A FedRAMP and NIST Aligned Framework for Securing AI Systems in Government Clouds

Md. Farhad Rahman[1]; Shamsad Binte Ehsan[2]; Tawhidur Rahman[3]; Monira Mostafa[4]

[1]Department of Technology, HSBC, Bangladesh, Dhaka, Bangladesh
[2]Dept. of CSE, Military Institute of Science & Technology, Dhaka, Bangladesh
[3]Dept. of CSE, Faculty of Science & Technology, Bangladesh University of Professionals (BUP), Dhaka, Bangladesh
[4]Department of Technology, BEXIMCO IT Division, Dhaka, Bangladesh

Publication Date: 2026/02/19

Abstract: The presence of Artificial Intelligence (AI) systems on clouds systems deployed on the cloud infrastructure is becom- ing critical to the activities of the U.S. government, the country's defense, and controlled industries. Though the FedRAMP ap- proved government cloud platforms offer minimum safeguards to infrastructure and data, they are not suitable to deal with AI specific risks like model poisoning, adversarial manipulation, training data breach, and AI supply chain risks. This breach injects the issue of national security with the integration of AI systems into mission critical cloud environments. This paper aims at creating a national security oriented framework of securing AI systems used in the U.S government cloud infrastructure. The framework will attempt to address the lack of connection between current federal cloud security mandates and the risk profile specific to AI systems that perform throughout the entire AI lifecycle. The main deliverable of this work is the standards aligned security framework, which incorporates AI specific threat modeling with the established federal guidelines, such as the NIST AI Risk Management Framework, NIST SP 800-53, Zero Trust Architecture principles, and FedRAMP security controls. The framework cross over AI induced risks to the technical, operational, and governance controls that apply to multi tenant government cloud environments. The results of this paper indicate that the existing FedRAMP based cloud deployments were greatly missing governance and control in the areas of AI model integrity, data provenance, and shared responsibility in terms of AI security. The discussion shows that the current models of cloud security need clarified extensions to AI to handle the risks of national security related to cloud hosted AI solutions.

Keywords: *Artificial Intelligence Security (AIS); Government Cloud Security; FedRAMP; NIST AI Risk Management Frame- work; Gov Cloud; NIST 800-53.*

How to Cite: Md. Farhad Rahman; Shamsad Binte Ehsan; Tawhidur Rahman; Monira Mostafa (2026) A FedRAMP and NIST Aligned Framework for Securing AI Systems in Government Clouds. *International Journal of Innovative Science and Research Technology*, 11(2), 911-918. https://doi.org/10.38124/ijisrt/26feb600

I. INTRODUCTION

Modern misinformation operations are increasingly aug- mented by generative AI, automated account orchestration, and algorithmic content amplification. For government and critical infrastructure missions, misinformation is not only a public facing risk; it is also an internal analytic risk: contaminated open source intelligence, manipulated sensor derived data products, and adversarially crafted content can influence AI enabled triage, prioritization, and decision sup- port. When AI pipelines ingest untrusted external sources or interoperate with tool using agents, the attack surface expands from classical cyber compromise to semantic manipulation of model behavior. Consequently, AI security in govern- ment cloud environments must treat information integrity as a first class security objective, alongside confidentiality and availability. Government cloud environments (e.g., AWS GovCloud and Azure Government) are engineered to satisfy high assurance requirements and commonly operate under FedRAMP baselines [7]–[9]. However, FedRAMP's inherited controls and many cloud shared responsibility models were designed for traditional workloads. AI systems introduce distinct failure modes: (i) the model can be compromised without compromising the host; (ii) training and evaluation pipelines constitute a supply chain; (iii) inference time inputs can induce policy bypass or data exfiltration (e.g., prompt injection); and (iv) model behavior can drift or be strategically manipulated even when infrastructure is hardened. These characteristics create a national security gap: the compliance posture may be "green" while the AI capability remains vulnerable to mission-impacting manipulation. This research develops a practical, standards aligned framework to secure AI systems deployed in U.S. government cloud environments. A national security focused AI lifecycle threat model tailored to U.S. government cloud constraints (multi tenancy, inherited controls, regulated data). A mapping AI RMF results to NIST SP 800-53 control families and operational evidence artifacts A FedRAMP aligned AI control overlay. The AI systems should have a shared responsibility decomposition (agency, CSP, AI supplier) whereby someone is responsible to manage the model integrity, provenance, and monitoring. A methodology and metrics of control coverage, detection efficacy and operation response to representative AI threats.

➢ *Objectives*

- *Threat Modeling:* Identify AI-specific threats and adversary tactics relevant to multi-tenant government cloud deployments across the AI lifecycle.
- *Standards Alignment:* Translate NIST AI Risk Management Framework (AI RMF) outcomes into assessable controls consistent with NIST SP 800-53 and FedRAMP processes.
- *Architecture and Operations:* Integrate Zero Trust principles with AI lifecycle assurance to improve prevention, detection, response, and recovery.

➢ *Paper Organization*

The paper is organized as follows: Section II reviews the lit- erature; Section III discusses related work; Section IV presents the proposed framework; Section V describes the methodology and system architecture; Section VI details the experimental setup; Section VII analyzes the results; Section VIII discusses the limitations; Section IX outlines future work; and Section X concludes the paper.

II. LITERATURE REVIEW

AI security research establishes a taxonomy of threats including poisoning, evasion, extraction, and privacy leakage [1]. Systematic reviews of machine learning security in cloud contexts highlight that MLOps introduces dependency and pipeline risks not captured by traditional application security controls [2]. NIST's AI RMF provides risk management functions and trustworthiness characteristics but is intentionally non prescriptive, requiring translation into implementation controls [3]. NIST's adversarial ML taxonomy standardizes terminology and supports consistent threat modeling [4]. MITRE ATLAS provides tactics and techniques enabling threat informed testing and defensive engineering [5]. Zero Trust literature emphasizes continuous verification and least privilege; however, ZTA does not, by itself, prevent manipulation of model behavior [6]. FedRAMP standardizes cloud security assessment and authorization using NIST SP 800-53 controls [7]. Provider guidance for government cloud offerings describes compliance capabilities but largely defers AI workload security to customer configuration and governance [8], [9].

Table I summarizes representative studies and standards, emphasizing datasets, models/systems, metrics, and findings relevant to building a FedRAMP aligned AI control overlay.

Table 1 Literature Review Summary

Source	Authors/Org	Datasets	Models/Systems	Metrics	Key Findings / Relevance
[1]	Survey	N/A	DL Systems	Threat Taxonomy	Establishes core ML threat classes and privacy risks; motivates lifecycle defenses beyond perimeter security.
[2]	Systematic Review	Multiple	Cloud ML/MLOps	Coverage Analysis	Identifies that cloud ML security depends on secure pipelines, provenance, isolation, and monitoring.
[3]	NIST	N/A	Risk Framework	Governance Mapping	Defines outcomes (Govern/Map/Measure/Manage); requires translation into assessable controls.
[4]	NIST	N/A	Threat Taxonomy	Terminology	Enables consistent threat modeling, supports control
[5]	MITRE/Com munity	N/A	AI ATT&CK-like TTPs	TTP Mapping	Supports threat-informed defense, red teaming, and incident response for AI systems.
[6]	SpringerOpen	N/A	ZTA architectures	Synthesis	ZTA improves access control; additional controls needed for model integrity and inference abuse.
[7]	FedRAMP	N/A	Authorization	Compliance	Federal baseline for cloud security; AI requires explicit overlay for artifacts/pipelines/inference.

III. RELATED WORK

Adversarial machine learning research demonstrates that attackers can manipulate AI behavior without compromising the underlying compute substrate [1]. Attacks can occur at data ingestion (poisoning), training (backdoors), evaluation (benchmark manipulation), and inference (evasion, prompt injection). Unlike conventional application flaws, these attacks often exploit statistical learning properties or human in the loop processes (labeling, prompt design, policy tuning). Con- sequently, effective defense requires controls on data prove- nance, training integrity, evaluation robustness, and inference time policy enforcement [15].

FedRAMP aligns cloud authorizations around standardized NIST SP 800-53 controls [7]. Government cloud offerings provide compliance enabling primitives (identity, encryption, logging, segmentation) [8], [9]. However, the AI workload security posture is subject to customer governance choices: model purchase, dependency management, and operational monitoring of AI particular misuse. Multi tenant and managed service dependencies further complicate accountability for artifacts, telemetry, and incident response [16]

Security, resilience, robustness, and transparency are some of the trustworthiness attributes outlined by NIST AI RMF [3]. But these outcomes need to be integrated into current practices of the RMF/FedRAMP by the agencies. One of the key issues is that ATO packages and SSPs often do not have clear evidence artifacts that allow testing model integrity, data provenance, and robustness [12].

ZTA can constrain access to datasets and inference endpoints through continuous authorization and least privilege [6]. Nonetheless, ZTA alone cannot prevent behavior manipulation (e.g., prompt injection into authorized sessions) or ensure model artifact integrity. Therefore, ZTA must be extended with AI-specific guardrails, policy enforcement points, and telemetry for behavior anomalies.

IV. PROPOSED NATIONAL SECURITY FRAMEWORK

The framework is designed for national security and regu- lated operations, emphasizing:

- *Lifecycle assurance:*
  Controls must cover data, model, pipeline, and operations.

- *Threat informed defense:*
  Controls and testing derived from recognized AI threat taxonomies [4], [5].

- *FedRAMP compatibility:*
  Use an overlay approach that augments (not replaces) SP 800-53/FedRAMP practices [7].

- *Zero Trust enforcement:*
  Continuous verification for humans, services, and agents.

- *Evidence driven security:*
  Define auditable artifacts (attestations, logs, test reports) suitable for ATO and continuous monitoring.

- *Framework Structure*

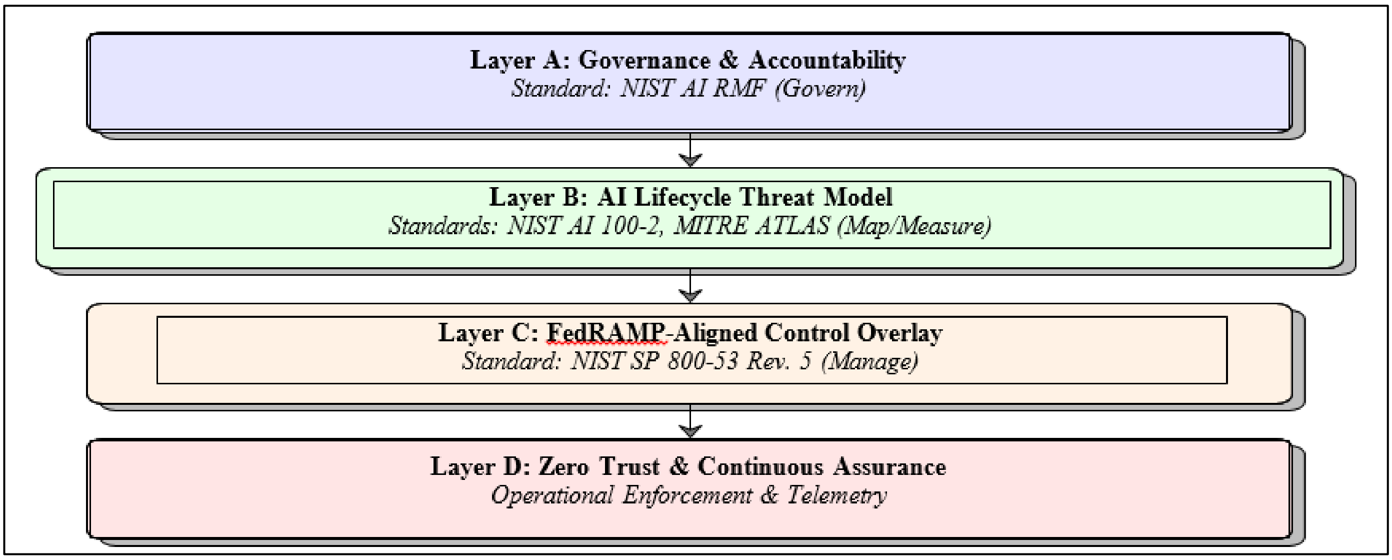
Layer A: Governance & Accountability
Standard: NIST AI RMF (Govern)
Layer B: AI Lifecycle Threat Model
Standards: NIST AI 100-2, MITRE ATLAS (Map/Measure)
Layer C: FedRAMP-Aligned Control Overlay
Standard: NIST SP 800-53 Rev. 5 (Manage)
Layer D: Zero Trust & Continuous Assurance
Operational Enforcement & Telemetry

Fig 1 Proposed National Security Framework Structure Integrating NIST Standards.

The framework consists of four integrated layers:

- Layer A: Governance and Accountability (AI RMF Govern outcomes) [3].
- Layer B: AI Lifecycle Threat Model (Map/Measure) using NIST AI 100-2 and ATLAS [4], [5].
- Layer C: FedRAMP-Aligned Control Overlay (Man- age) mapping threats to SP 800-53 control families and operational procedures [7].
- Layer D: Zero Trust + Continuous Assurance Architecture (operational enforcement and telemetry).

- *Threat Model Tailored to Government Cloud AI*

We model adversaries with capabilities relevant to national security settings. The threat model considers multiple ad- versary classes relevant to multi tenant government cloud environments. External adversaries interact with AI systems through inference APIs and publicly exposed interfaces, lever- aging open source intelligence and social engineering to con- duct model extraction, evasion, and prompt injection attacks. Supply chain adversaries compromise upstream dependencies such as container images, pretrained models, third party libraries, or data sources, enabling the insertion of hidden backdoors or malicious logic. Privileged or insider integrators are an issue as they manipulate training data, evaluation results, or deployment promotion processes, thus compro- mising the integrity of the models and compromising their reliability. Cross tenant adversaries use misconfigurations, shared services or hardware and software side channels to obtain unauthorized entry of sensitive data or model arti- facts. Examples of specific assets throughout the AI lifecycle are labeled datasets, feature stores, training pipelines, model weights, evaluation artifacts, prompts and system instructions, tool connectors, cryptographic secrets, inference logs, and mission critical outputs.

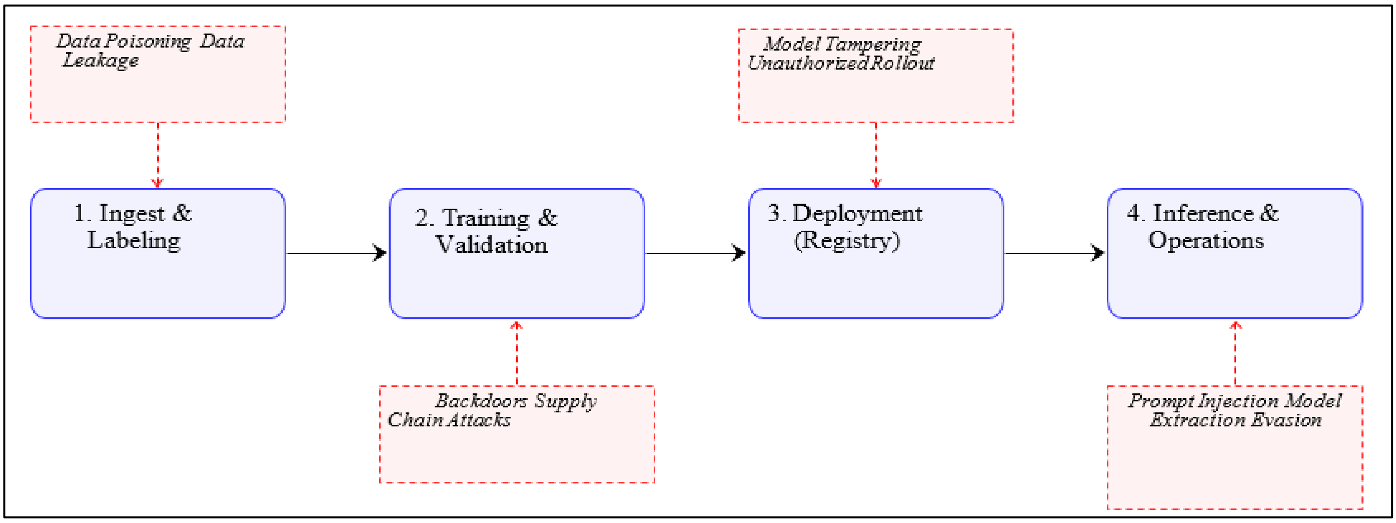
Data Poisoning Data Leakage
Model Tampering Unauthorized Rollout
1. Ingest & Labeling
2. Training & Validation
3. Deployment (Registry)
4. Inference & Operations
Backdoors Supply Chain Attacks
Prompt Injection Model Extraction Evasion

Fig 2 AI Lifecycle Threat Model mapped to operational phases.

➢ *FedRAMP Aligned AI Control Overlay (Concept)*

FedRAMP baselines typically address system security at the infrastructure/platform level. The overlay proposed here introduces *AI specific control interpretations* and *new evidence artifacts* while staying within the FedRAMP-aligned gover- nance model. Examples of overlay requirements include:

- *Dataset Provenance and Integrity Evidence:*

  Signed dataset manifests, lineage records, write-path restrictions, quarantine procedures.

- *Model Artifact Integrity Evidence:*

  Signed model arti- facts, registry access controls, reproducible build attesta- tions.

- *AI Evaluation Evidence:*

  Robustness testing reports, back- door checks, prompt injection test suites, drift monitoring thresholds.

- *Inference Misuse Monitoring:*

  Telemetry and detection rules for extraction, abuse, and policy bypass attempts.

➢ *AI Control Overlay Specification*

Operationalizing the overlay requires mapping threats to concrete control interpretations and evidence. The framework defines AI specific overlay requirements along three axes:

- *Artifact Scope:*

  Datasets, model weights, prompts and system instructions, evaluation artifacts, tool configura- tions, and retrieval connectors.

- *Lifecycle Phase:*

  Ingest, label, train, validate, deploy, operate, and retire.

- *Control Family Interpretation:*

  How SP 800-53 Rev. 5 families apply to AI artifacts and workflows.

For example, for data poisoning attacks during ingest and training, the overlay extends:

✓ SA (System and Services Acquisition) to require documented security requirements for data providers, contractual controls for labeling vendors, and acceptance criteria that include poisoning resilience testing.
✓ CM (Configuration Management) to require immutable storage for raw data, change controlled preprocessing code, and gated promotion of curated datasets into training catalogs.
✓ AU (Audit and Accountability) to require that all write path operations into training datasets are attributable, logged, and periodically reviewed for anomalous patterns.

Similarly, for prompt injection and tool misuse in agentic systems, the overlay interprets:

✓ AC (Access Control) to enforce least privilege tool bindings, separate roles for prompt authors versus tool integrators, and explicit allow lists for external connectors.
✓ SC (System and Communications Protection) to require isolation boundaries and sandboxes for tool executed actions, with constrained interfaces back into mission systems.
✓ *IR (Incident Response)* to define AI specific playbooks (e.g., disabling a compromised tool connector, revoking

prompt templates, or quarantining a memory store) as part of the overall cloud IR plan.

These interpretations remain compatible with existing Fe- dRAMP documentation artifacts (System Security Plan, con- tinuous monitoring reports, POA&M), but introduce AI spe- cific evidence such as SBOMs for models, red team test reports, and drift monitoring dashboards [3]–[5].

A. SP 800-53 Rev. 5 and FedRAMP Rev. 5 Context

NIST SP 800-53 Rev. 5 and its associated SP 800-53B baselines introduced several changes that are directly relevant to AI risk, including an expanded focus on privacy and the creation of the SR (Supply Chain Risk Management) control family [7], [13]. FedRAMP Rev. 5 baselines align with these updates, emphasizing updated configuration management dili- gence, enhanced privacy considerations, and structured supply chain risk management for cloud services [10], [11].

Within this context, the proposed overlay treats AI artifacts and pipelines as *first-class objects* subject to Rev. 5 expecta- tions:

- *Supply Chain Risk Management (SR):*
  SR controls are interpreted to cover *model supply chains*, including pre- trained models, third party datasets, open source libraries, and managed AI services. Evidence expectations include software bills of materials (SBOMs), model provenance attestations, vetted source repositories, and continuous monitoring of third party advisories.

- *System and Information Integrity (SI):*
  SI controls are extended from traditional malware and configuration anomalies to include AI specific anomalies such as unexpected behavior drift, backdoor activation patterns, and prompt injection signatures derived from NIST AI 100-2 and MITRE ATLAS taxonomies [4], [5].

- *Configuration Management (CM):*
  CM controls are ap- plied to the *full AI lifecycle*: data preprocessing pipelines, feature extractors, training configurations, evaluation scripts, and inference time policy templates. Promotion of AI artifacts across environments requires signed releases and reproducible pipelines.

- *Access Control (AC) and Audit and Accountability (AU):*
  AC and AU are interpreted for fine grained control of model registries, dataset catalogs, labeling tools, and inference endpoints, with audit trails designed to support reconstruction of AI relevant incidents (e.g., poisoning campaigns or misuse of agent tools).

By grounding the overlay in SP 800-53 Rev. 5 and Fe- dRAMP Rev. 5 baselines, agencies avoid introducing a separate AI compliance regime while still satisfying updated expectations for supply chain and integrity risk management in cloud environments [7], [11].

V. METHODOLOGY

➢ *Assurance Workflow*

We propose a repeatable workflow to integrate AI security into federal cloud governance:

- *Define Authorization Boundary:*
  Enumerate AI compo- nents (data pipelines, model registry, inference service, connectors) and inherited controls.

- *Threat Model Across Lifecycle:*
  Map threats to phases (data, training, eval, deploy, operate).

- *Derive Control Overlay:*
  Interpret SP 800-53 families for AI artifacts and pipelines; document responsibilities.

- *Implement ZTA Enforcement:*
  Identity, device posture, service to service auth, micro segmentation, policy en- forcement at AI interfaces.

- *Collect Evidence Artifacts:*
  Provenance manifests, at- testations, logs, test reports.

- *Continuous Monitoring:*
  Detection engineering for AI misuse, behavior drift, and supply chain anomalies.

➢ *Reference Architecture (Textual Description)*

A deployable reference architecture for securing AI systems in government cloud environments is organized into multiple interdependent planes that collectively support secure devel- opment, deployment, and operation across the AI lifecycle. The data plane provides foundational controls for data in- tegrity and provenance through the use of object storage with immutable buckets, structured dataset catalogs, and lineage tracking mechanisms. Data labeling workflows incorporate quality assurance gates to reduce the risk of data poisoning and ensure traceability. The MLOps plane enables secure model development by enforcing isolated training environments, continuous integration and delivery pipelines for machine learning workflows, strict dependency pinning, and the use of signed container images verified through an attestation service. The model plane manages model artifacts using a central- ized registry with cryptographic signing, controlled promotion across development, testing, and production environments, and rollback capabilities to support rapid recovery. The inference plane exposes model functionality through an API gateway that enforces strong authentication, rate limiting, and policy based controls on prompts, tools, and retrieval connectors operating under least privilege principles. Finally, the security plane provides centralized logging, integration with SIEM and SOAR platforms, detection rules tailored to AI specific threats, and predefined incident response playbooks to support timely detection, response, and recovery.

➢ *Shared Responsibility and Continuous Authorization*

Traditional FedRAMP shared-responsibility models distin- guish between CSP managed infrastructure/platform controls and customer managed application controls. AI workloads introduce a third actor: the AI supplier (e.g., foundation model provider, model hub, or managed AI service operator). The proposed framework decomposes responsibilities as follows:

- *CSP:*

  Enforces FedRAMP aligned controls for compute, storage, networking, identity, encryption, and logging; provides primitives for isolation, key management, and monitoring in government cloud regions [8], [9].

- *Agency (Mission Owner):*

  Owns AI use case definition, data acquisition and labeling governance, model selection and evaluation criteria, deployment policies, and incident response playbooks for AI misuse.

- *AI Supplier:*

  AI supplier is responsible for model development practices, pre deployment testing and red teaming, model documentation (e.g., cards and datasheets), vulnerability disclosure processes, and timely distribution of security advisories.

Continuous authorization to operate (cATO) is implemented by automating evidence generation across this shared responsi- bility stack. Build systems emit signed attestations for contain- ers, pipelines, and model artifacts; evaluation pipelines export robustness and leakage test results; and cloud native log- ging feeds AI specific detections into the existing FedRAMP continuous monitoring cadence. This supports risk based, incremental updates to AI components without requiring full re authorization for every model iteration, while preserving traceability to NIST AI RMF outcomes [3], [7].

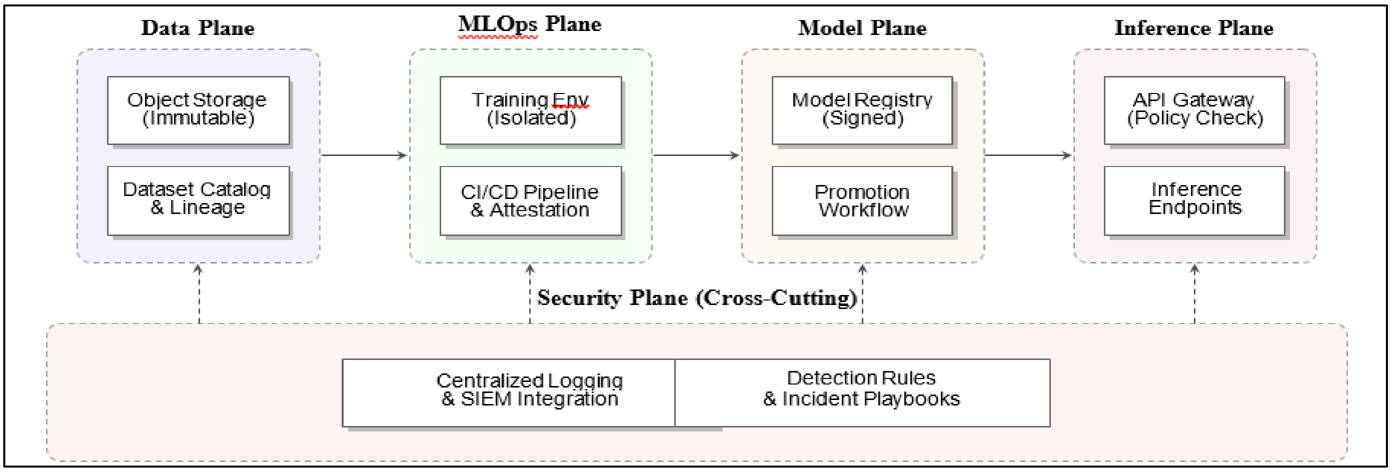
Data Plane
Object Storage (Immutable)
Dataset Catalog & Lineage
MLOps Plane
Training Env (Isolated)
CI/CD Pipeline & Attestation
Model Plane
Model Registry (Signed)
Promotion Workflow
Inference Plane
API Gateway (Policy Check)
Inference Endpoints
Security Plane (Cross-Cutting)
Centralized Logging & SIEM Integration
Detection Rules & Incident Playbooks

Fig 3 Reference Architecture for Securing Cloud AI Systems. The Security Plane Provides Cross-Cutting Monitoring and Enforcement Across All Lifecycle Phases.

VI. EXPERIMENTAL SETUP

Given the constraints of accessing real federal systems, we define a representative government-cloud AI service architec- ture and a set of simulated adversarial behaviors aligned to known threat classes [4], [5]. The intent of this section is to specify how the proposed framework *could* be evaluated in practice, rather than to report fully implemented experiments with quantitative measurements. The scenarios E1–E5 are therefore evaluation blueprints that agencies or practitioners can instantiate on their own infrastructures.

➢ *Evaluation Scenarios*

We outline five representative evaluation scenarios that exercise different parts of the overlay:

- *E1 (Poisoning Controls):*

  Introduce poisoned sam- ples and evaluate prevention/detection via provenance + anomaly gates.

- *E2 (Supply Chain Backdoor):*

  Insert a malicious depen- dency/model and evaluate attestation/signature enforce- ment.

- *E3 (Prompt Injection):*

  Craft injection payloads to in- duce tool misuse or sensitive retrieval; evaluate policy enforcement and logging.

- *E4 (Extraction/Abuse):*

  Simulate high rate and adaptive querying; evaluate throttling, anomaly detection, and response.

- *E5 (Drift + Integrity Monitoring):*

  Simulate distribution shift and unauthorized model promotion; evaluate drift alerts and CM gates.

Each experiment is instantiated on a reference architecture that reflects common patterns seen in FedRAMP authorized environments: an isolated government cloud VPC/VNet, man- aged identity and access management, an MLOps platform (e.g., a managed training service) scoped to the authorization boundary, a model registry, and an API gateway fronting inference endpoints. Data sources include a mix of synthetic mission like telemetry and curated open source datasets hosted in immutable object storage, with lineage tracked through a catalog and metadata service [8], [9].

➢ *Metrics*

- *Threat Coverage (TC):*
  Fraction of identified threats with at least one preventive and one detective control.

- *Detection Rate (DR):*
  Proportion of simulated attacks detected.

- *MTTD/MTTR:*
  operational timeliness for AI specific incidents.

- *Artifact Integrity Compliance (AIC):*
  Percent of de- ployments verified by signature + attestation.

- *Policy Enforcement Success (PES):*
  Percent of prohib- ited tool/data accesses blocked at runtime.

VII. RESULTS AND ANALYSIS

The following analysis is qualitative and expectation based: it describes the governance gaps the framework is designed to address and the types of improvements we anticipate when the overlay is instantiated in real deployments, rather than reporting measured metrics from a specific implementation.

➢ *Key Governance Gaps Observed*

Across typical "cloud compliant" deployments, three gaps consistently drive national security exposure: (1) Provenance and integrity gaps: datasets and model weights often lack chain of custody evidence and integrity validation, enabling subtle manipulation. (2) Inference time blind spots: standard logging may not capture prompt patterns, tool invocation, retrieval queries, or model behavior anomalies needed for investigations and (3) Responsibility ambiguity: when using managed AI services, responsibility for model security, eval- uation rigor, and incident handling is frequently unclear.

➢ *Overlay Effectiveness (Qualitative)*

The overlay improves assurance by converting AI RMF out- comes into auditable evidence artifacts suitable for FedRAMP style governance:

- *Stronger Prevention:*
  Signed artifacts and controlled promotion reduce unauthorized model changes.

- *Improved Detection:*
  Inference time telemetry and anomaly rules detect prompt injection and extraction behaviors earlier.

- *Faster Response:*
  AI specific IR playbooks (quarantine dataset, rollback model, rotate tool credentials) reduce MTTR.

➢ *Threat-to-Control Mapping (Operationalized)*

Table II provides an operational mapping between AI threat categories, AI lifecycle phase, example FedRAMP/NIST SP 800-53 control families, and required evidence artifacts. The mapping is presented at the family level to remain broadly applicable across baselines and agency overlays.

Table 2 AI Threats Mapped to FedRAMP/NIST SP 800-53 Control Families and Evidence Artifacts

Threat Category	Lifecycle Phase	Control Families (Examples)	Evidence Artifacts (Examples)
Data poisoning / label manipulation	Data, Training	SA, SI, AU, CM, MP	Dataset manifests, lineage records, labeling QA logs, write path IAM policies, quarantine procedures, anomaly gate reports.
Backdoored pretrained model / dependency compromise	Acquisition, Training, Deploy	SA, CM, SI	Approved source list, SBOM, signature verification logs, build attestation, registry access logs, reproducibility notes.
Model inversion / training data leakage	Training, Inference	AC, SC, SI, AU	Access policies to sensitive data, privacy risk assessment, inference logging policy, test results for leakage probes.
Model extraction / API abuse	Inference	AC, AU, SC, SI	API gateway policies, rate limits, anomaly detection alerts, investigation playbooks, throttling logs.
Prompt injection / tool misuse (agentic systems)	Inference	AC, AU, SC, SI, IR	Prompt/tool policy definitions, tool allow lists, retrieval access logs, blocked action logs, red team test reports.
Unauthorized model promotion / rollback failure	Deploy, Ops	CM, AU, IR, CP	Change tickets, signed releases, promotion approvals, rollback tests, incident records, continuity plans.
Behavior drift / concept drift	Ops	SI, CA, AU	Drift dashboards, monitoring thresholds, re-evaluation triggers, periodic robustness reports, audit logs.

VIII. LIMITATIONS

This paper is primarily a standards aligned framework and overlay design; limitations include:

➢ *Generalized Evaluation:*

Results are based on represen- tative architectures and simulated adversaries; validated operational outcomes require deployment specific testing.

➢ *Control ID Granularity:*

This version maps at the con- trol family level; agencies may require explicit control ID overlays tailored to FedRAMP Low/Moderate/High baselines [14]

IX. FUTURE WORK

The next stage in the research is to develop the proposed framework until it can be implemented in operational settings of the U.S. government clouds. One of these directions is the creation of a specialised FedRAMP AI Overlay Appendix that expressly aligns AI specific risks and mitigations to NIST SP-800-53 Revision 5 control identifiers and FedRAMP parameter specifications, which will help make decisions regarding authorization of AI enabled systems more uniformly. Further development is required to specify continued Authorization to Operate (cATO) through automating the collection of evidence at the AI pipelines such as cryptographic attestation of train- ing and inference environments, standardized model testing reports, as well as continuous monitoring of data and model drift. The structure must also be scaled to incorporate new AI agent architectures, which execute autonomous or semi autonomous, and focus on sandboxing execution, restricted and policy controlled tool application, and non repudiation logging to facilitate accountability and forensic examination. Lastly, formalized assessment criteria in determining prompt injection resistance, model integrity, and adversarial resistance in controlled cloud settings should be applied in future re- search.

X. CONCLUSION

To secure AI systems it is necessary to extend the lifecycle traditional cloud compliance postures to the specific AI related threats. This paper has suggested a framework related to a national security that operationalizes NIST AI RMF results in the FedRAMP aligned governance including an AI control overlay, threat informed testing, Zero Trust integration, and evidence based continuous assurance. The results suggest that the existing deployments lack adequate provenance and integrity controls of datasets and model artifacts, inference time telemetry of abuse and semantic attacks, and shared responsibility of AI supply chain risk clearly defined. The overlay suggested here offers a viable avenue through which the agencies, CSPs, and AI suppliers can enhance resilience and reliability of the cloud hosted AI functions without including a different compliance regime.

REFERENCES

[1]. N. Papernot *et al.*, “Security and Privacy Issues in Deep Learning,” *arXiv preprint arXiv:1807.11655*, 2018. [Online]. Available: https://arxiv.org/abs/1807.11655

[2]. “Securing Machine Learning in the Cloud: A Systematic Review,” *NCBI/PMC*, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7931962/

[3]. NIST, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” NIST AI 100-1, 2023. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

[4]. NIST, “Adversarial Machine Learning: A Taxonomy and Terminology,” NIST AI 100-2, 2025. [Online]. Available: https://csrc.nist.gov/pubs/ai/ 100/2/e2025/final

[5]. “MITRE ATLAS: Adversarial Threat Landscape for AI Sys- tems,” 2024. [Online]. Available: https://www.redhat.com/en/blog/ harden-your-ai-systems-applying-industry-standards-real-world

[6]. “The Role of AI in Zero Trust Architecture: A Review,” 2024. [Online]. Available: https://jesit.springeropen.com/articles/10. 1186/s43067-024-00155-z

[7]. FedRAMP, “FedRAMP Program,” 2026. [Online]. Available: https://www.fedramp.gov

[8]. Amazon Web Services, “AWS GovCloud (U.S.) Security Overview,” 2026. [Online]. Available: https://aws.amazon.com/govcloud-us/

[9]. Microsoft, “Azure Government Compliance Documentation,” 2026. [Online]. Available: https://learn.microsoft.com/en-us/azure/ azure-government/documentation-government-compliance

[10]. “FedRAMP Rev. 5 Baselines Transition,” 2025. [Online]. Available: https://www.fedramp.gov/documents-templates/

[11]. StandardFusion, “NIST SP 800-53 Rev. 5 and FedRAMP,” 2026. [Online]. Available: https://www.standardfusion.com/blog/ nist-sp-800-53-rev-5-and-fedramp

[12]. M. O. Faruq, “Vendor risk management in cloud-centric architectures: A systematic review of SOC 2, FedRAMP, and ISO 27001 practices,” *International Journal of Business and Economics Insights*, vol. 4, no. 1,pp. 1–32, 2024.

[13]. M. O. Faruq and M. J. I. Saidur, “Aligning FedRAMP and NIST frameworks in cloud-based governance models: Challenges and best practices,” *Review of Applied Science and Technology*, vol. 1, no. 1, pp. 1–37, 2022.

[14]. H. Teuscher, “Automating the RMF: Lessons from the FedRAMP 20x pilot,” *arXiv preprint arXiv:2510.09613*, 2025.

[15]. M. Sayduzzaman, S. Sazzad, M. Rahman, T. Rahman, and M. K. Uddin, “Managing escalating cyber threats: Perspectives and policy insights for Bangladesh,” Technical report, 2024.

[16]. M. Sayduzzaman and M. H. Nawab, “Blockchain-backed ML-based zero-trust honeypot for forensic-ready cyber-physical system security in Industry X,” *Journal of Computational Science and Applications*, vol. 2, no. 2, pp. 1–10, 2025.