclave attestation. The attestation report is itself signed with a PQC-safe key and appended to the MBOM-PQC record. Remote attestation enables downstream consumers to verify pipeline environment integrity independently of the signing organization.

6.1.5. Stage 5—Deployment

During deployment, the signed and attested artifact is released to the operational environment. The deployment system verifies the MBOM-PQC record, confirms signature validity, and checks the attestation report before permitting execution. Deployment gate checks are logged to the provenance record, creating an auditable trail from artifact origin through operational activation. Post-deployment, any model update triggers re-entry into the pipeline at Stage 1, ensuring that continuous learning environments maintain the same integrity guarantees as initial deployments.

6.2. PQC-Safe Signing Flow

The signing flow within Stage 3 implements a three-mode signing architecture that adapts to the artifact type, lifetime, and organizational PQC readiness level.

6.2.1. Hybrid Mode Signing

Hybrid mode is the recommended default during the current transition period, informed by CNSA 2.0 and federal PQC migration timelines [11]. Each artifact receives two concurrent signatures: a classical ECDSA or Ed25519 signature for backward compatibility with existing verification infrastructure, and a FIPS 204 (ML-DSA) signature providing quantum resistance. Both signatures are stored in the MBOM-PQC Cryptographic Integrity Fields and are independently verifiable. Verifiers that have not yet migrated to PQC tooling can still validate the classical signature, while PQC-capable verifiers validate both, gaining stronger assurance.

6.2.2. FIPS 204 (ML-DSA) Signing for Standard Artifacts

ML-DSA (FIPS 204) is used as the primary signing algorithm for model weights, packaging artifacts, and pipeline execution records. ML-DSA-65 is the recommended parameter set, balancing signature size (3,309 bytes, approximately 3.3 KB) against security level (NIST Level 3) [2]. Organizations with constrained distribution channels may use ML-DSA-44 (NIST Level 2) for internal artifacts, reserving ML-DSA-87 (NIST Level 5) for high-assurance artifacts deployed in national security or critical infrastructure contexts.

While ML-DSA-65 is presented as the default for the civilian/commercial profile throughout §6.2, organizations operating in National Security Systems (NSS) or other high-assurance contexts require distinct parameter selections. Table 4 specifies four algorithm profiles aligned with deployment context; profile selection is contextual, and one parameter set does not fit all use cases.

Table 4. Algorithm Profile Selection by Deployment Context. ML-DSA-65 is presented as the default civilian/commercial profile in this manuscript and underpins the cost analysis in §8.3.5 (Table 8). NSS deployments require ML-DSA-87 (NIST Level 5) per CNSA 2.0 [11]; FIPS 205 (SLH-DSA) is not currently

 doi:10.20944/preprints202603.1963.v2

approved for NSS use, so the NSS profile applies ML-DSA-87 to both operational signing and archival integrity. Analogous size and cost calculations apply with proportional but bounded increases.

Profile	Operational signing	Long-term archival	Hash	Use context
Constrained / Internal	ML-DSA-44 (NIST L2)	—	SHA-3-256 or SHAKE-256	Internal-only artifacts; short-lived non-regulated commercial settings
Civilian / Commercial (default)	ML-DSA-65 (NIST L3)	SLH-DSA-192s (NIST L3)	SHA-3-256 or SHAKE-256	Federal non-NSS; regulated industries (healthcare, finance) and critical infrastructure with multi-decade retention requirements; SLH-DSA-128s (NIST L1) may be applied as a documented profile exception for shorter-horizon (single-digit-year), non-regulated commercial archival per §6.2.3
High-Assurance / non-NSS	ML-DSA-87 (NIST L5)	SLH-DSA-256s (NIST L5)	SHA-3-512 or SHAKE-256	Highest-assurance non-NSS deployments; regulated industries and critical infrastructure requiring NIST Level 5 strength; SLH-DSA-256s suitable for long-term archival where hash-based forward security is desirable
NSS / CNSA 2.0	ML-DSA-87 (NIST L5)	ML-DSA-87 (NIST L5)	SHA-3-512 or SHAKE-256	National Security Systems; classified workloads; CNSA 2.0–mandated procurement (ML-DSA-87 / Category 5 [11]); FIPS 205 / SLH-DSA is not approved for NSS, so ML-DSA-87 covers both operational signing and archival integrity (re-signed at policy cadence)

6.2.3. FIPS 205 (SLH-DSA) for Long-Term Artifacts

SLH-DSA (FIPS 205) is reserved for artifacts that must remain verifiable over multi-year horizons: dataset manifests, provenance logs, training environment snapshots, and archival MBOM-PQC records. SLH-DSA's hash-based construction is conservative relative to ML-DSA's lattice-based assumptions, providing higher confidence in long-term security at the cost of larger signature sizes (approximately 8–50 KB depending on parameter set). Parameter set selection by retention period and mission criticality is specified in Table 4 (§6.2.2). FIPS 205 (SLH-DSA) is not currently approved for NSS use; NSS deployments apply ML-DSA-87 per CNSA 2.0 [11] for both operational signing and archival integrity. Across security categories, the 𝓼' (small-signature) variants are generally preferred over 𝓯' (fast-signing) variants for archival use, since signing is a one-time cost while signature size is amortized over the artifact's full retention period.

6.2.4. PQC-Safe Key Management

PQC signing keys must be generated, stored, and rotated in alignment with CNSA 2.0 transition expectations and NIST standardization guidance for FIPS 204 (ML-DSA) for National Security System contexts, and applicable NIST guidance for FIPS 205 (SLH-DSA) deployments in non-NSS federal

 doi:10.20944/preprints202603.1963.v2

and commercial contexts. FIPS 204 private keys must be stored in hardware security modules (HSMs) or secure enclaves that support PQC key operations; FIPS 205 keys, used for non-NSS archival integrity, are subject to the same hardware storage requirements. Key rotation schedules must be documented in the MBOM-PQC record, and historical signing keys must be retained and protected to support retrospective verification of previously signed artifacts. Organizations must plan for the larger key storage footprint introduced by PQC algorithms relative to classical counterparts.

6.2.5. Worked Example: 110M-Parameter Transformer Checkpoint

To make the signing flow concrete, this subsection traces the application of the schema (§5) and the pipeline (§6.1–§6.2.4) to a representative artifact: a fine-tuned 110M-parameter Transformer encoder distributed in the safetensors format (FP32, ≈ 440 MB).

The artifact ingestion stage (§6.1, Stage 1) catalogs three input files — the model weights file, the tokenizer configuration, and the model card — and computes SHA-3-256 hashes over each. (SHA-3 is selected over SHA-256 for forward compatibility with FIPS 204 and FIPS 205, per §5.3.3 and the qualifying note in the Figure 5 caption.) The hashes populate Component C1 (Model Metadata) and C7 (Cryptographic Integrity Fields) of the MBOM-PQC record. The dataset lineage component (C2) is populated from the model card's documented training data sources, with a separate hash recorded for each upstream dataset reference; pre-trained model dependencies (C3) record the upstream foundation model identifier and its own MBOM-PQC reference if available. The fine-tuning artifacts component (C4) records the fine-tuning dataset hash, the hyperparameter configuration, and a hash of the fine-tuning script. The training environment component (C5) records the build platform identifier, library versions, and the TPM-rooted attestation quote (§6.4.1) of the signing host.

The signing stage (§6.1, Stage 3) generates two signatures over the canonicalized MBOM-PQC record: (i) an ML-DSA-65 signature (3,309 bytes) for the operational artifact, per FIPS 204 [2]; and (ii) an ECDSA-P256 signature (≈ 71 bytes) over the same payload, for backward verifier compatibility. The two signatures, together with the certificate chain (which itself includes the PQC-safe public key and the issuing CA chain), form the hybrid bundle of total size ≈ 11.8 KB recorded in C7. For long-lived archival of the same artifact's provenance manifest (rather than the model weights themselves), an SLH-DSA-192s signature (16,224 bytes) is generated in addition, per the default civilian/commercial archival profile in Table 4; this artifact is recorded as a separate record in C7 with its own integrity field.

The attestation stage (§6.1, Stage 4) generates a TPM-rooted attestation report binding the signing operation to the verified-boot state of the signing host. The attestation report is itself signed with the host's PQC-safe attestation key and appended to the MBOM-PQC record. At the deployment stage (§6.1, Stage 5), the consuming system verifies the hybrid bundle against the recorded certificate chain, validates the attestation report, and consults the policy enforcement schema (§6.4) before permitting the artifact to load.

Across this end-to-end flow, the absolute bytes added to the artifact distribution by the MBOM-PQC record and the hybrid bundle total ≈ 11.8 KB regardless of the underlying model size. For the 440 MB Transformer in this example, the relative overhead is ≈ 0.0026% of the artifact size; for larger model classes the relative overhead decreases asymptotically. The full quantitative analysis across model scales is presented in §8.3.5. A machine-readable JSON-LD instantiation of the complete MBOM-PQC record for this Transformer example is provided in the Supplementary Materials.

6.3. Attestation Architecture

Attestation extends the assurance provided by signing by binding signed artifacts to the verified state of the environment in which they were processed. This transforms provenance from a claim about artifact content into a verifiable statement about the entire signing context.

6.3.1. Hardware Root of Trust

 doi:10.20944/preprints202603.1963.v2

The attestation architecture is anchored in a hardware root of trust, implemented via TPM 2.0 or equivalent secure enclave technology, consistent with platform resiliency guidance [53] (Figure 6). The hardware root provides an unforgeable measurement of the platform state at signing time, including firmware, bootloader, operating system, and signing application configuration. This measurement is recorded in the attestation report and signed with a PQC-safe endorsement key. Organizations migrating to PQC must ensure that their TPM or enclave firmware supports PQC endorsement key operations; hybrid (classical + PQC) attestation paths are available during the transition period; hardware upgrades may be required in legacy environments. Practical migration considerations for organizations whose existing TPM and HSM inventory does not yet support FIPS 204 / FIPS 205 natively — including phased adoption strategies and a qualitative cost matrix — are addressed in §8.3.6. The pipeline architecture treats hardware-root-of-trust upgrade as a configurable maturity dimension (mapping to SCAMM Level 4 in §7.2) rather than a Day-1 prerequisite, allowing organizations to advance through earlier maturity levels using software-rooted attestation while hardware refresh proceeds.

6.3.2. PQC-Safe Certificate Chains

Attestation reports and model signatures are bound to a certificate-chain design that can evolve toward PQC-safe or hybrid trust anchors during transition, extending from a trusted root certificate authority to the signing platform's endorsement key. During the transition period, hybrid certificate chains are used: each certificate in the chain carries both a classical and a PQC signature, allowing verification by both legacy and PQC-capable verifiers. Organizations operating in federal environments must align certificate chain updates with GSA and CISA PQC PKI migration guidance [42,43] and CNSA 2.0 timelines [11].

6.3.3. Remote Attestation

Remote attestation enables a relying party to verify the integrity of the signing environment without physical access to the signing platform. The attestation report—a signed, hardware-rooted statement of platform state [53]—is published alongside the signed artifact and can be independently verified by any party with access to the endorsement key certificate chain. This capability is particularly valuable in federated or multi-organization supply chains where trust must be established across organizational boundaries without requiring bilateral agreements about physical infrastructure.

6.4. Integration with Zero Trust Architecture and AI RMF

The PQC-safe signing and attestation pipeline is designed to integrate with existing enterprise governance frameworks rather than operate as a standalone security control. Two integration points are particularly important for federal and defense deployments.

6.4.1. Zero Trust Architecture Integration

Within a Zero Trust Architecture (ZTA), every model artifact is treated as untrusted until verified. The pipeline operationalizes this principle by producing a pipeline-verified trust score for each artifact, derived from signature validity, attestation report status, provenance completeness, and dependency verification outcome. This trust score is surfaced to ZTA policy decision points, enabling dynamic access controls that restrict model deployment or inference based on real-time supply chain integrity status. Models with incomplete provenance, expired signatures, or failed attestation are denied deployment regardless of network location or user identity (Figure 8). For multi-cloud and cloud-native AI deployments, NIST SP 800-207A [54] further specifies ZTA-based access-control mechanisms for which the pipeline-derived trust scores in this manuscript provide an AI-artifact-level assurance input.

 doi:10.20944/preprints202603.1963.v2

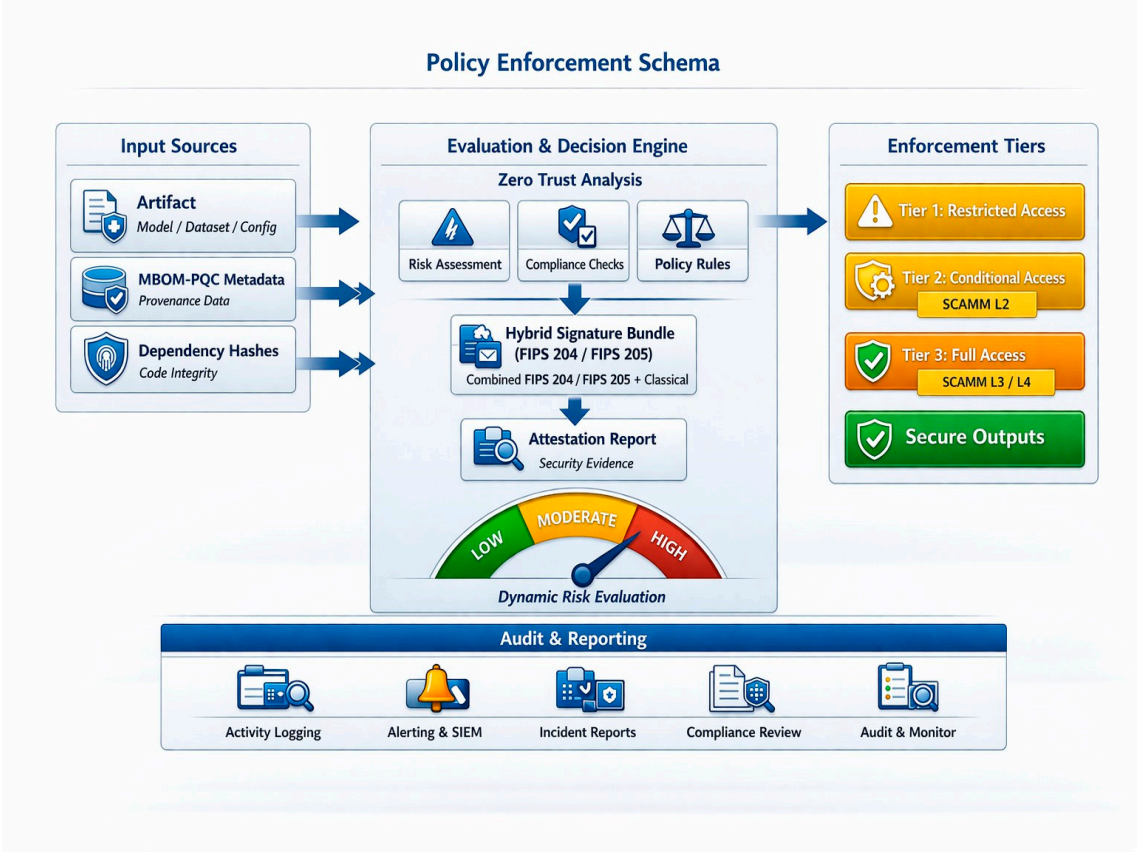

Figure 8. Proposed Policy Enforcement Schema. Zero Trust Evaluation & Decision Engine receives Input Sources (Artifact, MBOM-PQC Metadata, Dependency Hashes) and applies Risk Assessment, Compliance Checks, and Policy Rules. Dynamic Risk Evaluation (Low/Moderate/High) drives Enforcement Tiers: Tier 1 Restricted Access, Tier 2 Conditional Access (SCAMM L2), Tier 3 Full Access (SCAMM L3/L4), and Secure Outputs. Audit & Reporting: Activity Logging, Alerting & SIEM, Incident Reports, Compliance Review, and Audit & Monitor.

6.4.2. AI RMF Integration

The pipeline maps to the AI RMF GOVERN, MAP, MEASURE, and MANAGE functions [1]. GOVERN: pipeline policies are documented in organizational governance frameworks. MAP: provenance records identify AI artifacts and their dependencies. MEASURE: signature validity, attestation status, and provenance completeness provide measurable supply chain integrity indicators. MANAGE: trust scores and verification outcomes inform deployment, decommissioning, and incident response decisions. This integration ensures that pipeline outputs are immediately actionable within established AI governance processes, avoiding the creation of parallel security tracks competing for organizational attention.

6.5. Continuous-Learning Pipeline Modes

The five-stage pipeline of §6.1 specifies the discrete-release case, in which an artifact is acquired, verified, signed, attested, and deployed once. Continuous-learning deployments — incremental fine-tuning on streaming data, federated update aggregation, or scheduled retraining — require additional structure, because executing the full five-stage pipeline at every iteration may impose computational and signing-key-management overhead that exceeds the operational benefit. Three operational modes are defined to address this need: *full re-sign* (the complete pipeline is re-executed and a new MBOM-PQC record is generated), *delta-sign* (the weight delta against a prior signed checkpoint is signed and appended to the existing record), and *batched-checkpoint* (signing is applied at scheduled cumulative checkpoints rather than per-iteration). Mode selection is governed by a three-input decision rule on update frequency, criticality tier, and weight-change magnitude, with

 doi:10.20944/preprints202603.1963.v2

security-critical updates and L2-norm threshold violations forcing full re-sign regardless of frequency. The complete mode specifications, the formal decision rule, and the corresponding workflow diagram are provided in the Supplementary Materials. Open research questions concerning these modes — formal verification of delta-sign correctness, bounded-staleness guarantees for batched-checkpoint, and the interaction with SCAMM maturity levels — are discussed in §8.4.3.

7. SCAMM: Proposed Maturity Model

7.1. SCAMM Overview

The Supply Chain Assurance Maturity Model (SCAMM) defines five cumulative maturity levels that enable organizations to assess their current AI supply chain security posture, identify gaps relative to PQC transition requirements, and develop a structured improvement roadmap. Each level builds on the preceding one, reflecting increasing cryptographic assurance, provenance coverage, and governance integration. SCAMM is grounded in the requirements derived in Section 4.4 and is operationalized through the four assessment dimensions described in Section 7.3: Provenance Completeness, Cryptographic Integrity, Pipeline Attestation, and Lifecycle Governance (Figure 9).

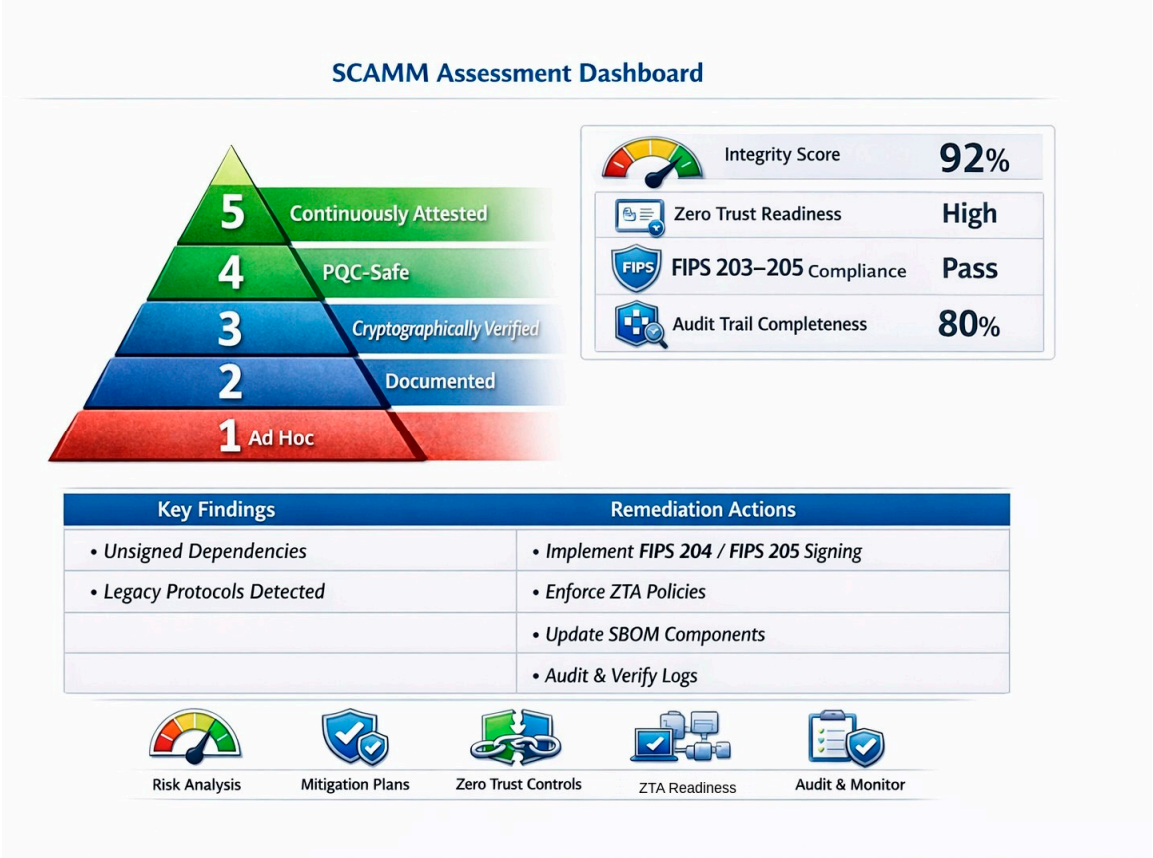

Figure 9. Illustrative SCAMM Assessment Dashboard from the worked example in §7.3.5. The displayed scores correspond to D_prov = 0.95, D_crypto = 0.92, D_attest = 0.88, D_gov = 0.93, yielding an arithmetic-mean aggregate of 92% and a weakest-link maturity placement at Level 4, with Level 5 blocked by Cryptographic Integrity, Pipeline Attestation, and Lifecycle Governance (Pipeline Attestation representing the largest gap) per Tables 5 and 6 (§7.3.5). The dashboard visualization is a *proposed reference workflow* for organizational reporting; the displayed indicator framing aligns with the four assessment dimensions defined in §§7.3.1–7.3.4. Note: signing and provenance requirements within the SCAMM model are scoped to FIPS 204 (ML-DSA) and FIPS 205 (SLH-DSA); FIPS 203 (ML-KEM) addresses key encapsulation rather than digital signatures and is included in the dashboard's PQC-coverage indicator for completeness of NIST PQC standards.

 doi:10.20944/preprints202603.1963.v2

7.2. Maturity Level Definitions

Level 1: Ad Hoc—Minimal Assurance. Organizations at Level 1 lack formal AI supply chain controls. There is no standardized provenance documentation, no model or dataset signing, no verification of third-party model sources, and no attestation of training or deployment pipelines. Cryptographic mechanisms, where present, rely solely on classical signatures without PQC planning. This level reflects the current state of many organizations adopting AI rapidly without corresponding security controls, and represents the baseline from which all maturity progression begins.

Level 2: Documented—Foundational Provenance. At Level 2, organizations begin establishing basic supply chain documentation: partial provenance records covering datasets, model versions, and dependencies; manual verification of third-party model sources; and classical signatures applied inconsistently to model artifacts. Training environment visibility is limited, and no PQC-safe mechanisms are in place, though PQC transition planning has begun. Level 2 provides foundational transparency but lacks cryptographic durability. The primary diagnostic indicator is whether provenance records exist at all—even partial, manually maintained records represent a meaningful advance over Level 1.

Level 3: Cryptographically Verified—Classical Integrity Controls. At Level 3, organizations implement consistent integrity mechanisms: complete MBOM-style provenance records (without PQC extensions), classical signatures applied to all model artifacts, automated verification of model and dataset integrity, and basic pipeline attestation such as container signing and build verification. Hash selection prioritizes SHA-3 or SHAKE for alignment with FIPS 205 and design-diversity future-proofing [2,3], with the chosen hash function meeting the required security strength for the artifact class. Level 3 provides strong classical integrity but remains vulnerable to quantum-enabled forgery via HNFL attacks. Organizations at this level have the operational infrastructure needed to transition directly to Level 4 through algorithm substitution and key migration activities.

Level 4: PQC-Safe—Quantum-Resistant Integrity. At Level 4, organizations adopt PQC-safe mechanisms across the AI lifecycle: the full MBOM-PQC schema is implemented; hybrid signature bundles combining classical and FIPS 204 (ML-DSA) signatures are applied to all artifacts; PQC-safe certificate chains are deployed for model signing keys; and training and deployment pipelines produce PQC-safe attestation records. Level 4 ensures that AI artifacts remain verifiable throughout the PQC transition and constitutes the minimum target posture for organizations operating AI systems in national security, defense, or critical infrastructure contexts under CNSA 2.0 timelines.

Level 5: Continuously Attested—Zero Trust-Aligned AI Supply Chain. At Level 5, organizations achieve continuous, end-to-end assurance: real-time verification of model integrity during deployment and inference; continuous attestation of training, deployment, and runtime environments; automated detection of provenance drift or unauthorized model updates; and full integration with Zero Trust Architecture trust scoring. PQC-only signatures using FIPS 204 (ML-DSA) are applied for all operational artifacts; for non-NSS long-term archival records, FIPS 205 (SLH-DSA) provides additional hash-based integrity anchoring. Continuous learning workflows include PQC-safe update signing and provenance extension at each update cycle. Level 5 represents a fully mature, Zero Trust-aligned AI supply chain with durable, PQC-safe integrity guarantees, and is the long-term target posture for high-assurance AI deployments in persistent threat environments.

7.3. SCAMM Indicators and Metrics

Each maturity level is evaluated using measurable indicators across four dimensions. These indicators enable repeatable, evidence-based assessment that can be reported to governance bodies and integrated into Authorization to Operate (ATO) packages, risk management frameworks, and supply chain security audits.

7.3.1. Provenance Completeness

 doi:10.20944/preprints202603.1963.v2

Provenance completeness is measured as the percentage of artifacts with MBOM-PQC coverage across all seven schema components, supplemented by dataset lineage completeness (proportion of training datasets with signed lineage records) and upstream model dependency transparency (proportion of ingested models with verified provenance chains). These metrics directly reflect the threat requirements identified in Section 4.4.1 and provide the primary evidence base for Levels 2 through 5 assessment.

7.3.2. Cryptographic Integrity

Cryptographic integrity is measured as the percentage of artifacts signed with PQC-safe or hybrid signatures, PQC-safe certificate chain coverage (proportion of signing key chains rooted in a PQC-safe CA), and key rotation and lifecycle management compliance against documented key management policy. These indicators distinguish Level 3 (classical-only) from Level 4 (hybrid/PQC-safe) and provide the quantitative evidence required to demonstrate CNSA 2.0 alignment in federal authorization packages.

7.3.3. Pipeline Attestation

Pipeline attestation is measured across three sub-dimensions: build attestation coverage (proportion of model artifacts with hardware-rooted signing environment records), training pipeline attestation coverage (proportion of training runs with verified environment attestation), and deployment and runtime attestation frequency (cadence of post-deployment integrity re-verification). These indicators become mandatory at Level 4 and must be continuous at Level 5, reflecting the shift from point-in-time controls to persistent operational verification.

7.3.4. Lifecycle Governance

Lifecycle governance is measured through four indicators: integration with ZTA trust scoring (whether supply chain risk scores from the pipeline are consumed by ZTA policy decision points), continuous learning update verification (whether all post-deployment model updates are re-processed through the full signing and attestation pipeline), policy enforcement for third-party model ingestion (whether organizational policy gates prevent deployment of unverified externally sourced models), and automated provenance drift detection (whether the organization has tooling to detect unauthorized changes to model artifacts or provenance records between validation cycles). These indicators distinguish Level 5 organizations from Level 4 by confirming that supply chain assurance is operationally embedded, continuously enforced, and lifecycle-spanning rather than periodically audited.

7.3.5. Scoring Methodology

To make the SCAMM evaluation mathematically reproducible, this subsection specifies the scoring formula used to compute each dimension score, the aggregation rule that determines the organization's maturity level from the four dimension scores, and the default sub-indicator weights. Organizations may customize the weights to reflect sector-specific priorities, provided the customization is documented and the aggregation rule is preserved.

Each of the four assessment dimensions — D_prov (Provenance Completeness, §7.3.1), D_crypto (Cryptographic Integrity, §7.3.2), D_attest (Pipeline Attestation, §7.3.3), and D_gov (Lifecycle Governance, §7.3.4) — is computed as a weighted sum of its sub-indicators:

$$D\_k = \Sigma\_j \; w\_\{k,j\} \cdot i\_\{k,j\} \text{ for } k \in \{\text{prov, crypto, attest, gov}\}$$

where $0 \leq i\_\{k,j\} \leq 1$ is the j-th sub-indicator value for dimension k (typically expressed as a proportion, such as the fraction of artifacts with verified provenance), and $0 \leq w\_\{k,j\} \leq 1$ is its weight, with $\Sigma\_j \; w\_\{k,j\} = 1$ within each dimension. Default weights are derived from the requirements-

 doi:10.20944/preprints202603.1963.v2

traceability mapping in §§5.4 and 7.4: indicators that satisfy more requirement classes receive proportionally higher weight. Default weights are tabulated in Table 5.

The aggregate organizational maturity level uses a *weakest-link rule* rather than an arithmetic mean over the four dimensions, reflecting the supply-chain principle that overall integrity is bounded by the weakest dimension:

$$L\_org = \max \{ L : \forall k, D\_k \geq \tau\_\{L,k\} \}$$

where τ_{L,k} is the threshold for dimension k at maturity level L. The thresholds are tabulated in Table 6. The weakest-link rule prevents a single strong dimension (for example, excellent provenance documentation) from masking a weakness in another dimension (for example, missing pipeline attestation).

Worked example. The illustrative dashboard in Figure 9 reports an aggregate score of 92%. Under the scoring methodology specified above, this score is the arithmetic mean of D_prov = 0.95, D_crypto = 0.92, D_attest = 0.88, D_gov = 0.93. Applying the weakest-link rule with the Table 6 thresholds: at Level 4, the per-dimension thresholds are τ_4 = (0.85, 0.85, 0.85, 0.85), all of which the four scores exceed; at Level 5, the thresholds tighten to τ_5 = (0.95, 0.95, 0.95, 0.95), and D_attest = 0.88 falls below the Level-5 attestation threshold; D_crypto = 0.92 and D_gov = 0.93 also fall below 0.95, and only D_prov = 0.95 just meets it. The organization is therefore placed at SCAMM Level 4, with Level 5 blocked by Cryptographic Integrity, Pipeline Attestation, and Lifecycle Governance — Pipeline Attestation representing the largest gap and therefore the priority for the dashboard's remediation guidance. The worked example illustrates how the scoring methodology produces both an aggregate score and an actionable remediation pointer.

Table 5. Default Sub-Indicator Weights for SCAMM Dimension Scoring. Weights sum to 1.000 within each dimension. Organizations may customize weights with documentation; sector-specific recommendations are provided in §7.3.5.

Dimension	Sub-indicator	Default weight
D_prov (Provenance Completeness)	Model metadata coverage	0.20
	Pre-training dataset lineage coverage	0.30
	Pre-trained model dependency coverage	0.20
	Fine-tuning artifact coverage	0.20
	Deployment packaging dependency coverage	0.10
D_crypto (Cryptographic Integrity)	Proportion of artifacts with PQC-safe signatures	0.40
	Proportion with hybrid signature bundles	0.20
	Certificate-chain validity rate	0.20
	Cryptographic agility readiness score	0.20
D_attest (Pipeline Attestation)	Build attestation coverage	0.40
	Training pipeline attestation coverage	0.30
	Deployment and runtime attestation cadence	0.30
D_gov (Lifecycle Governance)	ZTA trust scoring integration	0.30
	Continuous-learning update verification	0.20
	Third-party model ingestion policy enforcement	0.20
	Automated provenance drift detection	0.30

Table 6. Per-Level Dimension Thresholds (τ_{L,k}). An organization is placed at the highest level L for which all four dimension scores meet or exceed the corresponding row.

 doi:10.20944/preprints202603.1963.v2

Level	τ_prov	τ_crypto	τ_attest	τ_gov
L1 (Ad Hoc)	0.00	0.00	0.00	0.00
L2 (Documented)	0.50	0.40	0.30	0.50
L3 (Cryptographically Verified)	0.70	0.65	0.55	0.70
L4 (PQC-Safe)	0.85	0.85	0.85	0.85
L5 (Continuously Attested)	0.95	0.95	0.95	0.95

7.4. Requirements-to-Maturity Mapping

SCAMM directly reflects requirements derived from the threat and dependency analysis in Section 4. Table 7 maps each key requirement to its corresponding maturity levels and provides the derivation rationale grounding each assignment in documented threats and cryptographic dependencies.

Table 7. Requirements-to-SCAMM Maturity Level Traceability Matrix. Each row traces a requirement to (a) the threat class or cryptographic dependency that motivated it, (b) the evidence sources from the synthesis that support it, with confidence-tier labels per §3.7, (c) the schema component or pipeline stage that operationalizes it, and (d) the SCAMM indicator that measures it.

#	Requirement	Threat / dependency	Evidence (Tier)	Operationalization	SCAMM indicator
1	Pre-training dataset lineage capture	Training-time data poisoning (§4.1.1)	[13] T3, [14] T3	Schema C2	i_{prov,2}
2	Pre-trained model dependency tracking	Ingestion-time tampering (§4.1.2)	[4] T4, [16] T4, [19] T3	Schema C3	i_{prov,3}
3	Fine-tuning artifact provenance	Fine-tuning tampering (§4.3.2)	[14] T3, [44] T3	Schema C4	i_{prov,4}
4	Training environment attestation	Pipeline compromise (§4.2.3)	[51] T1, [55] T2	Schema C5; Pipeline Stage 4	i_{attest,2}
5	Deployment packaging integrity	Deployment-time tampering (§4.1.3)	[10] T1, [55] T2	Schema C6; Pipeline Stage 5	i_{prov,5}
6	PQC-safe signing of artifacts	Quantum-enabled forgery, HNFL (§§1, 4.2.1)	[2] T1, [11] T2	Schema C7; Pipeline Stage 3	i_{crypto,1}
7	Hybrid signature support during transition	Backward verifier compatibility (§5.3.1)	[56] T5 (primary); [31] T5, [32] T5 (transition context)	Schema C7; Pipeline Stage 3	i_{crypto,2}
8	Certificate-chain PQC-safe validation	Long-term chain integrity (§5.3.2)	[57] T1	Schema C7; Pipeline Stage 2	i_{crypto,3}
9	Lifecycle attestation cadence	Multi-stage compromise (§4.3)	[51] T1, [55] T2	Pipeline Stages 1–5	i_{attest,3}
10	Continuous verification	Continuous-learning integrity (§§4.3.4, 6.5)	[51] T1	Pipeline Stage 5; §6.5	i_{attest,3}
11	ZTA trust scoring integration	Governance gap (§4.4.4)	[1] T2, [23] T1, [58] T2	§6.4	i_{gov,1}

 doi:10.20944/preprints202603.1963.v2

#	Requirement	Threat / dependency	Evidence (Tier)	Operationalization	SCAMM indicator
12	Continuous-learning update verification and third-party model ingestion control	Operational risk drift (§4.3.4)	[23] T1, [24] T2	Discussion §8	i_{gov,2,3}
13	Automated provenance drift detection	Trust scoring (§§6.4, 8.2)	[51] T1, [54] T2	§6.4	i_{gov,4}

Tier labels (T1–T5) refer to the evidence-confidence tiers defined in §3.7. Section references correspond to the threat and dependency analysis (Section 4), schema/pipeline sections (Sections 5–6), and the new continuous-learning subsection (§6.5).

8. Discussion

8.1. Implications for AI Governance and Risk Management

The introduction of MBOM-PQC and the PQC-Safe Signing Pipeline has significant implications for AI governance. First, the framework operationalizes key elements of the NIST AI RMF [1]—particularly transparency, accountability, and security—by providing a structured method for documenting and verifying model lineage. Second, the integration of PQC-safe signatures ensures that provenance records remain trustworthy throughout the expected operational lifetime of AI systems, addressing a gap not currently covered by existing AI governance standards. Third, the SCAMM maturity model provides a roadmap for organizations to align AI supply chain assurance with broader modernization efforts such as Zero Trust Architecture and PQC migration. These implications extend beyond compliance. By enabling verifiable provenance and continuous attestation, the framework enhances organizational resilience against model poisoning, dependency compromise, and supply chain manipulation. It also supports mission-critical environments—such as defense, healthcare, and critical infrastructure—where long-term integrity guarantees are essential for ATO and ongoing risk management.

8.2. Integration with Zero Trust Architecture and Enterprise Security

Beyond the artifact-level ZTA integration described in §6.4.1, the pipeline supports enterprise-wide cryptographic agility. By abstracting PQC-safe signing and certificate management into shared services, organizations can modernize AI supply chains without requiring each development team to independently implement PQC-safe mechanisms. This reduces duplication, improves consistency, and accelerates compliance with CNSA 2.0 [11] and NIST PQC transition guidance. The result is a supply chain assurance capability that scales across large, heterogeneous AI portfolios without proportional increases in implementation complexity. Industry security frameworks, including Google's Secure AI Framework (SAIF) [59] and Microsoft's AI security research program [60], identify analogous supply chain integrity requirements, reinforcing the broader ecosystem need for systematic provenance controls and cryptographic assurance.

8.3. Implementation Challenges

Despite its benefits, implementing the proposed framework presents several challenges that organizations must plan for explicitly.

8.3.1. Performance and Storage Overhead

PQC signatures—particularly FIPS 204 (ML-DSA) and FIPS 205 (SLH-DSA)—are significantly larger than classical signatures, increasing storage requirements for provenance records, bandwidth

 doi:10.20944/preprints202603.1963.v2

consumption during model distribution, and verification time during deployment and inference. While these overheads are manageable in enterprise environments with adequate storage and network capacity, they may pose challenges for constrained or tactical systems where bandwidth is limited and latency is critical. Organizations should benchmark signature verification performance against their deployment SLAs before selecting parameter sets. Quantitative analysis of the overhead across model scales is provided in §8.3.5.

8.3.2. Legacy System Compatibility

Many existing AI pipelines rely on classical cryptographic libraries, legacy certificate chains, and proprietary model formats that do not support PQC operations. Integrating PQC-safe signing requires updating build systems, modifying deployment workflows, replacing or upgrading cryptographic libraries, and ensuring backward compatibility during transition. Hybrid signature modes mitigate some compatibility challenges by allowing legacy verifiers to validate the classical signature component, but they do not eliminate the need for infrastructure updates. Organizations operating long-lived AI systems should develop explicit migration roadmaps that sequence infrastructure updates ahead of CNSA 2.0 compliance deadlines.

8.3.3. Provenance Completeness

Capturing complete provenance—particularly for pre-trained models and third-party datasets—may be difficult when upstream providers do not supply sufficient metadata or signatures. Public model repositories vary widely in provenance transparency, and many commercially available pre-trained models are distributed without signed lineage records or training environment documentation. Organizations may need to establish procurement requirements or contractual obligations specifying provenance standards as a condition of third-party model acquisition. Until ecosystem-level transparency norms mature, organizations operating at SCAMM Level 4 or above may need to generate best-effort provenance records through reverse-engineering and internal attestation of ingested models.

8.3.4. Organizational Maturity and Skill Gaps

Achieving higher SCAMM levels requires cryptographic expertise, secure pipeline engineering, governance alignment, and cross-team coordination that many organizations currently lack. AI development teams are typically focused on model performance rather than supply chain security, and cryptography teams may not have deep familiarity with ML pipeline architectures. Bridging this gap requires deliberate workforce development, cross-functional security reviews, and organizational structures that embed supply chain security requirements into AI development workflows from the outset rather than as a post-deployment retrofit.

8.3.5. Performance Overhead Across Model Scales

The MBOM-PQC overhead is bounded in absolute terms regardless of model size, because the schema and the hybrid bundle attach to the artifact rather than scaling with it. Table 8 makes this concrete by tabulating the bytes added and the verification time across five representative model scales spanning four orders of magnitude. ML-DSA cycle counts in this analysis are taken from Table 1 of the CRYSTALS-Dilithium Round 3 Specification [21], which reports median cycles for the AVX2-optimized implementation on Intel Skylake (median of 1,000 executions, 32-byte message). The Round 3 specification figures are used here rather than later third-party benchmarks because they are authoritative for the algorithm as standardized in FIPS 204 and are reproducible from the public reference implementation at https://github.com/pq-crystals/dilithium. Conversion to wall-clock time uses a 3.3 GHz reference clock. SLH-DSA signature sizes are taken directly from FIPS 205 [3]. SHA-3-256 throughput for hash-cost estimation is taken at approximately 500 MB/s software (commodity x86 without SHA-3 hardware acceleration) and approximately 5 GB/s with hardware acceleration

 doi:10.20944/preprints202603.1963.v2

available on ARMv8 cores with the SHA3 cryptographic extension; these figures correspond to typical eBASH/SUPERCOP-class measurements for Keccak-f-1600.

Table 8. Performance Overhead Across Model Scales for MBOM-PQC Hybrid Signing. ML-DSA-65 verify time computed from 179,424 cycles on Skylake AVX2 [21, Table 1] at 3.3 GHz reference clock = 54 µs. Bundle overhead is the sum of ML-DSA-65 signature (3,309 B per FIPS 204), ECDSA-P256 signature (≈ 71 B), hybrid PQC certificate chain (≈ 6.4 KB), and JSON-LD framing overhead (≈ 2 KB). The verify column reports cryptographic-core overhead only — SHA-3-256 hash computation over the artifact plus ML-DSA-65 signature verification — and does not include JSON-LD canonicalization, certificate path validation, MBOM-PQC schema parsing, I/O, registry/transparency-log latency, or policy-engine (ZTA) evaluation. SHA-3-256 hash times are computed at assumed throughput of approximately 500 MB/s for software-only execution on commodity x86 without SHA-3 hardware acceleration, and approximately 5 GB/s with hardware acceleration available on ARMv8 cores with the SHA3 cryptographic extension (FEAT_SHA3); both figures are typical of published Keccak-f-1600 benchmarks reported in the eBASH/SUPERCOP measurement suite (https://bench.cr.yp.to) and on the Keccak Team reference page (https://keccak.team/sw_performance.html), and are presented here as order-of-magnitude assumptions for asymptotic comparison across model scales rather than absolute performance predictions for any specific platform; per-deployment benchmarking on the target hardware is required for SLA planning. These end-to-end pipeline costs are largely independent of model size, are bounded in absolute terms, and are deferred for empirical characterization to the validation roadmap in §8.4.4.

Model class	Artifact size	Bundle overhead	Relative overhead	SHA-3-256 hash time (sw / hw)	ML-DSA-65 verify [21]	Cryptographic-core verify (sw / hw)
Small Transformer (BERT-base)	50 MB	11.8 KB	0.023%	100 ms / 10 ms	54 µs	≈ 100 ms / ≈ 10 ms
Mid-tier classifier	500 MB	11.8 KB	0.0023%	1.0 s / 100 ms	54 µs	≈ 1.0 s / ≈ 100 ms
7B-parameter LLM (FP16)	14 GB	11.8 KB	8.0×10^{-5} %	28 s / 2.8 s	54 µs	≈ 28 s / ≈ 2.8 s
70B-parameter LLM (FP16)	140 GB	11.8 KB	8.0×10^{-6} %	280 s / 28 s	54 µs	≈ 280 s / ≈ 28 s
Frontier checkpoint (FP16)	350 GB	11.8 KB	3.2×10^{-6} %	700 s / 70 s	54 µs	≈ 700 s / ≈ 70 s

Two observations follow from Table 8. First, the bytes added by MBOM-PQC are O(1) in model size: the hybrid bundle totals approximately 11.8 KB regardless of whether the underlying artifact is 50 MB or 350 GB. (The Round 3 specification [21, Table 1] reports a slightly smaller signature size of 3,293 bytes; the final FIPS 204 [2] size of 3,309 bytes reflects post-Round-3 refinements and is used in this analysis.) Second, **the verification time is dominated entirely by hash computation over the artifact, not by signature verification**: even at 350 GB, the ML-DSA-65 verify operation contributes only 54 µs out of a cryptographic-core verification time measured in tens to hundreds of seconds, with hash computation over the artifact dominating that cost. The PQC contribution to cryptographic-core verification cost is therefore negligible at all model scales considered, and the practical bottleneck for large-artifact cryptographic-core verification is the choice of hash function and the availability of hash hardware acceleration, not the choice of post-quantum signature algorithm. End-to-end pipeline verification additionally incurs JSON-LD canonicalization, certificate path validation, MBOM-PQC schema parsing, I/O, registry/transparency-log latency, and policy-engine evaluation; these are largely model-size-independent fixed costs and are deferred for empirical characterization to the validation roadmap in §8.4.4.

For long-lived archival artifacts, FIPS 205 SLH-DSA [3] is the appropriate selection and provides hash-based security as a hedge against future cryptanalytic results against lattice-based schemes.

 doi:10.20944/preprints202603.1963.v2

SLH-DSA signatures are larger than ML-DSA: 7,856 bytes for SLH-DSA-128s, 16,224 bytes for SLH-DSA-192s (the default civilian/commercial archival profile per Table 4), 29,792 bytes for SLH-DSA-256s, and 49,856 bytes for the maximum-security SLH-DSA-256f parameter set (FIPS 205 Table 1). Verification is correspondingly slower, with published x86 verify-time benchmarks for SLH-DSA-128s in the low-millisecond range. The same asymptotic conclusions hold: PQC contribution to verification cost is bounded by a few milliseconds and is dominated by hash-computation cost for any artifact above approximately 100 MB. For tactical-edge deployments operating under sub-100 KB bandwidth budgets and intermittent connectivity, ML-DSA-44 [2] (signature size 2,420 bytes; verify time ≈ 36 μs at 3.3 GHz from the 118,412-cycle Skylake AVX2 figure in [21, Table 1]) is a viable selection at the cost of reduced security margin (NIST Level 2 vs. Level 3); the pipeline architecture supports parameter-set selection per artifact class.

8.3.6. Hardware Root-of-Trust Migration

The pipeline architecture in §6 assumes that signing and attestation infrastructure can produce ML-DSA and SLH-DSA signatures with hardware-rooted assurance. As of the publication of this manuscript, the Trusted Computing Group is integrating ML-DSA support into the PC Client TPM Profile (2025 draft revision) [41], but the installed base of fielded TPM 2.0 hardware does not yet provide native FIPS 204 or FIPS 205 capabilities; the HSM landscape is similarly transitional, with PQC-capable FIPS 140-3 modules lagging the publication of the underlying NIST standards. Organizations operating long-lived AI systems may therefore face hardware-refresh costs, firmware-update requirements, or both as a precondition for reaching SCAMM Level 4. Three transition strategies — *software-rooted attestation* (interim, no hardware refresh, software-managed PQC keys), *hybrid-bundle bridging* (hardware-rooted classical signature plus software-bound PQC leg, valid through the CNSA 2.0 transition window), and *phased hardware refresh* (PQC-capable HSMs and TPMs deployed in line with existing hardware lifecycle) — together with a qualitative cost matrix comparing them along capital expenditure, operational expenditure, timeline, PQC durability, and hardware-rooted assurance dimensions are presented in the Supplementary Materials. The framework treats hardware-root-of-trust upgrade as a configurable maturity dimension (mapping to SCAMM Level 4 in §7.2) rather than a Day-1 prerequisite, allowing organizations to advance through earlier maturity levels using software-rooted attestation while hardware refresh proceeds.

8.4. Limitations of the Proposed Framework

The framework has several limitations that constrain its current scope and warrant future research.

8.4.1. Evolving PQC Standards

PQC algorithms and certificate formats continue to evolve. While ML-DSA (FIPS 204) [2], SLH-DSA (FIPS 205) [3], and ML-KEM (FIPS 203) [22] are standardized, additional algorithms may be introduced, and performance optimizations may change implementation guidance. IETF drafts for hybrid key exchange [31,32], PQC certificate profiles, and composite signature formats are still in progress. Specific implementation choices made today may require revision as standards mature. The framework is designed to be algorithm-agnostic at the architectural level, but specific parameter set recommendations should be reviewed against current NIST and IETF guidance at the time of implementation.

8.4.2. Dependency on Upstream Transparency

The framework assumes that upstream model providers, dataset curators, and library maintainers supply sufficient metadata and signatures to populate MBOM-PQC records. In practice, transparency varies widely across providers, and many commercially distributed models lack the provenance documentation necessary to achieve SCAMM Level 3 or above without supplementation.

 doi:10.20944/preprints202603.1963.v2

Adoption may require ecosystem-level incentives, regulatory requirements mandating provenance disclosure for AI artifacts used in regulated sectors, or the development of community-maintained provenance registries analogous to existing software vulnerability databases. Without broader ecosystem participation, the framework's effectiveness will be bounded by the provenance quality of the most opaque components in the supply chain.

8.4.3. Continuous-Learning Open Questions

The operational modes for continuous-learning deployments are specified in §6.5. Three open research questions remain. First, the formal verification of *delta-sign correctness* — that is, the conditions under which the integrity guarantee provided by signing the weight delta against a prior signed checkpoint is equivalent to the integrity guarantee that would be provided by signing the full updated artifact — has not been treated rigorously in the published cryptographic literature for the AI-checkpoint case. Second, *bounded-staleness guarantees* for the batched-checkpoint mode require a specification of the maximum acceptable delay between an integrity-relevant event (a data-poisoning attempt, a model-tampering attempt) and its detection at the next signed checkpoint; the appropriate bound depends on the deployment criticality and is not yet codified. Third, the *interaction between continuous-learning mode and SCAMM maturity level* deserves explicit treatment: an organization at SCAMM Level 5 ("Continuously Attested") operating a high-frequency online-learning system needs guidance on how the continuous-attestation requirement maps onto the batched-checkpoint mode, and whether such mapping admits a Level-5 placement at all. These three questions together form a coherent agenda for follow-on research.

8.4.4. Empirical Validation Roadmap

The artifacts proposed in this manuscript are design-science constructions and have not yet been empirically validated. A revised version of this work, and follow-on publications, will pursue empirical validation along three coordinated tracks. **Track 1: Proof-of-concept implementation.** A reference implementation of the five-stage signing and attestation pipeline (§6.1) targeting the ML-DSA-65 and SLH-DSA-192s parameter sets, instrumented with liboqs and OpenSSL 3.x with PQC providers, will be evaluated against the Transformer-class models tabulated in §8.3.5. Reported metrics will include signing latency, verification latency, attestation round-trip time, and MBOM-PQC payload size, with comparison against the analytical bounds in §8.3.5. **Track 2: Schema-coverage evaluation.** The MBOM-PQC schema will be applied to a sample of publicly available pre-trained model artifacts spanning three model classes (encoder-only, decoder-only LLM, vision foundation model) and three distribution channels (Hugging Face Hub, GitHub Releases, vendor model registries) to evaluate schema-population coverage and identify the upstream-transparency gaps that require ecosystem-level remediation. **Track 3: SCAMM Delphi-style expert evaluation.** A multi-round Delphi study with practitioners from defense, healthcare, and regulated commercial sectors will evaluate the SCAMM dimension definitions, default weights (Table 5), and per-level thresholds (Table 6) for face validity, sectoral applicability, and inter-rater reliability. The Delphi protocol will be specified in a separate methodological paper. The three tracks together will move the framework from a design-science contribution to an operationally validated standard.

Ahead of this broader program, the first concrete validation step will deliver a reference implementation over a small model registry testbed of approximately five publicly available pre-trained models, exercising four core capabilities together: JSON-LD record generation per Listing S1 of the Supplementary Materials, ML-DSA-65 signing of the canonicalized record, hybrid verification combining ML-DSA-65 with ECDSA-P256, and SCAMM scoring against the four assessment dimensions of §7.3. This minimum viable demonstration is scoped to expose any integration gaps among the schema (§5), the signing pipeline (§6.1), and the maturity model (§7), and to provide concrete code and data artifacts that reviewers and prospective adopters can examine before the broader Track 1 performance evaluation and Track 2 schema-coverage study are completed.

 doi:10.20944/preprints202603.1963.v2

8.5. Critical Considerations and Boundary Conditions

The discussion to this point has emphasized obstacles to adopting the proposed framework. A complete critical assessment also requires examining whether parts of the framework may be over-engineered for some deployment classes, whether adoption is contingent on ecosystem conditions that may not materialize, and how the framework would manifest differently across plausible operational scenarios. This subsection addresses each in turn.

8.5.1. Where MBOM-PQC May Be Over-Engineered

The full schema and pipeline target high-assurance environments — defense, healthcare, regulated financial services, critical infrastructure — where AI artifacts have multi-year operational lifetimes and where the cost of integrity failure is high. Several deployment classes lie outside this assurance envelope and may not warrant the full framework. *Short-lived models in non-regulated settings* (consumer-facing recommendation systems retired within a 12-month window, internal-only research artifacts with no production deployment) may not require PQC-safe signing because the operational lifetime expires before HNFL becomes a credible threat; for these cases, the schema's classical-signature-only mode (a hybrid bundle without the PQC leg) provides documentation value without the full cryptographic apparatus. *Pre-deployment research artifacts* (training-loop checkpoints, intermediate fine-tuning states) may be adequately covered by lightweight provenance (Components C1, C2, C5 of the schema only) without the full seven-component record. The framework's modularity supports these reduced configurations, and §7.2's Level-2 ("Documented") definition explicitly accommodates organizations whose risk profile does not justify higher levels.

8.5.2. Ecosystem Conditions That May Not Materialize

Several adoption pathways depend on ecosystem-level developments that are uncertain. First, *upstream provenance disclosure norms*: many publicly distributed pre-trained models lack signed lineage records or training-environment documentation, and adopting organizations are dependent on upstream providers to populate Components C2, C3, and C5 of the schema. Without procurement-driven incentives or regulatory mandates, ecosystem-wide provenance disclosure may not emerge at the scale required for the framework to be broadly applicable. Second, *PQC-capable hardware-root-of-trust availability*: as discussed in §8.3.6, native PQC support in TPMs and HSMs is currently lagging the standardization of the underlying algorithms; if this gap persists, organizations will be forced into the software-rooted-attestation strategy as a long-term posture rather than as an interim measure, with reduced assurance properties. Third, *regulator-driven adoption pressure*: the framework's value at SCAMM Levels 4 and 5 depends on regulatory or contractual pressure that justifies the implementation cost. If regulatory pressure emerges only in narrow sectors (defense, perhaps healthcare), adoption beyond those sectors may be slower than the threat-landscape analysis in §4 would warrant. The framework remains useful in each of these conditional cases — but its impact is bounded by the ecosystem conditions rather than by its internal coherence alone.

8.5.3. Illustrative Scenarios

Two illustrative scenarios — neither empirically grounded, both intended as design-time aids for organizations applying SCAMM — show how the framework would manifest differently across plausible deployment contexts.

Scenario A: Defense-mission AI deployment. A federal program office deploys a fine-tuned vision-language model to a tactical edge environment. Operational lifetime is bounded by the deployment cycle (≈ 5 years). The classification regime requires verifiable provenance and supply chain integrity. The CNSA 2.0 timeline applies. Under SCAMM, this organization targets Level 4 (PQC-Safe) at deployment with a Level-5 (Continuously Attested) target by the end of the deployment cycle. Schema components C1–C7 are fully populated; the pipeline operates in full re-sign mode (per §6.5)

 doi:10.20944/preprints202603.1963.v2

with daily attestation refresh; HRoT migration is on the phased-hardware-refresh strategy of §8.3.6. The full framework applies.

Scenario B: Commercial generative-AI service. A consumer-facing software vendor distributes a fine-tuned LLM for use in a hosted productivity service. Operational lifetime per model version is short (≈ 6 months between major version increments). Regulatory pressure is limited to the OWASP-style and ISO/IEC 42001 governance categories. Under SCAMM, this organization targets Level 2 (Documented) initially, with Level 3 (Cryptographically Verified) targeted within an 18-month modernization horizon. Schema components C1, C2, C3, and C7 are fully populated; C4 and C5 are partially populated with best-effort metadata; C6 is fully populated through the existing CI/CD pipeline. Signing uses classical algorithms initially, transitioning to hybrid bundles via the bridging strategy of §8.3.6 ahead of the broader CNSA 2.0 sunset. The framework applies in a reduced configuration appropriate to the assurance requirement.

8.6. Opportunities for Future Research

The framework opens several avenues for future research. Automated provenance extraction tools could reduce the manual effort required to populate MBOM-PQC records from legacy or third-party artifacts. PQC-safe model registries and distribution protocols could establish community infrastructure for signed model hosting analogous to package repositories in software development. Integration with confidential computing and secure enclaves could extend attestation guarantees to model execution environments, not just signing environments. AI-specific certificate profiles for PQC-safe signing could standardize how FIPS 204 and FIPS 205 keys are bound to organizational identities in AI supply chain contexts. Empirical validation of certificate-chain interoperability across heterogeneous PKI environments and end-to-end attestation workflow performance would provide the operational evidence needed to transition the framework from architectural prescription to deployment-ready standard. Formal verification of provenance completeness could provide mathematical guarantees that a given MBOM-PQC record captures all material influences on model behavior. Finally, systematic benchmarking of PQC signature performance across representative AI pipeline configurations would provide the empirical data needed to finalize parameter set recommendations and validate the framework's practical applicability at scale.

9. Conclusions

The growing dependence of AI systems on multi-stage supply chains — pre-trained foundation models, third-party datasets, open-source libraries, and automated training and deployment pipelines — has created a class of integrity vulnerabilities that existing AI governance and software supply chain frameworks do not fully address. The transition to post-quantum cryptography compounds this gap: classical digital signatures used to attest AI artifact integrity will not provide durable assurance over the operational lifetimes typical of AI deployments in defense, healthcare, transportation, and critical infrastructure. The harvest-now, forge-later pattern proposed in §1 makes the temporal mismatch explicit.

The MBOM-PQC schema, the PQC-safe signing and attestation pipeline, and the SCAMM maturity model proposed here address this gap as integrated design-science artifacts. The contributions are *prescriptive rather than empirically demonstrated*. The framework is grounded in a structured evidence synthesis (54 sources, traceable via the requirements-to-architecture matrices in §§5.4, 7.4) and is internally consistent; it has not yet been validated through proof-of-concept implementation, performance benchmarking against real AI pipelines, or expert evaluation. Section 8.4.4 specifies the empirical validation roadmap that constitutes immediate follow-on work. Section 8.5 acknowledges the boundary conditions — ecosystem-level provenance disclosure norms, PQC-capable hardware-root-of-trust availability, regulator-driven adoption pressure — under which the framework's practical impact is realized.

Within those boundary conditions, the framework offers a structured approach to AI supply chain assurance that connects provenance, cryptographic durability, and organizational maturity

 doi:10.20944/preprints202603.1963.v2

into a single auditable architecture. As AI systems are increasingly embedded in mission-critical environments and as the post-quantum transition advances along the timelines specified by NSA CNSA 2.0 [11] and OMB M-23-02 [12], the case for cryptographically durable AI provenance grows stronger. The framework presented here is offered as a foundation for AI supply chain assurance in the Future Internet ecosystem of cloud-connected, mission-critical, and distributed smart systems, contingent on the validation activities identified in §8.4.4.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. Supplementary File S1 contains four supporting sections: S1 — PRISMA-screened evidence bibliography; S2 — continuous-learning pipeline modes; S3 — hardware-root-of-trust migration strategies; and S4 — worked MBOM-PQC JSON-LD instantiation.

Author Contributions: Conceptualization, R.C.; methodology, R.C.; investigation, R.C.; writing—original draft preparation, R.C.; writing—review and editing, R.C. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new primary data were created or analyzed in this study. The evidence synthesis draws from publicly available policy documents, standards publications, peer-reviewed literature, and documented incident records. A selected subset of sources is cited directly in the manuscript to support specific claims, while the complete evidence set is provided in the Supplementary Materials. All 60 references — comprising the 54 PRISMA-screened sources plus six revision-time additions [21,28,29,41–43] added during revision — are listed in the References section per MDPI citation requirements.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AP	Analytical Proposition
ATO	Authorization to Operate
AVX2	Advanced Vector Extensions 2
CNSA	Commercial National Security Algorithm Suite
CSF	Cybersecurity Framework
CVE	Common Vulnerabilities and Exposures
ECDSA	Elliptic Curve Digital Signature Algorithm
FIPS	Federal Information Processing Standards
HNDL	Harvest-Now, Decrypt-Later
HNFL	Harvest-Now, Forge-Later
HRoT	Hardware Root of Trust
HSM	Hardware Security Module
IETF	Internet Engineering Task Force
ISO/IEC	International Organization for Standardization / International Electrotechnical Commission
MBOM	Model Bill of Materials
ML-DSA	Module-Lattice-Based Digital Signature Algorithm
ML-KEM	Module-Lattice-Based Key-Encapsulation Mechanism
NIST	National Institute of Standards and Technology
NSA	National Security Agency
NSS	National Security System
OMB	Office of Management and Budget
PQC	Post-Quantum Cryptography

 doi:10.20944/preprints202603.1963.v2

PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
SBOM	Software Bill of Materials
SCAMM	Supply Chain Assurance Maturity Model
SLH-DSA	Stateless Hash-Based Digital Signature Algorithm
SSDF	Secure Software Development Framework
TPM	Trusted Platform Module
ZTA	Zero Trust Architecture

References

1. NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0); National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023.
2. NIST. ML-DSA: Module-Lattice-Based Digital Signature Algorithm; FIPS 204; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024.
3. NIST. SLH-DSA: Stateless Hash-Based Digital Signature Algorithm; FIPS 205; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024.
4. ReversingLabs. Malicious Machine Learning Packages Targeting ML Developers in PyPI; ReversingLabs Threat Research: Boston, MA, USA, 2023.
5. CISA. Software Supply Chain Attacks: Threat Landscape and Mitigations; Cybersecurity and Infrastructure Security Agency: Washington, DC, USA, 2023.
6. MITRE. ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems; The MITRE Corporation: McLean, VA, USA, 2024. Available online: https://atlas.mitre.org (accessed on 30 April 2026).
7. OWASP. Machine Learning Security Top 10: ML06 — AI Supply Chain Attacks; OWASP Foundation, 2023. Available online: https://owasp.org/www-project-machine-learning-security-top-10/ (accessed on 30 April 2026).
8. OWASP. Top 10 for Large Language Model Applications, Version 2025: LLM03 — Supply Chain; OWASP GenAI Security Project, 2025. Available online: https://genai.owasp.org/ (accessed on 30 April 2026).
9. ISO/IEC 42001:2023; Information Technology — Artificial Intelligence — Management System; International Organization for Standardization: Geneva, Switzerland, 2023.
10. NIST. Secure Software Development Framework (SSDF), Version 1.1; SP 800-218; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2022.
11. NSA. Commercial National Security Algorithm Suite 2.0 (CNSA 2.0); National Security Agency: Fort Meade, MD, USA, 2022.
12. OMB. Memorandum M-23-02: Migrating to Post-Quantum Cryptography; Office of Management and Budget: Washington, DC, USA, 2022. Implements National Security Memorandum 10 (NSM-10).
13. Carlini, N.; Jagielski, M.; Choquette-Choo, C.A.; Paleka, D.; Pearce, W.; Anderson, H.; Terzis, A.; Thomas, K.; Tramèr, F. Poisoning Web-Scale Training Datasets is Practical. In Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP); IEEE: San Francisco, CA, USA, 2024; pp. 407–425. https://doi.org/10.1109/SP54263.2024.00179
14. Goldblum, M.; Tsipras, D.; Xie, C.; Chen, X.; Schwarzschild, A.; Song, D.; Mądry, A.; Li, B.; Goldstein, T. Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses. IEEE Trans. Pattern Anal. Mach. Intell. 2023, 45, 1563–1580. https://doi.org/10.1109/TPAMI.2022.3162397
15. Jiang, W.; Synovic, N.; Hyatt, M.; Schorlemmer, T.R.; Sethi, R.; Lu, Y.-H.; Thiruvathukal, G.K.; Davis, J.C. An Empirical Study of Pre-Trained Model Reuse in the Hugging Face Deep Learning Model Registry. In Proceedings of the 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE); IEEE: Melbourne, Australia, 2023; pp. 2463–2475. https://doi.org/10.1109/ICSE48619.2023.00206
16. PyTorch. TorchServe Security Advisory: Server-Side Request Forgery and Model Loading Vulnerabilities (CVE-2023-43654); PyTorch Foundation: San Francisco, CA, USA, 2023. Available online: https://github.com/pytorch/serve/security/advisories/GHSA-8fxr-qfr9-p34w (accessed on 30 April 2026).
17. Ladisa, P.; Plate, H.; Martinez, M.; Barais, O. SoK: Taxonomy of Attacks on Open-Source Software Supply Chains. In Proceedings of the 2023 IEEE Symposium on Security and Privacy (SP); IEEE: San Francisco, CA, USA, 2023; pp. 1509–1526. https://doi.org/10.1109/SP46215.2023.10179304

 doi:10.20944/preprints202603.1963.v2

18. Li, Y.; Wang, H.; Barni, M. A Survey of Deep Neural Network Watermarking Techniques. Neurocomputing 2021, 461, 171–193. https://doi.org/10.1016/j.neucom.2021.07.051
19. Zhao, J.; Wang, S.; Zhao, Y.; Hou, X.; Wang, K.; Gao, P.; Zhang, Y.; Wei, C.; Wang, H. Models Are Codes: Towards Measuring Malicious Code Poisoning Attacks on Pre-trained Model Hubs. In Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering (ASE 2024); ACM: Sacramento, CA, USA, 2024. https://doi.org/10.1145/3691620.3695271
20. Rieger, P.; Krauß, T.; Miettinen, M.; Dmitrienko, A.; Sadeghi, A.-R. CrowdGuard: Federated Backdoor Detection in Federated Learning. In Proceedings of the Network and Distributed System Security Symposium (NDSS 2024); Internet Society: San Diego, CA, USA, 2024. https://doi.org/10.14722/ndss.2024.23233
21. Bai, S.; Ducas, L.; Kiltz, E.; Lepoint, T.; Lyubashevsky, V.; Schwabe, P.; Seiler, G.; Stehlé, D. CRYSTALS-Dilithium Algorithm Specifications and Supporting Documentation (Version 3.1); NIST Post-Quantum Cryptography Standardization Round 3 Submission, 8 February 2021. Available online: https://pq-crystals.org/dilithium/data/dilithium-specification-round3-20210208.pdf (accessed on 26 April 2026).
22. NIST. ML-KEM: Module-Lattice-Based Key-Encapsulation Mechanism; FIPS 203; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024.
23. NIST. Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations; SP 800-161r1; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2022 (updated 2024). https://doi.org/10.6028/NIST.SP.800-161r1-upd1
24. NIST. The NIST Cybersecurity Framework (CSF) 2.0; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024.
25. NIST. SP 800-204D: Strategies for the Integration of Software Supply Chain Security in DevSecOps CI/CD Pipelines; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024.
26. DoD CDAO. Responsible Artificial Intelligence (RAI) Toolkit; Chief Digital and Artificial Intelligence Office: Arlington, VA, USA, 2023. Available online: https://www.ai.mil/Latest/Blog/Article-Display/Article/3940314/responsible-ai-toolkit/ (accessed on 30 April 2026).
27. Cooper, D.; Apon, D.; Dang, Q.; Davidson, M.; Dworkin, M.; Miller, C. Recommendation for Stateful Hash-Based Signature Schemes; NIST Special Publication (SP) 800-208; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2020. https://doi.org/10.6028/NIST.SP.800-208
28. Hevner, A.R.; March, S.T.; Park, J.; Ram, S. Design Science in Information Systems Research. MIS Quarterly 2004, 28, 75–105.
29. Peffers, K.; Tuunanen, T.; Rothenberger, M.A.; Chatterjee, S. A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems 2007, 24, 45–77.
30. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021, 372, n71. https://doi.org/10.1136/bmj.n71
31. IETF. Hybrid Key Exchange in TLS 1.3; Internet-Draft draft-ietf-tls-hybrid-design-16; Internet Engineering Task Force, 2026. (Work in Progress.) Available online: https://datatracker.ietf.org/doc/draft-ietf-tls-hybrid-design/ (accessed on 30 April 2026).
32. IETF. Post-quantum Hybrid ECDHE-MLKEM Key Agreement for TLSv1.3; Internet-Draft draft-ietf-tls-ecdhe-mlkem-04; Internet Engineering Task Force, 2026. (Work in Progress.) Available online: https://datatracker.ietf.org/doc/draft-ietf-tls-ecdhe-mlkem/ (accessed on 30 April 2026).
33. NIST. Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile; SP 800-218A; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024. Available online: https://csrc.nist.gov/pubs/sp/800/218/a/final (accessed on 30 April 2026).
34. NIST. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile; AI 600-1; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024.
35. NIST. Security and Privacy Controls for Information Systems and Organizations; SP 800-53, Rev. 5; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2020. https://doi.org/10.6028/NIST.SP.800-53r5
36. ISO/IEC 23894:2023; Information Technology — Artificial Intelligence — Guidance on Risk Management; International Organization for Standardization: Geneva, Switzerland, 2023.

 doi:10.20944/preprints202603.1963.v2

37. CISA. Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Security-by-Design and -Default; Cybersecurity and Infrastructure Security Agency: Washington, DC, USA, April 2023. Available online: https://www.cisa.gov/securebydesign (accessed on 30 April 2026).
38. DoD. Data, Analytics, and Artificial Intelligence Adoption Strategy; Department of Defense: Washington, DC, USA, 2023.
39. OWASP Foundation; Ecma TC54. CycloneDX Bill of Materials Standard (ECMA-424); 2023. Available online: https://cyclonedx.org/specification/overview/ (accessed on 30 April 2026).
40. Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; Ma, X. Anti-Backdoor Learning: Training Clean Models on Poisoned Data. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021); Curran Associates: Red Hook, NY, USA, 2021; pp. 14900–14912.
41. Trusted Computing Group. PC Client Platform TPM Profile (PTP) Specification, Family 2.0; Level 00, 2025 draft revision; Trusted Computing Group: Beaverton, OR, USA, 2025. Available online: https://trustedcomputinggroup.org/resource/pc-client-platform-tpm-profile-ptp-specification/ (accessed on 26 April 2026).
42. GSA. Post-Quantum Cryptography Buyer's Guide; U.S. General Services Administration: Washington, DC, USA, 2025. Available online: https://buy.gsa.gov/api/system/files/documents/final-508c-pqc_buyer-s_guide_2025.pdf (accessed on 26 April 2026).
43. CISA. Product Categories for Technologies That Use Post-Quantum Cryptography Standards; Cybersecurity and Infrastructure Security Agency: Washington, DC, USA, 23 January 2026; published per Executive Order 14306. Available online: https://www.cisa.gov/resources-tools/resources/product-categories-technologies-use-post-quantum-cryptography-standards (accessed on 26 April 2026).
44. Machado, G.R.; Silva, E.; Goldschmidt, R.R. Adversarial Machine Learning in Image Classification: A Survey Toward the Defender's Perspective. ACM Comput. Surv. 2023, 55, 1–38. https://doi.org/10.1145/3485133
45. Gu, T.; Dolan-Gavitt, B.; Garg, S. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. IEEE Access 2019, 7, 47230–47244. https://doi.org/10.1109/ACCESS.2019.2909068
46. Wu, B.; Chen, H.; Zhang, M.; Zhu, Z.; Wei, S.; Yuan, D.; Shen, C. BackdoorBench: A Comprehensive Benchmark of Backdoor Learning. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Datasets and Benchmarks Track; Curran Associates: Red Hook, NY, USA, 2022.
47. Ohm, M.; Plate, H.; Sykosch, A.; Meier, M. Backstabber's Knife Collection: A Review of Open Source Software Supply Chain Attacks. In Proceedings of the Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA 2020); Maurice, C., Bilge, L., Stringhini, G., Neves, N., Eds.; Lecture Notes in Computer Science 12223; Springer: Cham, Switzerland, 2020; pp. 23–43. https://doi.org/10.1007/978-3-030-52683-2_2
48. PyTorch Foundation. Compromised PyTorch-nightly Dependency Chain Between December 25th and December 30th, 2022; PyTorch Blog, December 2022. Available online: https://pytorch.org/blog/compromised-nightly-dependency/ (accessed on 30 April 2026).
49. Ultralytics. GitHub Issue #18027: Published Wheel 8.3.41 Contained Code Not Present in GitHub and Appeared to Invoke an XMRig Miner; December 2024. Available online: https://github.com/ultralytics/ultralytics/issues/18027 (accessed on 30 April 2026).
50. Open Container Initiative. Image Format Specification, Version 1.1.0; OCI Working Group, 2024. Available online: https://github.com/opencontainers/image-spec (accessed on 30 April 2026).
51. NIST. Zero Trust Architecture; SP 800-207; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2020.
52. Hugging Face. Hub Security Documentation: Pickle Scanning, Malware Scanning, and Repository Trust Controls; 2024. Available online: https://huggingface.co/docs/hub/en/security (accessed on 30 April 2026).
53. Red Hat. Strengthen Security in Your Software Supply Chain; Red Hat: Raleigh, NC, USA, 2024. Available online: https://www.redhat.com/en/solutions/trusted-software-supply-chain (accessed on 30 April 2026).
54. NIST. A Zero Trust Architecture Model for Access Control in Cloud-Native Applications in Multi-Cloud Environments; SP 800-207A; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023.

 doi:10.20944/preprints202603.1963.v2

55. NSA; CISA. Deploying AI Systems Securely: Best Practices for Deploying Secure and Resilient AI Systems; National Security Agency and Cybersecurity and Infrastructure Security Agency, 2024. Available online: https://media.defense.gov/2024/Apr/15/2003439257/-1/-1/0/CSI-DEPLOYING-AI-SYSTEMS-SECURELY.PDF (accessed on 30 April 2026).
56. IETF. Composite ML-DSA for Use in X.509 Public Key Infrastructure; Internet-Draft, current IETF LAMPS working-group draft. (Work in Progress.) Available online: https://datatracker.ietf.org/doc/draft-ietf-lamps-pq-composite-sigs/ (accessed on 30 April 2026).
57. Massimo, J.; Kampanakis, P.; Turner, S.; Westerbaan, B.E. Internet X.509 Public Key Infrastructure — Algorithm Identifiers for the Module-Lattice-Based Digital Signature Algorithm (ML-DSA); RFC 9881; Internet Engineering Task Force, October 2025. https://doi.org/10.17487/RFC9881. Available online: https://www.rfc-editor.org/rfc/rfc9881 (accessed on 30 April 2026).
58. The White House. Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence; Washington, DC, USA, 30 October 2023.
59. Google. Secure AI Framework (SAIF); Google LLC: Mountain View, CA, USA, 2023. Available online: https://safety.google/cybersecurity-advancements/saif/ (accessed on 30 April 2026).
60. Microsoft. AI Security Research Program; Microsoft Corporation: Redmond, WA, USA, 2024. Available online: https://www.microsoft.com/en-us/security/business/solutions/security-for-ai (accessed on 30 April 2026).