Preprints.org

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Not peer-reviewed version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: 25 March 2026

doi: 10.20944/preprints202603.1963.v1

Keywords: model integrity; post-quantum cryptography; model provenance; MBOM-PQC; ML-DSA; model signing; attestation pipeline; SCAMM; zero trust integration; cryptographic agility

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

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; rc@medcybersecurity.com

Abstract

Artificial intelligence systems increasingly depend on complex, multi-stage supply chains that incorporate pre-trained models, third-party datasets, open-source libraries, and automated training pipelines. This dependency creates a rapidly expanding attack surface in which model poisoning, dependency compromise, and provenance manipulation can undermine system integrity long before deployment. Existing AI governance frameworks—including the NIST AI Risk Management Framework and NIST's Secure Software Development Framework—acknowledge supply chain risks but do not define a verifiable model provenance structure or cryptographically durable integrity guarantees. Simultaneously, the transition to post-quantum cryptography (PQC) introduces new requirements for long-lived AI artifacts: classical digital signatures used to verify model lineage, dataset integrity, and pipeline attestation will become vulnerable to quantum-enabled forgery within the expected operational lifetime of many AI systems. This paper synthesizes evidence from policy, standards, and benchmark sources to characterize the emerging AI supply chain threat landscape and identify cryptographic dependencies that the PQC transition disrupts. We propose a formal Model Bill of Materials with PQC-safe extensions (MBOM-PQC), a unified model signing and attestation pipeline based on ML-DSA and hybrid signature modes, and a five-level Supply Chain Assurance Maturity Model (SCAMM) enabling repeatable organizational assessment. The framework provides a cryptographically resilient foundation for AI provenance, ensuring that model integrity, lineage, and trustworthiness remain verifiable throughout the PQC transition and beyond. The principal contributions are: (1) the MBOM-PQC schema for structured AI provenance; (2) a PQC-safe signing and attestation pipeline; and (3) the SCAMM five-level maturity model for organizational assessment.

Keywords: model integrity; post-quantum cryptography; model provenance; MBOM-PQC; ML-DSA; model signing; attestation pipeline; SCAMM; zero trust integration; cryptographic agility

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. 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 [1–3]. The MITRE ATLAS knowledge base further systematizes these attack patterns by cataloging adversarial tactics and techniques relevant to AI-enabled systems [4].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

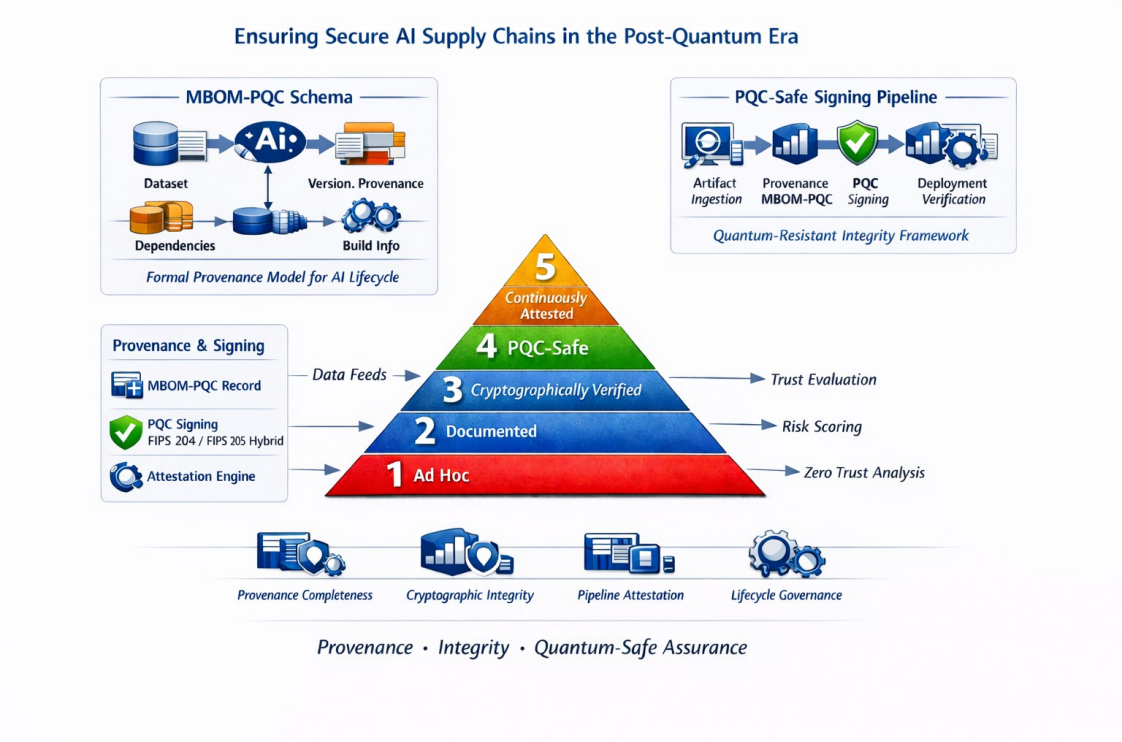

Figure 1. Graphical Abstract. The integrated MBOM-PQC Schema (top left), PQC-Safe Signing Pipeline (top right), and SCAMM Maturity Pyramid (center) provide a unified framework for AI supply chain assurance grounded in 54-source evidence synthesis. Assessment dimensions—Provenance Completeness, Cryptographic Integrity, Pipeline Attestation, and Lifecycle Governance—underpin all five maturity levels.

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 [5] 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 [6] 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.

This gap becomes more acute as the global cryptographic ecosystem transitions to post-quantum cryptography (PQC). AI models, datasets, and provenance records often have operational lifetimes measured in years, making them vulnerable to what may be termed "harvest-now, forge-later" (HNFL) attacks: a class of threat in which adversaries collect signed AI artifacts under current classical signature regimes and subsequently forge counterfeit signatures once cryptographically relevant quantum computers become available. HNFL is analogous to the well-known store-now, decrypt-later (SNDL) pattern for encrypted data, but targets signature integrity rather than confidentiality. Classical digital signatures used to verify model lineage, dataset integrity, and pipeline attestation will not provide long-term protection against quantum-enabled forgery. The National Security Agency's Commercial National Security Algorithm Suite 2.0 (CNSA 2.0) establishes a quantum-resistant signature direction for National Security Systems that aligns with ML-DSA as standardized in FIPS 204 [7,8], while NIST's broader PQC standards additionally include SLH-DSA (FIPS 205) [9] for non-NSS federal and commercial use. Yet no current AI assurance framework incorporates PQC-safe signing or hybrid signature modes into model provenance or supply chain validation.

The absence of a standardized, cryptographically resilient provenance model creates operational challenges for organizations deploying AI systems in high-assurance environments. Without verifiable

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

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 quantum-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 provide a durable foundation for AI supply chain integrity, ensuring that model lineage, authenticity, and trustworthiness remain verifiable throughout the PQC transition and beyond.

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 [10]. 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. 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 during distribution or deployment, in which a legitimate model is substituted with a manipulated version, undermine trust in the entire pipeline. These risks are amplified by the opacity of AI artifacts: unlike source code, model weights and training datasets are difficult to inspect, making tampering hard to detect without cryptographic verification or provenance metadata. The MITRE ATLAS framework catalogs these and related adversarial techniques across the AI lifecycle [4]. As a result, 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. Digital signing mechanisms can be used to verify the authenticity and integrity of model artifacts, providing a cryptographic basis for detecting tampering and confirming provenance [11]. Dataset integrity verification ensures training data has not been altered or poisoned. Secure enclaves and confidential computing protect model

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

execution environments. Federated learning authentication ensures that model updates originate from trusted participants. 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. However, most current implementations use classical cryptographic algorithms—such as RSA, ECDSA, and 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. NIST's standardization of ML-KEM (FIPS 203) [12] for key establishment and ML-DSA (FIPS 204) [7] and SLH-DSA (FIPS 205) [9] for digital signatures fundamentally changes the cryptographic landscape. PQC algorithms have significantly larger key and signature sizes than their classical counterparts, affecting model signing workflows, certificate chains, secure enclaves, federated learning protocols, and long-lived artifacts such as model checkpoints and provenance records. The NSA's CNSA 2.0 [8] establishes a quantum-resistant signature direction for National Security Systems that aligns with ML-DSA and emphasizes that classical signatures will not provide long-term integrity protection. SLH-DSA (FIPS 205), 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 an urgent need for PQC-safe signing and hybrid signature modes in AI supply chains, ensuring that integrity guarantees established today remain valid in a post-quantum environment.

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 AI Risk Management Framework (AI RMF) [5] provides governance and risk categories but does not define a model provenance structure or cryptographic requirements. NIST's Secure Software Development Framework (SSDF) [6] addresses software supply chain risks but does not extend to AI-specific artifacts such as model weights or training datasets. NIST SP 800-204D [13] 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 [14] focuses on ethical and operational considerations rather than cryptographic integrity. Established 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 [15] and CNSA 2.0 [8], does not address AI artifacts or model provenance. The absence of a Model Bill of Materials (MBOM) standard, combined with the lack of PQC-safe signing and attestation pipelines, leaves organizations without a structured method for verifying the authenticity, lineage, and integrity of AI systems. This gap motivates the need for a formal, cryptographically anchored framework that integrates provenance, PQC-safe signatures, and supply chain assurance.

3. Materials and Methods

3.1. Research Design and Contribution Type

This study employs a dual-method research design combining systematic evidence synthesis with prescriptive architectural development. The first component synthesizes authoritative policy documents, standards publications, and documented AI supply chain incidents to characterize the current state of AI supply chain risk and identify cryptographic dependencies relevant to post-quantum transition. The second component derives a formal provenance schema, PQC-safe signing pipeline, and

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

maturity model from the synthesized requirements. This approach is appropriate because the research questions concern (1) what is documented about AI supply chain vulnerabilities and cryptographic transition requirements, and (2) what architectural structures are needed to ensure durable provenance and integrity. The prescriptive components are derived artifacts intended for future operational validation rather than empirical evaluation. The methodology follows PRISMA 2020 guidelines [16] adapted for policy, standards, and security-incident synthesis. This adaptation is necessary because the evidence base consists primarily of normative requirements, technical specifications, and documented attack patterns rather than experimental studies. Accordingly, this study is best characterized as a design-oriented evidence synthesis: a prescriptive systems-architecture study informed by systematic evidence screening, rather than a conventional effectiveness or intervention review. The PRISMA-adapted process provides methodological rigor and transparency for the evidence-gathering phase, while the primary contributions—the MBOM-PQC schema, signing pipeline, and SCAMM maturity model—are design-science artifacts derived from that evidence base and intended for future empirical validation. Traceability between included sources, analytical propositions, extracted requirements, and architectural components is documented through the supplementary evidence bibliography, extraction matrix, exclusion ledger, and confidence-tier summary. All extraction decisions, coding rules, and architectural traceability mappings are documented in the Supplementary Materials.

3.2. Analytical Propositions

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 and December 2025 (IETF Internet-Drafts [17,18] were updated to their current versions during manuscript revision). Standards bodies provided NIST CSRC publications, NSA CNSA 2.0 guidance, IETF PQC drafts, and ISO/IEC AI standards. Government policy sources included the NIST Secure Software Development Framework (SSDF), NIST AI RMF, DoD CDAO Responsible AI Toolkit, and federal PQC transition memoranda. AI supply chain incident records documented cases of model poisoning, dependency compromise, malicious ML libraries, and tampered pre-trained models. Academic literature contributed peer-reviewed papers on adversarial ML, supply chain attacks, provenance modeling, and PQC performance benchmarks. Industry reports provided security analyses from reputable organizations documenting real-world AI supply chain failures. Primary search strings included: “AI supply chain” AND “provenance”; “model signing” AND “integrity”; “post-quantum” AND “AI assurance”; “ML-DSA” OR “SLH-DSA” AND “signature size”; and “model poisoning” OR “model swap attack.” Searches were executed across IEEE Xplore, ACM Digital Library, IACR ePrint, NIST CSRC, NSA.gov, CISA.gov, and curated incident repositories. To support architectural synthesis and policy alignment, the search and source-review process also incorporated authoritative government, standards, and industry materials from organizations such as the White House, NTIA, ISO/IEC, OWASP, CycloneDX, Google, Microsoft, Red Hat, and related incident documentation when those sources met the inclusion criteria.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

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. [17] and [18], 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.

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: 31 policy and standards documents, 15 peer-reviewed academic papers, and 8 documented supply chain incidents. 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 2. After full-text review, 54 sources were included in the evidence synthesis. A selected subset (28 of 54) is cited directly in the manuscript body to support specific argumentative claims, while the remainder informed the synthesis, analytical propositions, and architectural derivation. All 54 sources are listed in the References section to comply with MDPI citation requirements for supplementary materials. The complete 54-source evidence bibliography with tier assignments and extraction metadata is provided in the Supplementary Materials.

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 (A, B, or C); 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

Sources were classified into three confidence tiers. Tier A (High) encompasses authoritative standards from NIST and NSA, mature IETF working-group drafts relevant to PQC transition, peer-reviewed benchmarks, and documented incidents with forensic evidence; Tier A sources support quantitative or normative claims. Tier B (Medium) encompasses authoritative guidance from CISA and DoD CDAO and industry analyses with verifiable methodology; Tier B sources support architectural requirements. Tier C (Low) encompasses contextual reports, preliminary findings, and sources that identify gaps without prescriptive detail; Tier C sources are used to contextualize emerging risks. Tier assignments are recorded in the extraction template for each source and are referenced in the traceability matrix to indicate the evidential strength behind each architectural decision.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

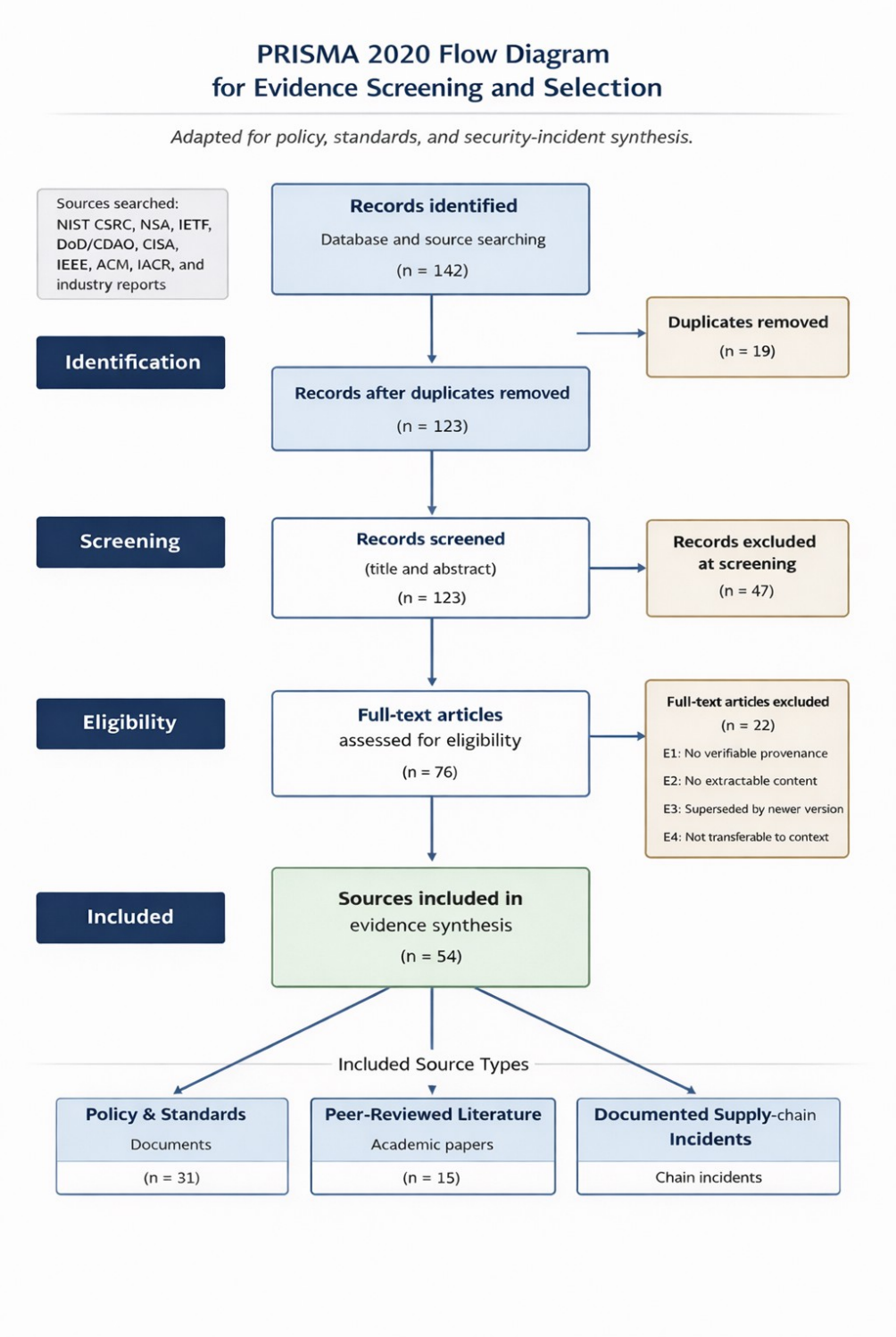

Figure 2. PRISMA 2020 Flow Diagram—Evidence Screening and Selection. Adapted for policy, standards, and security-incident synthesis [16]. The initial search identified 142 records across five source classes (NIST CSRC, NSA, IETF, DoD CDAO, CISA, IEEE Xplore, ACM Digital Library, IACR ePrint, and industry reports). 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: 31 policy and standards documents, 15 peer-reviewed academic papers, and 8 documented supply chain incidents.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

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 provided in the Supplementary Materials and is summarized in Section 5.4.

The results of the evidence synthesis are presented in four stages. Section 4 characterizes the AI supply chain threat landscape and derives four requirement classes from documented threats and cryptographic dependencies. Section 5 translates those requirements into the MBOM-PQC schema, a structured provenance model for AI artifacts. Section 6 operationalizes the schema through a PQC-safe signing and attestation pipeline. Section 7 introduces the Supply Chain Assurance Maturity Model (SCAMM), which enables repeatable organizational assessment against the requirements, schema, and pipeline defined in the preceding sections.

4. Results: Threat and Dependency Analysis

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 (Figures 3 and 4).

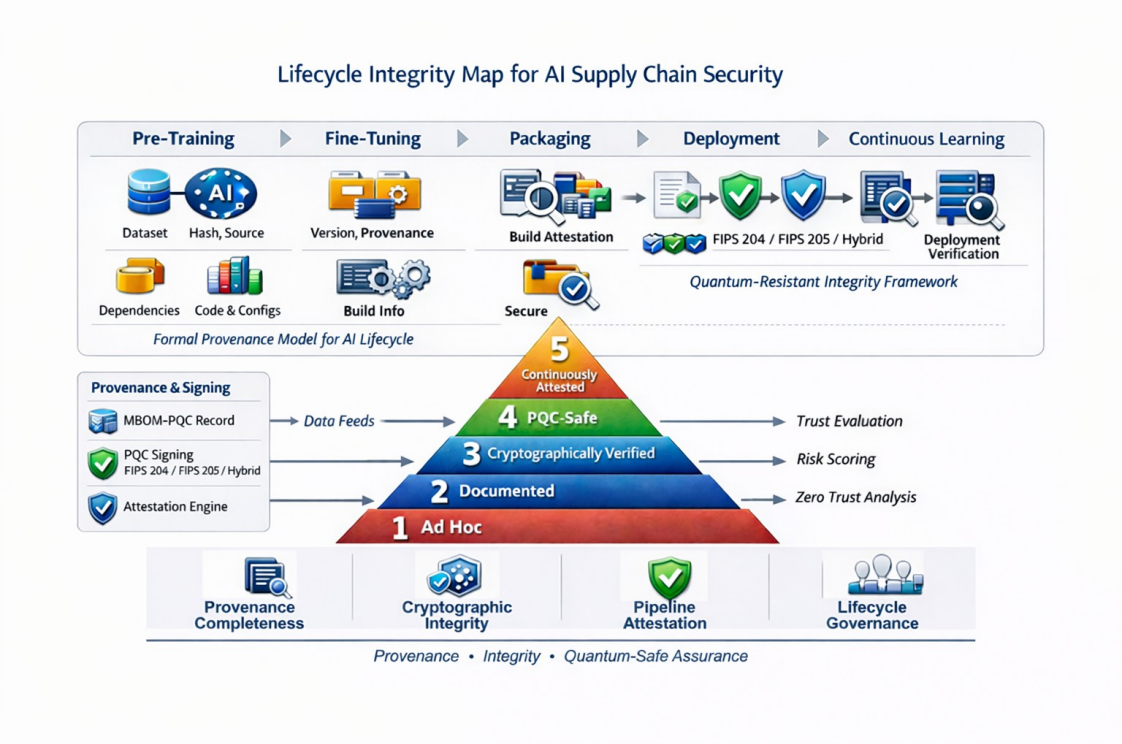

Figure 3. Lifecycle Integrity Map for AI Supply Chain Security. The five-stage AI lifecycle (Pre-Training → Fine-Tuning → Packaging → Deployment → Continuous Learning) with MBOM-PQC provenance capture at each stage. FIPS 204 (ML-DSA), FIPS 205 (SLH-DSA), and hybrid signatures are applied through the PQC-Safe Signing Pipeline. The SCAMM pyramid and four assessment dimensions—Provenance Completeness, Cryptographic Integrity, Pipeline Attestation, and Lifecycle Governance—govern organizational readiness.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

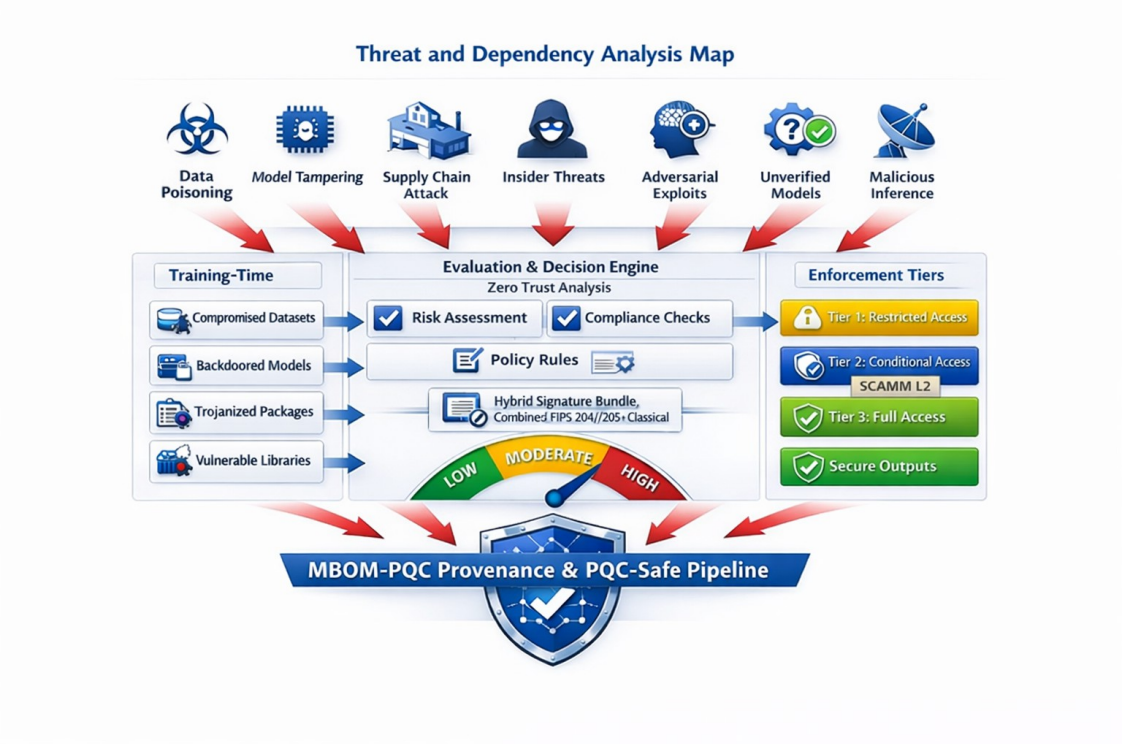

Figure 4. Threat and Dependency Analysis Map. Seven threat vectors—Data Poisoning, Model Tampering, Supply Chain Attack, Insider Threats, Adversarial Exploits, Unverified Models, and Malicious Inference—converge on Training-Time vulnerabilities (Compromised Datasets, Backdoored Models, Trojanized Packages, 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 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 [19,20]. 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, a threat class comprehensively evaluated in controlled benchmarks [21]. 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 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 [1,2,22]. Without verifiable provenance, organizations cannot reliably determine whether ingested models originate from trusted sources, making ingestion-time verification a critical control gap.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

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 [7–9]. 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 [17,18]. 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 [23]. 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

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.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

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 [24]. 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 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 [5,25]. 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

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

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. Results: The MBOM-PQC Schema—A Structured Framework for AI Provenance and Integrity

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.

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 5).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

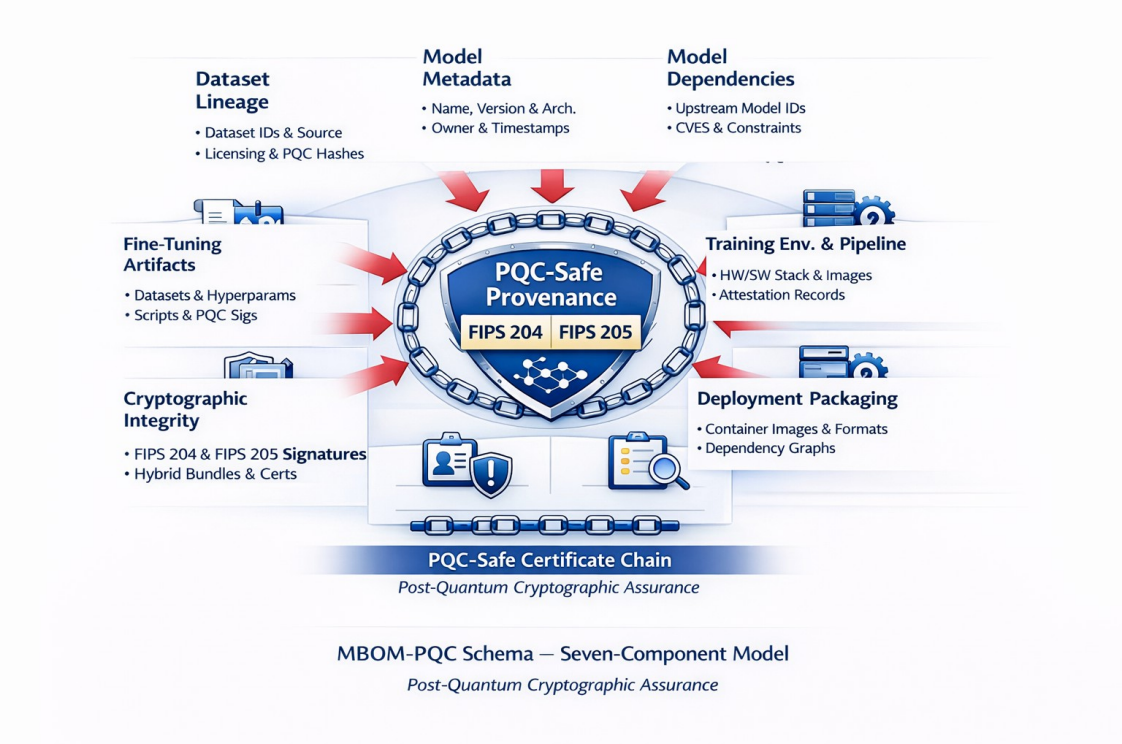

Figure 5. MBOM-PQC Schema—Seven-Component Provenance Model. Seven provenance components—C1: Model Metadata (Name, Version, Architecture), C2: Dataset Lineage (Dataset IDs, Source, Licensing), C3: Model Dependencies (Upstream Model IDs, CVEs), C4: Fine-Tuning Artifacts (Datasets, Hyperparameters, Scripts), C5: Training Environment & Pipeline (HW/SW Stack, Attestation Records), C6: Deployment Packaging (Container Images, Dependency Graphs), and C7: Cryptographic Integrity (FIPS 204 & FIPS 205 Signatures, Hybrid Bundles, PQC-Safe Certificates)—converge into the PQC-Safe Provenance shield signed with FIPS 204 (ML-DSA) and FIPS 205 (SLH-DSA). The PQC-Safe Certificate Chain anchors the entire provenance record, providing Post-Quantum Cryptographic Assurance.

5.2.1. Component 1: Model Metadata

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 HuggingFace or vendor registries, 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

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

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 6).

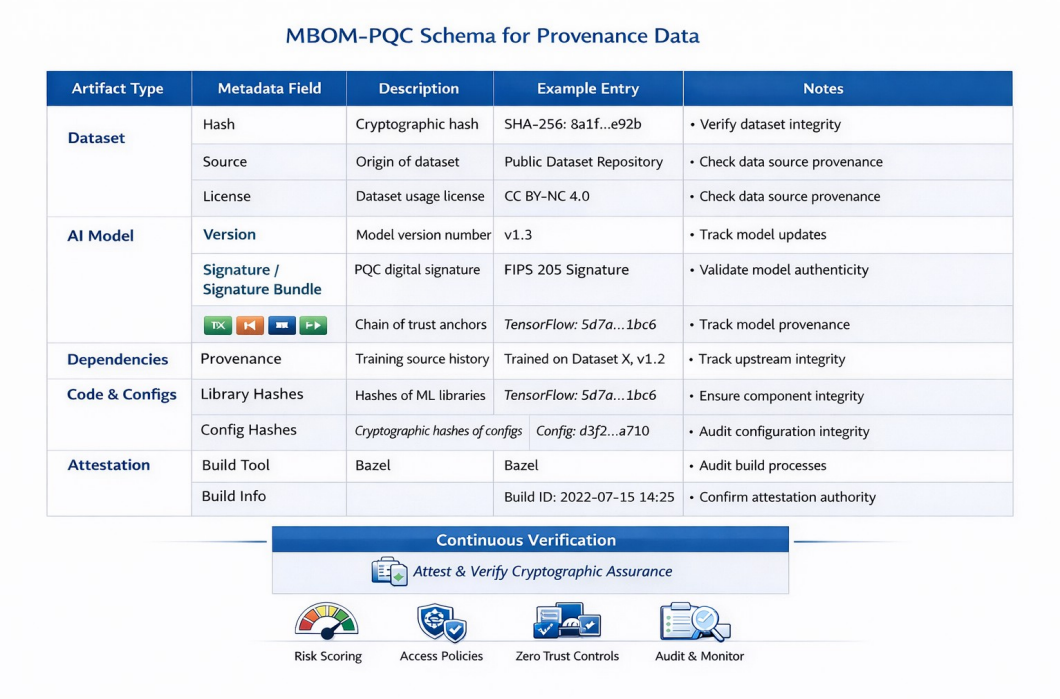

Figure 6. MBOM-PQC Schema for Provenance Data. The schema captures five artifact types—Dataset (Hash, Source, License), AI Model (ML-DSA/PQC Signing, FIPS 204/FIPS 205 Signatures), Dependencies (Provenance/training source history), Code & Configs (Library Hashes, Config Hashes), and Attestation (Build 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-256 hash shown is illustrative; PQC-forward deployments should use SHA-3 or SHAKE for forward compatibility with FIPS 204 and FIPS 205 (see Section 7.2, Level 3).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

Figure 5 presents the full seven-component conceptual provenance model, while Figure 6 provides a schema-oriented abstraction that groups those components into five artifact classes for implementation.

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 [8], 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 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. The traceability matrix below (Table 1) demonstrates that every schema component is grounded in documented supply chain risks and cryptographic dependencies rather than abstract design choices. This mapping provides a structured basis for auditing schema completeness and for extending the schema as new threat classes or PQC standards emerge.

Table 1. Requirements-to-MBOM-PQC Schema Traceability Matrix.

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

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

5.5. Summary

The MBOM-PQC schema provides a comprehensive, cryptographically anchored provenance model for AI supply chains. By integrating PQC-safe signatures, hybrid modes, and lifecycle-wide transparency across seven structured components, it enables organizations to verify the authenticity, integrity, and lineage of AI artifacts throughout the post-quantum transition. The traceability matrix in Table 1 confirms

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

that every schema element is grounded in documented threats and cryptographic requirements. The MBOM-PQC schema forms the foundation for the PQC-safe signing and attestation pipeline introduced in Section 6 and the Supply Chain Assurance Maturity Model presented in Section 7.

6. Results: PQC-Safe Model Signing and Attestation Pipeline

The MBOM-PQC schema defined in Section 5 provides the provenance structure for AI supply chain artifacts. To operationalize that structure, organizations require a corresponding pipeline: a defined sequence of cryptographic operations that generates, applies, and verifies PQC-safe signatures and attestations across the model lifecycle. This section introduces the PQC-Safe Model Signing and Attestation Pipeline—a five-stage operational architecture that integrates FIPS 204 (ML-DSA) signing for operational artifacts, FIPS 205 (SLH-DSA) signing for non-NSS archival records, hybrid signature modes, hardware-rooted attestation, and Zero Trust-aligned trust scoring into a coherent, implementable workflow.

6.1. Pipeline Overview

The pipeline comprises five sequential stages: Ingestion, Verification, Signing, Attestation, and Deployment (Figure 7). 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 8) illustrates how Key Management, Signing Service, Attestation Engine, and Policy Enforcement components support this workflow.

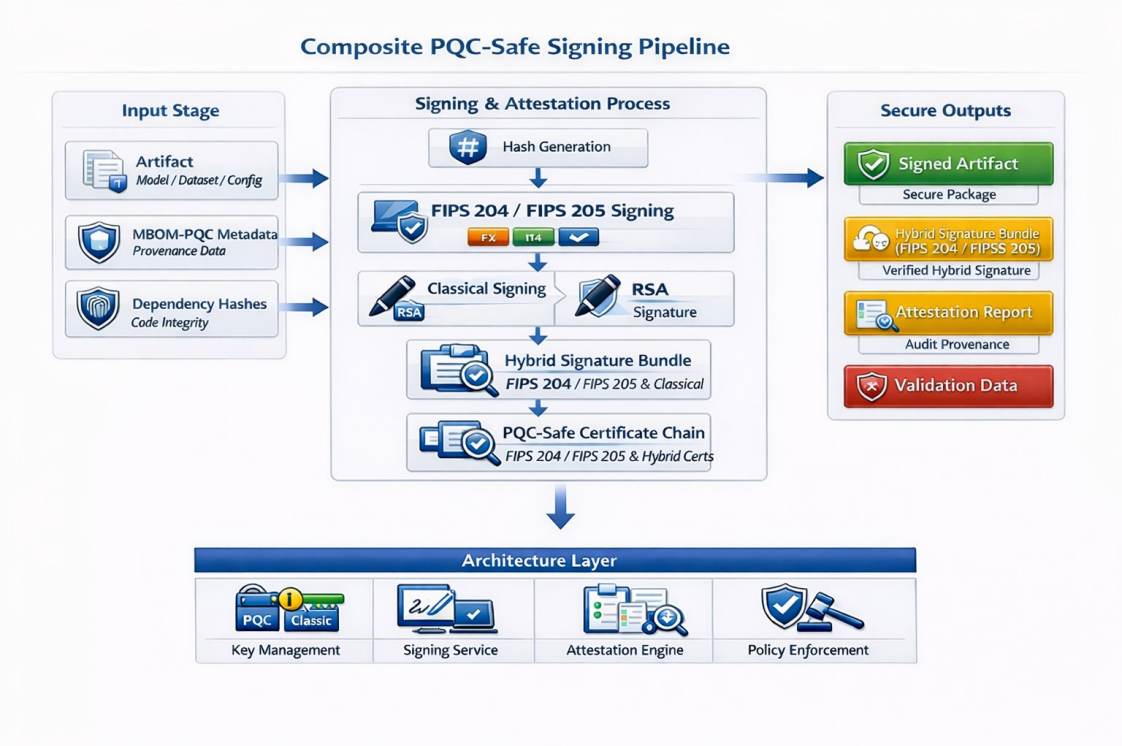

Figure 7. Composite PQC-Safe Signing Pipeline. Input Stage (Artifact, MBOM-PQC Metadata, Dependency Hashes) feeds the Signing & Attestation Process: Hash Generation → PQC Signing (FIPS 204/ML-DSA for operational artifacts; FIPS 205/SLH-DSA for non-NSS archival) and Classical Signing (RSA) in parallel → Hybrid Signature Bundle → Attestation Report → PQC-Safe Certificate Chain. Secure Outputs: Signed Artifact, Hybrid Signature Bundle, Attestation Report, and Validation Data. The three visual columns map to the five pipeline stages defined in Section 6.1: Input Stage encompasses Ingestion (Stage 1) and Verification (Stage 2); Signing & Attestation Process encompasses Signing (Stage 3) and Attestation (Stage 4); Secure Outputs encompasses Deployment (Stage 5).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

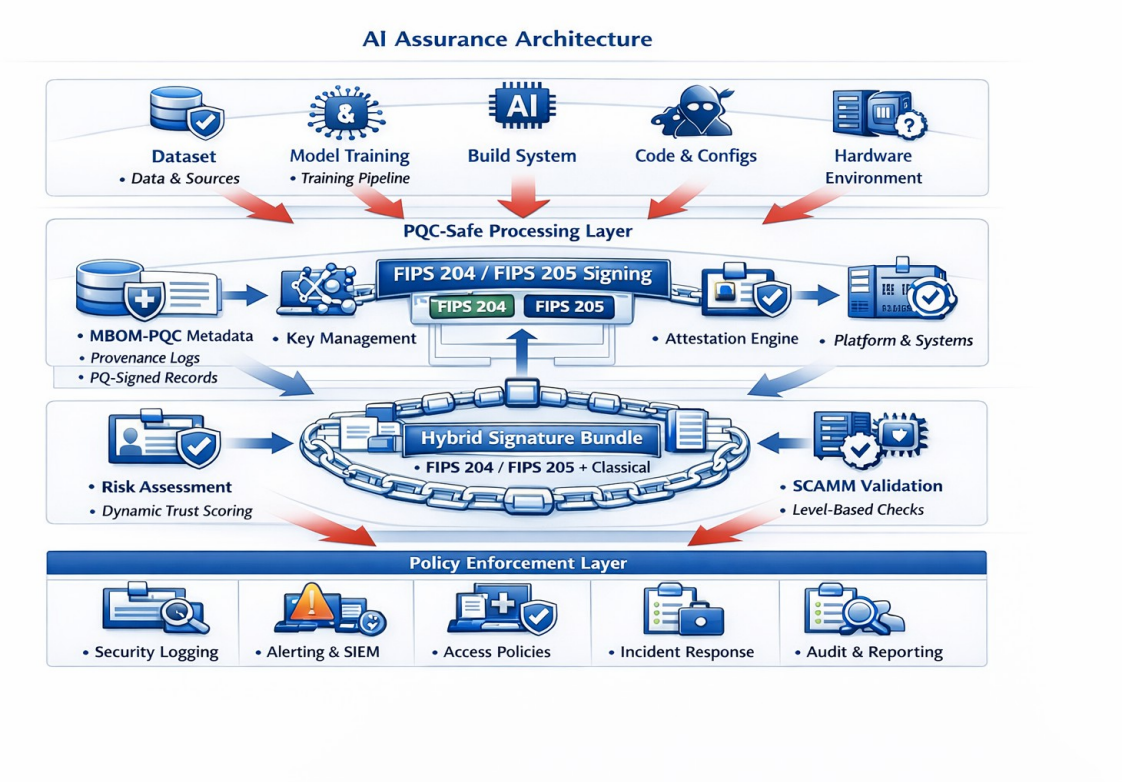

Figure 8. AI Assurance Architecture: Layered View of PQC-Safe Processing, Signing, and Policy Enforcement. The architecture spans four tiers: input artifacts (Dataset, Model Training, Build System, Code & Configs, Hardware Environment), PQC-Safe Processing Layer (PQC Signing Service with FIPS 204 and FIPS 205, Key Management, Attestation Engine, Platform & Systems), Hybrid Signature Bundle (PQC & Classical Signing with Risk Assessment, Dynamic Trust Scoring, and SCAMM Validation), and Policy Enforcement Layer (Security Logging, Alerting & SIEM, Access Policies, Incident Response, Audit & Reporting).

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, both classical and PQC signatures are checked. 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 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).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

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 [8]. Hybrid-signature considerations in this paper are presented as an architectural transition pattern rather than as a finalized IETF normative construct. The cited IETF drafts [17,18] primarily inform hybrid key-establishment and TLS migration mechanics, while signature and attestation design choices are grounded in the NIST FIPS-standardized PQC algorithms and associated assurance requirements (Figure 7). 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. Refs. [17] and [18] are cited as current, transition-relevant IETF work-in-progress on hybrid TLS key-establishment and deployment mechanics, not as primary normative authority for the manuscript's hybrid-signature or attestation architecture. IETF drafts on composite digital signatures and PQC X.509 certificate profiles are progressing separately and should be reviewed against current working-group output at the time of deployment.

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) [7]. 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.

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

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

8–50 KB depending on parameter set). The SLH-DSA-SHA2-128s or SLH-DSA-SHAKE-128s parameter sets are recommended for archival use given their balance of security and signature size.

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 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.

6.3. 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

The attestation architecture is anchored in a hardware root of trust, implemented via Trusted Platform Module (TPM) 2.0 or equivalent secure enclave technology, consistent with platform resiliency guidance [26] (Figure 7). 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; hardware upgrades may be required in legacy environments.

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 and CNSA 2.0 timelines [8].

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 [26]—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.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

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 9).

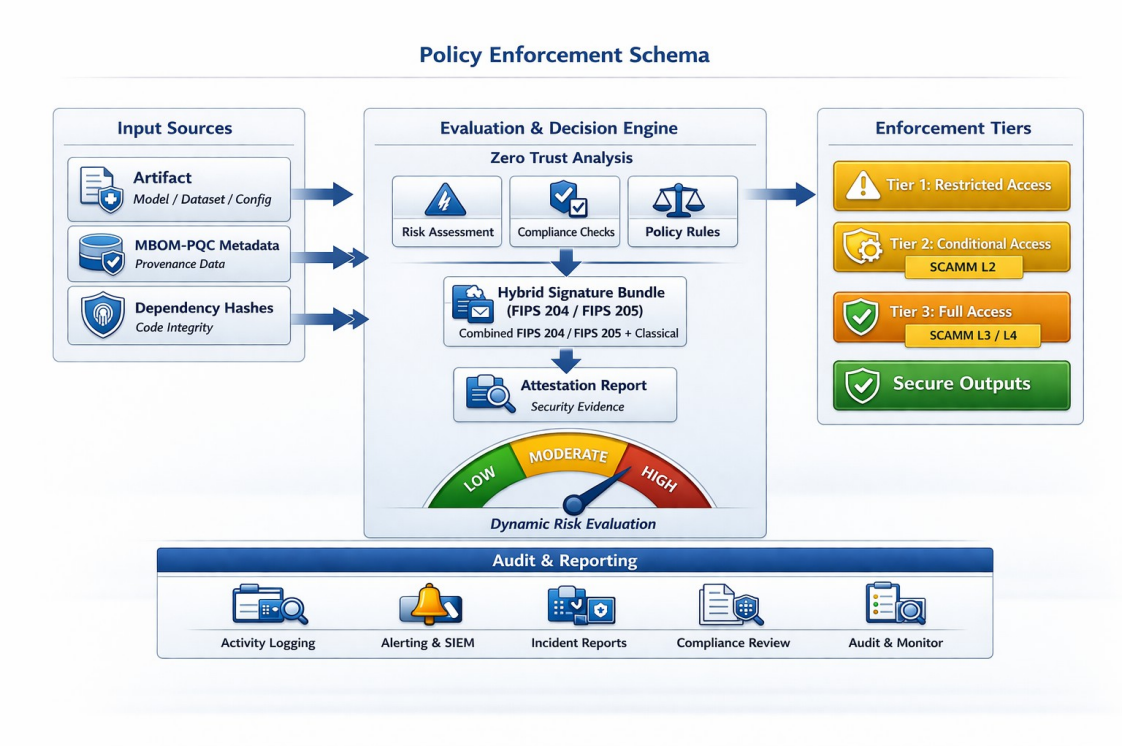

Figure 9. Policy Enforcement Schema. Zero Trust Evaluation & Decision Engine receives Input Sources (Artifact, MBOM-PQC Metadata, Dependency Hashes) and applies Risk Assessment, Compliance Checks, and Policy Rules. Dynamic Risk Evaluation (Low/Moderate/High) drives Enforcement Tiers: Tier 1 Restricted 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.

Whereas Figure 4 maps threat vectors to the decision engine, Figure 9 details the policy enforcement workflow that operationalizes trust decisions within the ZTA framework.

6.4.2. NIST AI RMF Integration

The pipeline also integrates with the NIST AI Risk Management Framework (AI RMF) [5], which organizes AI risk across four core functions: Govern, Map, Measure, and Manage. The MBOM-PQC schema operationalizes the Map function by providing structured artifact provenance that enables organizations to identify and document supply chain dependencies. The PQC-Safe Signing Pipeline supports the Measure function by generating verifiable integrity metrics—signature validity, attestation coverage, and provenance completeness—that can be incorporated into AI RMF measurement plans. The SCAMM maturity model aligns with the Govern and Manage functions by establishing repeatable organizational controls and a roadmap for improving supply chain assurance posture. Together, the pipeline and MBOM-PQC schema provide the technical infrastructure needed to operationalize AI RMF requirements in cryptographically high-assurance deployment environments.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

7. Results: Supply Chain Assurance Maturity Model (SCAMM)

The MBOM-PQC schema and PQC-Safe Signing and Attestation Pipeline introduced in Sections 5 and 6 provide the technical mechanisms for capturing and verifying AI supply chain provenance. However, technical mechanisms alone do not determine whether an organization achieves durable supply chain integrity. Organizational capability—the policies, processes, tooling, and governance structures that govern how provenance is documented, how signatures are applied, and how attestation is maintained across the model lifecycle—determines whether those mechanisms are consistently and correctly applied. Organizations at different stages of AI adoption and cryptographic maturity will implement the framework at different levels of completeness, and they require a structured method for assessing their current posture, identifying gaps, and developing a prioritized improvement roadmap.

This section introduces the Supply Chain Assurance Maturity Model (SCAMM): a five-level framework that enables repeatable, evidence-based assessment of AI supply chain security posture aligned with PQC transition requirements and Zero Trust principles. SCAMM is grounded in the four requirement classes derived in Section 4.4—provenance completeness, cryptographic integrity, pipeline attestation, and lifecycle governance—and is operationalized through measurable indicators that correspond directly to the schema components defined in Section 5 and the pipeline stages defined in Section 6. Each maturity level builds cumulatively on the preceding one, enabling organizations to confirm that foundational controls are in place before advancing to higher-assurance postures. The model is designed to function both as a standalone assessment instrument and as an input to broader governance activities, including Authorization to Operate (ATO) packages, Zero Trust Architecture implementation plans, and CNSA 2.0 compliance roadmaps.

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 10).

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. PQC-safe hashing using SHA-3 or SHAKE is introduced for forward compatibility [7,9] with FIPS 204

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

and FIPS 205. 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.

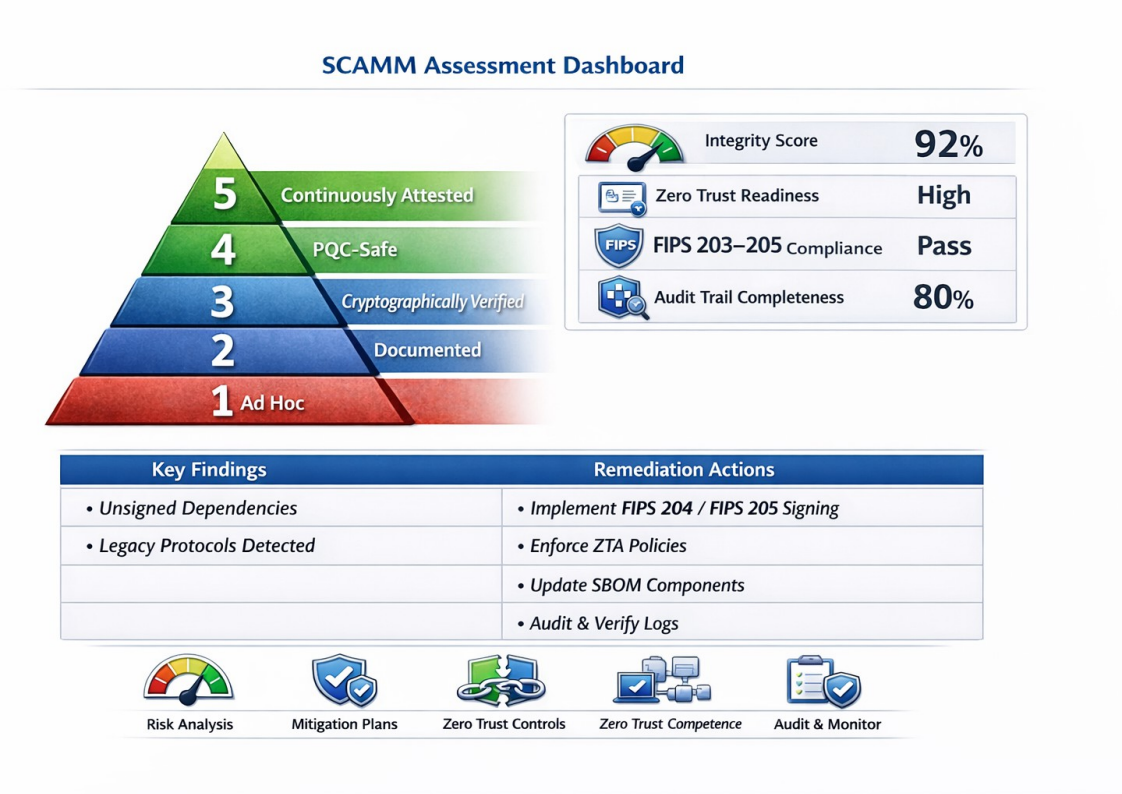

Figure 10. Illustrative SCAMM Assessment Dashboard. A representative organizational assessment output showing Integrity Score (92%), Zero Trust Readiness (High), PQC Compliance (Pass), and Audit Trail Completeness (80%). Key findings—Unsigned Dependencies and Legacy Protocols Detected—map to four remediation actions: Implement PQC-Safe Signing, Enforce ZTA Policies, Update SBOM Components, and Audit & Verify Logs. Governance functions at the base span Risk Analysis, Mitigation Plans, Zero Trust Controls, Zero Trust Competence, and Audit & Monitor. The displayed indicators are representative outputs derived from the SCAMM assessment dimensions and organizational maturity evidence. Note: The dashboard graphic references "FIPS 203–205 Compliance" broadly; within this framework, signing and provenance requirements are scoped to FIPS 204 (ML-DSA) and FIPS 205 (SLH-DSA), as FIPS 203 (ML-KEM) addresses key encapsulation rather than digital signatures.

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. PQC-safe key management integrating FIPS 204/ML-DSA (for all artifacts, and aligned with NSS transition expectations under CNSA 2.0) and FIPS 205/SLH-DSA (for non-NSS archival artifacts under NIST FIPS 205) is built into KMS and HSM systems. 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

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

operational artifacts; for non-NSS long-term archival records, FIPS 205 (SLH-DSA) provides additional hash-based integrity anchoring (note: NSS environments must use FIPS 204 (ML-DSA) in accordance with CNSA 2.0, as FIPS 205 (SLH-DSA) is not approved for NSS use). 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 ATO packages, risk management frameworks, and supply chain security audits.

7.3.1. Provenance Completeness

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.

7.4. Requirements-to-Maturity Mapping

SCAMM directly reflects requirements derived from the threat and dependency analysis in Section 4. Table 2 maps each key requirement to its corresponding maturity levels and provides the derivation rationale grounding each assignment in documented threats and cryptographic dependencies.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

Table 2. Requirements-to-SCAMM Maturity Level Traceability Matrix.

Requirement	SCAMM Level	Rationale
Dataset lineage	Levels 2–5	Required for poisoning detection across all lifecycle stages (Section 4.1.1, Section 4.3.1)
PQC-safe signatures	Levels 4–5	Required for PQC-safe integrity; driven by emerging CNSA 2.0 and federal PQC transition timelines (Section 4.2.1, Section 4.4.2)
Pipeline attestation	Levels 3–5	Required for supply chain transparency and deployment-time tampering detection (Section 4.1.3, Section 4.4.4)
Continuous verification	Level 5	Required for Zero Trust alignment and continuous learning integrity (Section 4.4.4; continuous learning lifecycle context: Section 4.3.4)
Hybrid signature modes	Level 4	Required during PQC transition to maintain backward verifier compatibility (Section 4.2.1, Section 5.3.1, Section 6.2.1)

Section references correspond to the threat and dependency analysis (Section 4) and schema/pipeline sections (Sections 5–6).

7.5. Summary

The Supply Chain Assurance Maturity Model (SCAMM) provides a structured, evidence-based framework for assessing and improving AI supply chain integrity. By defining five cumulative maturity levels aligned with PQC transition requirements and Zero Trust principles, SCAMM enables organizations to evaluate their current posture, prioritize modernization activities, and ensure durable, cryptographically anchored trust in AI systems. The four assessment dimensions—provenance completeness, cryptographic integrity, pipeline attestation, and lifecycle governance—translate the technical requirements of the MBOM-PQC schema and PQC-safe signing pipeline into measurable organizational indicators. The requirements-to-maturity traceability matrix in Table 2 confirms that every SCAMM level is grounded in documented threats and cryptographic dependencies established in earlier sections. Together, Sections 5–7 form a complete, interlocking framework for AI supply chain security that is simultaneously architecturally prescriptive and organizationally actionable.

8. Discussion

The proposed MBOM-PQC schema, PQC-Safe Signing and Attestation Pipeline, and SCAMM maturity model collectively address a critical gap in current AI assurance practices: the absence of a unified, cryptographically durable framework for verifying the provenance, integrity, and trustworthiness of AI supply chains. While existing governance frameworks emphasize transparency, robustness, and ethical considerations, they do not provide the technical mechanisms required to ensure long-term integrity in the face of quantum-enabled threats. This section discusses the broader implications of the proposed framework, the challenges organizations may face during adoption, and the limitations that warrant future research.

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 [5]—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

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

chain manipulation. It also supports mission-critical environments—such as defense, healthcare, and critical infrastructure—where long-term integrity guarantees are essential for Authorization to Operate and ongoing risk management.

8.2. Integration with Zero Trust Architecture and Enterprise Security

The framework aligns naturally with Zero Trust principles. ZTA requires continuous verification of users, devices, and workloads [25]; the proposed pipeline extends this requirement to AI artifacts. Provenance completeness, PQC-safe signatures, and attestation records become inputs to trust scoring, enabling dynamic policy enforcement based on the integrity and lineage of AI components. This integration is particularly important for environments where AI models influence access decisions, anomaly detection, or mission-critical automation. Moreover, 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 [8] 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) [27] and Microsoft's AI security research program [28], 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 consumption during model distribution, and verification time during deployment and inference. ML-DSA-65 signatures are 3,309 bytes (approximately 3.3 KB), compared to 64 bytes for Ed25519, as specified in FIPS 204 [7], Table 2 therein. SLH-DSA signatures range from approximately 8 KB to 50 KB depending on the parameter set [9]. 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.

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.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

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.

8.4. 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) [7], SLH-DSA (FIPS 205) [9], and ML-KEM (FIPS 203) [12] are standardized, additional algorithms may be introduced, and performance optimizations may change implementation guidance. IETF drafts for hybrid key exchange [17,18], PQC certificate profiles, and composite signature formats are still in progress. Because Internet-Drafts are work-in-progress artifacts that may be revised or obsoleted, they are not used here as primary normative authority for hybrid-signature architecture. Instead, they inform the operational discussion of PQC transition pathways and hybrid transport negotiation. 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. Lack of Empirical Validation

The framework is prescriptive and architecture-driven. While grounded in systematic evidence synthesis from 54 sources, it has not yet been validated through large-scale operational deployments. Future work should evaluate performance impacts of PQC signing in real-world AI pipelines, usability of the MBOM-PQC schema across diverse model formats and development toolchains, integration with commercial AI platforms and model registries, and the effectiveness of the pipeline in detecting supply chain compromise in controlled red-team environments. Empirical validation will be essential for establishing the framework as an operational standard rather than a theoretical contribution.

8.4.3. 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. 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.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

8.4.4. Continuous Learning Complexity

Continuous learning systems introduce unique challenges around provenance management: frequent model updates create high-velocity signing requirements, dynamic provenance records must track evolving model state, and dependencies may change between update cycles. While the pipeline nominally supports continuous learning by re-entering updated models at Stage 1, the practical overhead of signing and attesting every incremental update in high-frequency learning systems may require architectural adaptations such as batched provenance recording, delta-signing schemes, or checkpoint-based attestation. Further research is needed to develop efficient provenance management and attestation patterns suited to continuous learning workloads.

8.5. 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.

8.6. Summary

The proposed framework provides a comprehensive, cryptographically anchored approach to AI supply chain security. By integrating PQC-safe signatures, hybrid modes, provenance modeling, and continuous attestation, it addresses critical gaps in current AI assurance practices that existing governance frameworks—including the NIST AI RMF [5], NIST SSDF [6], and DoD CDAO Responsible AI Toolkit [14]—do not fill. Implementation challenges around performance overhead, legacy compatibility, provenance completeness, and organizational maturity are real but tractable; the SCAMM framework provides a structured path for navigating them incrementally. Limitations around empirical validation, evolving standards, upstream transparency, and continuous learning complexity identify a productive research agenda that can be pursued in parallel with practitioner adoption. The framework offers a practical and theoretically grounded path toward durable, Zero Trust-aligned AI supply chain integrity—an essential requirement as organizations across defense, healthcare, critical infrastructure, and enterprise sectors prepare for the post-quantum era.

9. Conclusions

The growing dependence of artificial intelligence systems on multi-stage supply chains—spanning pre-trained models, third-party datasets, open-source libraries, and automated pipelines—has created systemic vulnerabilities that existing governance frameworks are not equipped to address. The transition to post-quantum cryptography (PQC) further complicates this landscape by rendering classical digital signatures insufficient for long-term integrity protection. As AI models and datasets often have operational lifetimes extending well beyond the anticipated arrival of cryptographically relevant quantum computers, organizations require provenance and integrity mechanisms that remain trustworthy in a post-quantum world.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

This paper addressed these challenges by introducing a comprehensive, cryptographically anchored framework for AI supply chain assurance (Figure 3). The MBOM-PQC schema provides a structured, lifecycle-wide provenance model that captures the full set of artifacts influencing model behavior, including datasets, upstream model dependencies, fine-tuning workflows, and training environments. The PQC-Safe Signing and Attestation Pipeline operationalizes this schema by integrating FIPS 204 (ML-DSA), FIPS 205 (SLH-DSA) for non-NSS archival integrity, hybrid signing, PQC-oriented certificate-chain evolution, and attestation into a unified mechanism for verifying artifact authenticity and pipeline integrity. The Supply Chain Assurance Maturity Model (SCAMM) complements these technical components by offering a structured, evidence-based method for assessing organizational readiness and guiding modernization efforts aligned with PQC transition timelines.

Together, the three contributions—(1) the MBOM-PQC schema, (2) the PQC-Safe Signing and Attestation Pipeline, and (3) the SCAMM maturity model—provide a durable foundation for AI supply chain security. They enable organizations to detect tampering, prevent model swap attacks, verify dataset lineage, and ensure that AI artifacts remain trustworthy throughout training, deployment, and continuous learning. They also align naturally with Zero Trust Architecture (ZTA) principles by enabling continuous verification of model integrity and provenance as part of broader enterprise trust decisions. The requirements-to-architecture traceability matrices in Sections 5.4 and 7.4 confirm that every schema component, pipeline stage, and maturity indicator is grounded in documented threats and cryptographic dependencies established through systematic evidence synthesis across policy, standards, and benchmark sources.

While the framework addresses critical gaps in current AI assurance practices, several challenges remain. PQC algorithms introduce performance and storage overheads that may impact constrained environments; upstream transparency varies widely across model providers; and continuous learning systems require further research to optimize provenance management and attestation frequency. Future work should focus on empirical validation of the proposed pipeline in operational AI environments, development of automated provenance extraction tools, integration with confidential computing and secure enclave technologies, and the establishment of community-maintained AI provenance registries that can reduce the ecosystem dependency burden identified in Section 8.4.3.

Despite these limitations, the framework presented in this paper offers a practical and forward-looking approach to securing AI supply chains in the post-quantum era. By combining rigorous provenance modeling, PQC-safe cryptographic mechanisms, and a structured maturity model, it provides organizations with the tools needed to build resilient, trustworthy AI systems capable of withstanding both current and emerging threats. As AI becomes increasingly embedded in mission-critical and high-assurance environments, such durable, cryptographically grounded supply chain assurance will be essential for maintaining operational integrity and strategic advantage.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, The Supplementary Materials include the complete 54-source evidence bibliography with tier assignments, the full extraction matrix, exclusion ledger, and confidence-tier summary.

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 54-source evidence set is provided in the Supplementary Materials. All 54 sources, including those not in-text cited in the manuscript body, are listed in the References section per MDPI requirements.

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

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

References

1. ReversingLabs. Malicious Machine Learning Packages Targeting ML Developers in PyPI; ReversingLabs Threat Research: Boston, MA, USA, 2023.
2. 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-xcvg-c98v-hjqc (accessed on 15 January 2026).
3. CISA. Software Supply Chain Attacks: Threat Landscape and Mitigations; Cybersecurity and Infrastructure Security Agency: Washington, DC, USA, 2023.
4. MITRE. ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems; MITRE Corporation: McLean, VA, USA, 2024. Available online: https://atlas.mitre.org/ (accessed on 15 January 2026).
5. NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0); National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023.
6. NIST. Secure Software Development Framework (SSDF), Version 1.1; SP 800-218; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2022.
7. NIST. ML-DSA: Module-Lattice-Based Digital Signature Algorithm; FIPS 204; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024.
8. NSA. Commercial National Security Algorithm Suite 2.0 (CNSA 2.0); National Security Agency: Fort Meade, MD, USA, 2022.
9. NIST. SLH-DSA: Stateless Hash-Based Digital Signature Algorithm; FIPS 205; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024.
10. 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
11. Liu, Y.; et al. A Survey on Model Watermarking and Provenance for Deep Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 987–1005.
12. NIST. ML-KEM: Module-Lattice-Based Key-Encapsulation Mechanism; FIPS 203; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024.
13. 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.
14. DoD CDAO. Responsible Artificial Intelligence (RAI) Toolkit; Chief Digital and Artificial Intelligence Office: Arlington, VA, USA, 2024. Available online: https://www.ai.mil/rai.html (accessed on 15 January 2026).
15. NIST. SP 800-208: Recommendation for Stateful Hash-Based Signature Schemes (XMSS and LMS); National Institute of Standards and Technology: Gaithersburg, MD, USA, 2020.
16. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, n71. https://doi.org/10.1136/bmj.n71
17. 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 15 January 2026).
18. 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 15 January 2026).
19. Carlini, N.; et al. Poisoning Web-Scale Training Datasets Is Practical. In Proceedings of the 2023 IEEE Symposium on Security and Privacy; IEEE: San Francisco, CA, USA, 2023.
20. Goldblum, M.; et al. Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses. *ACM Comput. Surv.* **2024**, *56*, 1–42.
21. Wu, B.; Chen, H.; Zhang, M.; Zhu, J.; Wei, S.; Yuan, C.; Shen, C. BackdoorBench: A Comprehensive Benchmark of Backdoor Learning. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022); Curran Associates: Red Hook, NY, USA, 2022.
22. Pearce, A.; et al. Model Inversion, Extraction, and Supply Chain Attacks on Machine Learning Systems. *IEEE Trans. Dependable Secur. Comput.* **2024**, *21*, 1123–1138.
23. Rieger, P.; Krauß, T.; Miettinen, M.; Dmitrienko, A.; Sadeghi, A.-R. DeepSight: Mitigating Backdoor Attacks in Federated Learning Through Deep Model Inspection. In Proceedings of the 2022 Network and Distributed System Security Symposium (NDSS); Internet Society: San Diego, CA, USA, 2022.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

24. Red Hat. Securing the Modern Software Supply Chain: AI Models and Container Images; Red Hat Blog: Raleigh, NC, USA, 2025. Available online: https://www.redhat.com/en/blog/securing-modern-software-supply-chain-ai-models-container-images (accessed on 15 January 2026).
25. NIST. Zero Trust Architecture; SP 800-207; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2020. https://doi.org/10.6028/NIST.SP.800-207
26. NIST. SP 800-193: Platform Firmware Resiliency Guidelines; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2022.
27. Google. Secure AI Framework (SAIF); Google: Mountain View, CA, USA, 2023. Available online: https://safety.google/cybersecurity-advancements/saif/ (accessed on 15 January 2026).
28. Kumar, R.S.S.; Lopez Munoz, G.; Maitre, M.; Minnich, A.; Chawla, S.; Dheekonda, R.S.R.; Zhang, L.; Siska, C.; Rakshit, S. New Research, Tooling, and Partnerships for More Secure AI and Machine Learning; Microsoft Security Blog: Redmond, WA, USA, 2023. Available online: https://www.microsoft.com/en-us/security/blog/2023/03/02/new-research-tooling-and-partnerships-for-more-secure-ai-and-machine-learning/ (accessed on 15 January 2026).
29. 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 15 January 2026).
30. NIST. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile; AI 600-1; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024.
31. 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
32. NIST. The NIST Cybersecurity Framework (CSF) 2.0; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024.
33. The White House. Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence; Washington, DC, USA, 30 October 2023.
34. 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
35. IETF. Internet X.509 Public Key Infrastructure — Algorithm Identifiers for the Module-Lattice-Based Digital Signature Algorithm (ML-DSA); RFC 9881; Internet Engineering Task Force, 2025. Available online: https://www.rfc-editor.org/rfc/rfc9881 (accessed on 15 January 2026).
36. 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 15 January 2026).
37. 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 15 January 2026).
38. 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 15 January 2026).
39. ISO/IEC 42001:2023; Information Technology — Artificial Intelligence — Management System; International Organization for Standardization: Geneva, Switzerland, 2023.
40. ISO/IEC 23894:2023; Information Technology — Artificial Intelligence — Guidance on Risk Management; International Organization for Standardization: Geneva, Switzerland, 2023.
41. 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 15 January 2026).
42. 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 15 January 2026).
43. 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.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2026 doi:10.20944/preprints202603.1963.v1

44. DoD. Data, Analytics, and Artificial Intelligence Adoption Strategy; Department of Defense: Washington, DC, USA, 2023.
45. 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).
46. OWASP Foundation; Ecma TC54. CycloneDX Bill of Materials Standard (ECMA-424); 2023. Available online: https://cyclonedx.org/specification/overview/ (accessed on 15 January 2026).
47. 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
48. 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.
49. 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); Springer: Cham, Switzerland, 2020.
50. Ladisa, P.; Plate, H.; Martinez, M.; Barber, O. A Taxonomy of Attacks on Open-Source Software Supply Chains. In Proceedings of the 2023 IEEE Symposium on Security and Privacy; IEEE: San Francisco, CA, USA, 2023.
51. 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 15 January 2026).
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 15 January 2026).
53. 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 15 January 2026).
54. Zhu, J.; et al. 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.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.