The Runtime Verification Gap in Federal AI Deployment

A Reference Primer

The Barrier Between Accredited Classified Deployment and Verifiable Classified Deployment

Ziru Labs Research Publication · Version 1.0 · 2026 · Published under Ziru Labs corporate byline with contribution from Daniel Martin and Jan de Voor · Distributed under Creative Commons Attribution 4.0 International (CC BY 4.0)

Citable reference: Ziru Labs. *The Runtime Verification Gap in Federal AI Deployment: A Reference Primer*, v1.0. Published at zirulabs.com/research.

ABSTRACT

The United States Department of Defense classifies data sensitivity along a four-level Impact Level scale defined in the Defense Information Systems Agency Cloud Computing Security Requirements Guide. Commercial cloud and AI infrastructure has now been authorized into the classified tiers: Azure OpenAI Service holds DISA Impact Level 6 authorization for Department of Defense workloads and ICD 503 authorization for Top Secret workloads. That authorization is an accreditation decision about the deployment environment, achieved predominantly through physically isolated, government-accredited cloud regions. It is not hardware-rooted runtime verification of the AI computation itself. The barrier between accredited classified deployment and verifiable classified deployment is specific and technically diagnosable. No program office has publicly named it as a blocking constraint; the claim of this primer is that the barrier is structurally present across the most consequential federal and allied AI programs and will surface as procurement frameworks, most directly the FY2026 NDAA Section 1513 implementation, codify physical-security and cybersecurity requirements for AI systems.

This primer names that barrier the runtime verification gap: the distance between accreditation, which grants the authority to operate at a classification tier, and verification, which would prove what the AI computation did once it got there. It locates the gap in the AI stack, surveys the publicly disclosed deployment programs within which the gap operates, and identifies the substrate category required to close it. The substrate category is the trust layer for AI, developed analytically in the companion Reference Framework. The threat classes the substrate addresses are catalogued in the companion Threat Taxonomy. This primer concretizes the substrate-and-threat-class analytical material in the specific deployment context of federal classified AI workloads.

Ziru Labs authored this primer as one operator within the trust layer for AI category. The primer is offered as a reference on the federal deployment problem rather than as a description of any single product, and the analysis of the gap stands independently of any specific operator’s response to it. Where Ziru Labs technology is referenced, it is referenced as one operational response to the runtime verification gap.

1. The Impact Level Ladder

The Defense Information Systems Agency Cloud Computing Security Requirements Guide (DISA, current revision) defines four active security Impact Levels for DoD cloud service offerings: Impact Levels 2, 4, 5, and 6. The Impact Level scale is a principal reference for DoD cloud authorization, including cloud-hosted AI workloads. Each level corresponds to a category of data sensitivity and a corresponding set of technical, physical, and procedural controls under which infrastructure handling data at that sensitivity must operate.

Impact Level 2 covers non-controlled unclassified information, including information approved for public release and some non-public and non-sensitive data, typical of administrative systems and public-facing services. Impact Level 4 covers Controlled Unclassified Information and other mission-critical information, including non-classified defense operations and logistics. Impact Level 5 covers Controlled Unclassified Information requiring greater protection than Impact Level 4, as well as unclassified National Security Systems. Impact Level 6 covers classified information up to and including Secret, the tier at which intelligence analysis, classified operations, and the AI workloads supporting them run. Top Secret and Special Access Program data sit above Impact Level 6 and typically require government-operated infrastructure under specialized frameworks beyond the DISA Cloud Computing SRG Impact Level scale (DISA, current revision; ODNI ICD 503; CNSS Instruction 1253).

Commercial cloud providers including AWS GovCloud, Azure Government, Google Cloud for Government, and Oracle Government Cloud hold IL4 and IL5 authorization broadly, and classified-tier authorization has now extended to AI services: Azure OpenAI Service received DISA IL6 authorization and ICD 503 Top Secret authorization in early 2025, making it available for Department of Defense workloads across all government classification levels. That classified-tier authorization is achieved predominantly through physically isolated, government-accredited cloud regions with cleared personnel and facility controls, rather than through hardware-rooted verification of the AI computation on commercial silicon. What the authorization establishes is the authority to operate at the classification tier; what it does not establish is cryptographic evidence, at the moment of inference, that the AI computation ran under its authorized constraints. That distinction is the operational signature of the runtime verification gap at the deployment surface.

Allied classification systems map approximately onto the U.S. Impact Level scale at the highest tiers. The United Kingdom NCSC OFFICIAL, SECRET, and TOP SECRET tiers correspond approximately to the IL4, IL6, and Top Secret tier envelopes (UK NCSC, current). The NATO classification tiers (NATO UNCLASSIFIED, NATO RESTRICTED, NATO CONFIDENTIAL, NATO SECRET, COSMIC TOP SECRET) overlay onto member-state classification systems through NATO Standardization Agreement frameworks (NATO, current). The Australian Protective Security Policy Framework and the Canadian Government of Canada Security Policy establish analogous classifications. The runtime verification gap operates across all of these allied frameworks at structurally equivalent positions in their respective ladders.

 © 2026 Ziru Labs · CC BY 4.0

2. Why Accreditation Reaches the Classified Tier and Runtime Verification Does Not

Classified-tier accreditation is reachable today because the controls it requires are satisfied through physical isolation, accredited facilities, and software-boundary protection. Confidential computing, in the form of NVIDIA Confidential Computing, AMD SEV-SNP, Intel TDX, and ARM Confidential Compute Architecture, extends the boundary of trust around the running operating system and protects workload data from a compromised host (Costan and Devadas, 2016; NVIDIA Corporation, 2024). Confidential computing also produces hardware-rooted runtime attestation: cryptographic evidence about measured components, generated while the workload runs, used to prove that the platform is genuine and the environment is unmodified. That attestation is increasingly marketed under language such as runtime assurance and continuous compliance. In this primer, **runtime verification** refers to cryptographic evidence about the **AI computation itself** (model/weights/constraints) at the moment of inference, not only attestation that an execution environment is genuine. Confidential-computing attestation, deployed carefully, can extend beyond the platform itself: a measured workload launched inside an attested trusted execution environment can bind a hash-verified model artifact to the attested session, which is a meaningful form of workload-identity evidence. That binding, however, is only as strong as the software stack performing the measurement, holds only while the session is active, and terminates at the accelerator die and the confidential-computing session boundary. It does not provide evidence that the authorized constraints held during the computation, and it does not survive the physical-layer, interconnect, and governance-persistence threat surfaces that define the IL6+ threat model developed below.

The interconnect surface in particular is narrowing without closing. PCIe Integrity and Data Encryption (IDE), the TEE Device Interface Security Protocol (TDISP), the DMTF Security Protocol and Data Model (SPDM) for device attestation, CXL IDE, and protected multi-accelerator NVLink on current-generation platforms encrypt and authenticate traffic between attested endpoints (PCI-SIG, current; DMTF, current; CXL Consortium, 2022). Where deployed, these specifications materially reduce the passive-interception surface between silicon components. What they do not provide is evidence of the AI computation itself, protection against extraction from powered-off or physically accessed silicon, coverage of pre-session and post-session windows, or governance enforcement that persists across firmware and operating-system compromise. The companion Threat Taxonomy (Ziru Labs, 2026) scopes these confidential-computing and link-protection interactions threat class by threat class.

In parallel, a growing set of vendors market “verifiable compute,” “verifiable runtime,” “AI integrity,” or “continuous assurance” control planes that cryptographically log policy, code, and data lineage during training and inference and provide runtime compliance checks. These approaches are directionally aligned with the problem this primer names, but they primarily produce evidence about workflows and environments. Several of these control planes also anchor their attestations to an external ledger or network service to construct the audit trail, a dependency that is structurally unavailable inside an air-gapped classified enclave, where the verification evidence must root locally. They do not, by themselves, close the IL6+ threat surfaces that sit outside the confidential-computing boundary: physical extraction, bus-level interception, and compromise-resistant governance enforcement persistence. Hardware root of trust, present on every modern GPU and CPU, validates firmware integrity at boot (NIST, 2018; Trusted Computing Group, 2003). These capabilities are sufficient for

 © 2026 Ziru Labs · CC BY 4.0

workloads whose threat model assumes a software-layer adversary with limited physical access to the deployed silicon.

IL6 and above add three categories of threat that sit outside this boundary.

Physical-layer extraction. Physical extraction attacks against deployed hardware in contested or partially trusted environments may expose protected workload data, including memory interposition that compromises encrypted data during runtime and cold boot recovery of residual memory contents, such as model weights or inference state where present (Chuang et al., 2026; Halderman et al., 2008; Yitbarek et al., 2017). Exposure depends on the hardware, memory technology, and protection configuration, with the threat operating against silicon contents during runtime and after the trusted environment is no longer running. These attacks may fall outside the deployed trusted environment's protection guarantees and are structurally outside what software-layer defenses address.

Bus-level interception. Bus-level interception of inference traffic between silicon components, including memory bus, PCIe, and high-bandwidth interconnect surfaces (Markettos et al., 2019; CXL Consortium, 2022). Link-protection and device-attestation specifications (PCIe IDE, TDISP, SPDM, CXL IDE, protected NVLink) authenticate and encrypt traffic between attested endpoints where deployed; the residual surface comprises interconnects and transfer phases outside an active protected session, interposer-level and probe-level physical access to links and memory, and the absence of evidence about the computation the protected links carried.

Governance enforcement persistence. Governance enforcement that must persist across operating system and firmware compromise, such that compromise of the software stack does not bypass the constraints under which the AI workload is authorized to operate. The adversarial-input literature demonstrates that software-layer constraint mechanisms can be defeated at the model layer (Zou, Wang, Kolter, and Fredrikson, 2023; Anil et al., 2024); the governance-decoupling literature demonstrates that an enforcement mechanism can be made to assert compliance while behavior diverges (Contag et al., 2017). These studies illustrate failures of model-level safeguards and compliance reporting, respectively. This primer names a related systems-security concern: whether governance enforcement, and the evidence that it held, remain trustworthy when components below the enforcement layer are compromised. The companion Threat Taxonomy develops this surface as Attack Surface 5.

These threats span physical access to the silicon, interception on the paths between components, and compromise of the software and firmware beneath the enforcement layer; none can be addressed by a defense whose own integrity depends on that software stack. The substrate that addresses them must root in physical hardware characteristics rather than in software signatures, and must produce cryptographic evidence at runtime not merely that the platform is genuine, which confidential-computing attestation already provides, but that the AI computation itself is operating under its authorized constraints, across the physical, bus, and governance-persistence surfaces that platform attestation does not reach. This requirement is the runtime verification gap.

The Physics-Layer Threat Taxonomy for AI Infrastructure (Ziru Labs, 2026) catalogs the threat classes the runtime verification gap encompasses across five attack surfaces (physical access, bus and interconnect, firmware and boot, inference integrity, AI governance persistence). The Taxonomy is the canonical reference for the threat vocabulary the runtime verification gap operates against.

 © 2026 Ziru Labs · CC BY 4.0

3. The Programs in the Deployment Landscape

Several publicly disclosed federal and allied AI programs illustrate the environment in which the runtime verification gap operates. Each is moving AI into classified or high-assurance environments through infrastructure accreditation and physical isolation; per the distinction established in Section 1, none of the public disclosures speaks to hardware-rooted runtime verification of the AI computation itself.

The NVIDIA AI Factory for Government (NVIDIA Corporation, 2025), announced in October 2025 with Palantir Technologies, Lockheed Martin, CrowdStrike, and ServiceNow as named participants, is a full-stack reference design for deploying AI in FedRAMP-authorized clouds and high-assurance environments, including classified and mission-critical environments. The announcement describes accreditation-aligned and confidential-computing security at the platform layer; it does not address hardware-rooted runtime verification of the AI computation, the surface this primer identifies.

Microsoft Azure has authorized Azure OpenAI Service into the classified tiers. It received ICD 503 authorization for Top Secret workloads, announced January 16, 2025, as part of a twenty-six-product authorization on the Azure for U.S. Government Top Secret cloud, and DISA Impact Level 6 authorization for Department of Defense workloads, announced April 16, 2025, with the underlying authorization granted in early February 2025 (Microsoft Corporation, 2025). Azure OpenAI is therefore available for Department of Defense use across all government classification levels. That availability is established through accreditation and physically isolated government cloud regions.

Amazon Web Services classified regions and frontier-model availability. AWS has operated accredited classified infrastructure for the U.S. intelligence community since the Commercial Cloud Services arrangement of 2013, with the AWS Secret Region following in 2017, and frontier AI models are now available within those accreditation boundaries: Anthropic announced Claude Gov models for U.S. national security customers in June 2025 (Anthropic, 2025), and OpenAI announced OpenAI for Government in the same period (OpenAI, 2025). The Department of Defense Joint Warfighting Cloud Capability, the multi-vendor vehicle spanning AWS, Google, Microsoft, and Oracle through which DoD components procure cloud and AI compute across classification levels (U.S. Department of Defense, 2022), is the procurement surface much of this deployment moves through. As with the Azure authorizations, each of these is an accreditation-and-isolation posture; the public disclosures do not address hardware-rooted runtime verification of the AI computation.

The Palantir Technologies United States Army Enterprise Service Agreement, publicly announced on July 31, 2025 (U.S. Army, 2025), consolidates roughly seventy-five prior contracts into a ten-year purchasing vehicle valued at up to ten billion dollars, with the Army noting it is not committed to spending the full ceiling. It is a procurement and volume-discount instrument rather than a classified-deployment authorization; the Army announcement does not characterize a classified AI portion as constrained by a technical deployment gap. It is cited here as an indicator of the scale of federal AI procurement into which runtime verification will be relevant, not as evidence of a blocked workload.

The United Kingdom Ministry of Defence Palantir Technologies engagement comprises a strategic partnership announced in September 2025, under which Palantir committed to invest up to 1.5 billion pounds in the United Kingdom, and a three-year enterprise data-services contract worth 240.6 million pounds signed in

 © 2026 Ziru Labs · CC BY 4.0

December 2025 (United Kingdom Ministry of Defence, 2025). The 1.5 billion pound figure is an inward-investment commitment rather than a contract value. The engagement is cited here as part of the allied AI deployment landscape, not as evidence of a workload constrained by the runtime verification gap.

Allied sovereign AI programs across Five Eyes member states, NATO member states, and Gulf Cooperation Council member states are extending AI into classified and high-assurance environments under their domestic classification frameworks. Each nation's classification structure maps approximately onto the U.S. Impact Level structure and produces analogous runtime verification requirements at the highest tiers. Public disclosures from allied programs are fragmentary by classification posture; the runtime verification requirement is inferred from the structural parallel to the U.S. frameworks rather than from program-specific disclosure.

In addition to these accredited deployment programs, the commercial AI ecosystem is rapidly adding **runtime governance and integrity layers** (policy enforcement, lineage capture, and continuous assurance) aimed at agentic and autonomous workloads. The most visible recent instance is the NVIDIA Open Agent Safety Platform (NVIDIA Corporation, 2026), which pairs the open-source NVIDIA OpenShell runtime with NVIDIA Sentry, a reference system design that monitors agent activity and enforces access policy from BlueField-4 DPUs, outside the agent and host software. NVIDIA positions it as a response to agents circumventing application-layer controls. The announcement describes enforcement and attested telemetry for agent activity; it does not address evidence about the inference computation itself or the physical-extraction and bus-level surfaces named in Section 2. This market activity, including the move to hardware-based enforcement, is an important signal that runtime assurance is becoming a mainstream requirement. It does not eliminate the classified-tier requirement for cryptographic evidence that persists across the physical, bus and interconnect, and governance-persistence threat surfaces.

4. What Closes the Gap

The substrate category required to close the runtime verification gap is the hardware-rooted trust layer for AI, developed analytically in The Trust Layer for AI: A Reference Framework for the Category (Ziru Labs, 2026). The substrate's structural characteristics are specified in the Reference Framework at category level. At the operational level relevant to federal classified deployment, the substrate carries four distinguishing properties.

Physical hardware rooting. The substrate roots in physically derivable hardware characteristics rather than in software signatures, so that the integrity of the substrate is recoverable through attestation against hardware properties an adversary cannot reproduce or simulate at software register.

Runtime cryptographic evidence. The substrate produces cryptographic evidence at the moment of inference that the model, the weights, the operational constraints, and the silicon all match the authorized configuration. The evidence is verifiable by parties evaluating the inference output independently of the operator that produced it.

Compromise-resistant persistence. The substrate persists across operating system compromise and firmware compromise that exceed the design envelope of software trusted execution environments. The integrity of the substrate is not predicated on the integrity of the layers above it, which is the structural property that the IL6 and above threat model requires.

 © 2026 Ziru Labs · CC BY 4.0

Additive composition with existing silicon. The substrate composes additively with the existing silicon-vendor confidential computing layer rather than replacing it, so that deployment proceeds on commercial silicon rather than on bespoke replacement hardware. The deployment economics are tractable at the scale federal and allied programs require.

No deployment program in Section 3 has published requirements in these terms, and this primer does not claim otherwise. The four properties are offered as the substrate-level requirements that follow from the IL6+ threat model developed in Section 2, and they are the properties against which procurement frameworks, most directly the FY2026 NDAA Section 1513 framework as it codifies into the Defense Federal Acquisition Regulation Supplement, can be expected to evaluate candidate mechanisms. The substrate that satisfies them is the trust layer for AI in operational deployment; the analytical foundation for the category is in the Reference Framework.

5. Standards and Regulatory Trajectory

The standards-and-regulatory environment around the runtime verification gap is actively codifying across multiple bodies. The codification activity is consequential because the substrate-integrity specifications adopted during the codification window become the operational definition under which federal and allied classified AI deployment proceeds for the subsequent decade.

U.S. federal frameworks. FY2026 National Defense Authorization Act Section 1513, Physical and Cybersecurity Procurement Requirements for Artificial Intelligence Systems (P.L. 119-60, 2025), directs the Department of Defense to develop a risk-based framework for cybersecurity and physical-security standards for AI and machine-learning systems and to amend the Defense Federal Acquisition Regulation Supplement to incorporate the requirements. Executive Order 14179 (signed January 23, 2025) directs the development of the federal AI Action Plan, released July 2025, and Office of Management and Budget memoranda M-25-21 and M-25-22 (April 2025) set federal AI use and acquisition policy (The White House, 2025; OMB, 2025). Executive Order 14365 (Ensuring a National Policy Framework for Artificial Intelligence, signed December 11, 2025) shapes the broader national AI regulatory environment within which these activities proceed. On the civilian-agency side, the General Services Administration FedRAMP program prioritizes AI cloud services for authorization through the Emerging Technology Prioritization Framework (GSA FedRAMP, 2024) and is modernizing continuous monitoring under FedRAMP 20x. The National Security Agency and the Cybersecurity and Infrastructure Security Agency have published joint guidance on deploying AI systems securely (NSA and CISA, 2024), and the Department of Defense Chief Digital and Artificial Intelligence Office develops the operational guidance under which DoD components procure AI infrastructure.

European frameworks. Under the EU AI Act, Article 40 (harmonised standards and standardisation deliverables) establishes the presumption-of-conformity route, and Article 43 sets out the conformity assessment procedures themselves (European Commission, 2024). The Article 40 harmonised standards are in development through CEN-CENELEC JTC 21 in response to the European Commission’s standardisation request, and the technical implementation activities at the European Commission AI Office are developing the operational framework for European deployment.

 © 2026 Ziru Labs · CC BY 4.0

Allied frameworks. NATO STANAG work on AI trust (NATO Allied Command Transformation, current) and Five Eyes intelligence community AI framework development (current) develop the allied frameworks. The Gulf Cooperation Council Joint AI Coordination activities develop the corresponding regional framework. Each allied framework maps onto the substrate-integrity requirements that close the runtime verification gap at the corresponding classification tier.

Industry standards bodies. ISO/IEC 27090 (Cybersecurity, Artificial intelligence, Addressing security threats and compromises to artificial intelligence systems; at Final Draft International Standard stage as of the first half of 2026) develops AI security guidance under the ISO/IEC JTC 1/SC 27 information-security committee. Consistent with the Reference Framework's treatment, ISO/IEC 27090 addresses the software and data layers and is an adjacent standard the trust layer composes with rather than a hardware-substrate standard. ISO/IEC JTC 1/SC 42 working groups (ISO/IEC, current) are developing AI trustworthiness specifications including ISO/IEC 23894 (AI risk management). At the hardware and interconnect layer, PCI-SIG (PCIe IDE and TDISP), the DMTF (SPDM), the CXL Consortium (CXL IDE), and the IETF (Remote Attestation Procedures architecture, RFC 9334) develop the link-protection and device-attestation specifications the substrate composes with at the trusted execution environment boundary; the Confidential Computing Consortium and the Trusted Computing Group develop the reference-architecture and trusted-platform specifications at the same boundary; and the NSA Commercial National Security Algorithm Suite 2.0 sets the cryptographic transition baseline that evidence chains for national security systems must satisfy.

The codification window for these frameworks is open through approximately eighteen to thirty-six months from the publication date of this primer, a horizon anchored by the Section 1513 implementation-plan reporting to Congress beginning June 2026, the CEN-CENELEC JTC 21 harmonised-standards drafting schedule, and the expected publication of ISO/IEC 27090. The substrate-integrity requirements established during the codification window determine the operational definition of the runtime verification gap and the technical specifications under which the gap is closed.

6. Why the Gap Remains Structurally Open

The runtime verification gap is among the most concrete and most analytically available problems in federal artificial intelligence. Each named program in Section 3 has publicly disclosed its accreditation status and its deployment activity. A reader unfamiliar with federal classification categories who reads the primary sources cited in the bibliography can independently confirm that AI is being authorized into the classified tiers through accreditation and physical isolation, and can see that none of those disclosures speaks to hardware-rooted runtime verification of the AI computation, which is the distinct layer this primer addresses.

The reason the gap remains structurally open is that the substrate category required is a deep-tech hardware-rooted infrastructure layer, produced on a development cadence and through a capital-formation structure distinct from the dominant AI security commercial pattern. The dominant pattern is software-layer detection and monitoring (and, increasingly, governance-layer control planes that log and enforce policies at runtime), which reaches commercial and IL4 and IL5 markets, but cannot reach the runtime verification requirement at the physical layer. Hardware-based enforcement of agent behavior is now reaching the commercial market; the substrate this gap

 © 2026 Ziru Labs · CC BY 4.0

requires is distinct, because it must produce evidence about the computation itself and hold against physical access to the silicon. The substrate category requires patent-stage deep-tech engineering, hardware integration discipline, and cleared-handling capacity that the software-layer pattern is not configured to produce.

The Reference Framework's analysis of the trust layer category as a foundational substrate category (Sections 2 and 3 of Ziru Labs, 2026) develops the broader analytical reason why substrate categories at this depth do not emerge through commercial competition at the layers above. The reason in summary: the substrate's value is realized through application enablement across multiple markets rather than through direct competition within an established market, which produces the capital-formation and development-cadence mismatch with the dominant commercial pattern at the application layers above.

7. Operator Profile: Ziru Labs

Ziru Labs is an early-stage operator within the trust layer for AI category, developing a hardware-rooted substrate as one operational response to the runtime verification gap. The company has an intellectual property position established through filed provisional patent applications; it has not shipped a commercial product. Its initial work is focused on the security application surface, with the verification, provenance, compliance, identity, and agency surfaces of the category to follow as commercial configurations mature. The company offers this primer in a convening posture, inviting federal and allied program offices, silicon and platform vendors, standards bodies, and other operators to engage with the analysis.

The company is the author of the Physics-Layer Threat Taxonomy for AI Infrastructure (Ziru Labs, 2026), the Trust Layer for AI: A Reference Framework for the Category (Ziru Labs, 2026), the AI Infrastructure Stack and the Trust Layer Position: A Reference Primer (Ziru Labs, 2026), the Trust Layer Category Map: A Reference Primer (Ziru Labs, 2026), the Ziru Labs Capability Posture: A Reference Primer (Ziru Labs, 2026), and this primer. Together these documents establish the analytical foundation for the runtime verification gap as a category-level problem, the trust layer as the substrate-level response, the operator landscape adjacent to it, and the company's specific posture within the category. The founding team's prior experience spans U.S. federal classified AI, signals intelligence, and cryptographic warfare, including senior service in the U.S. intelligence community and the U.S. Navy, alongside experience in financial services and AI infrastructure; future versions of the corpus will carry named contributions from members of the team and from external contributors as the company's public posture develops. Ziru Labs intends to contribute to the standards work surveyed in Section 5 and is preparing submissions to relevant bodies.

Continued research is published at zirulabs.com/research, with the planned quarterly State of Physics-Layer AI Trust report tracking the evolution of the runtime verification gap and the substrate response.

Counterparties evaluating engagement with Ziru Labs route through the Engage pathway at zirulabs.com.

8. Bibliography

References are organized into five categories: academic literature (A), standards and regulatory documents (B), industry and vendor technical documentation (C), historical and analytical reporting (D), and framework

 © 2026 Ziru Labs · CC BY 4.0

integration references (E).

A. ACADEMIC LITERATURE

Anil, C., Durmus, E., Panickssery, N., Sharma, M., and co-authors (2024). “Many-shot Jailbreaking.” Advances in Neural Information Processing Systems 37 (NeurIPS 2024).

Chuang, J., Seto, A., Berrios, N., van Schaik, S., Garman, C., and Genkin, D. (2026). “TEE.fail: Breaking Trusted Execution Environments via DDR5 Memory Bus Interposition.” Proceedings of the IEEE Symposium on Security and Privacy.

Contag, M., Li, G., Pawlowski, A., Domke, F., Levchenko, K., Holz, T., and Savage, S. (2017). “How They Did It: An Analysis of Emission Defeat Devices in Modern Automobiles.” Proceedings of the IEEE Symposium on Security and Privacy.

Costan, V. and Devadas, S. (2016). “Intel SGX Explained.” IACR Cryptology ePrint Archive, Report 2016/086.

Halderman, J.A., Schoen, S.D., Heninger, N., Clarkson, W., Paul, W., Calandrino, J.A., Feldman, A.J., Appelbaum, J., and Felten, E.W. (2008). “Lest We Remember: Cold Boot Attacks on Encryption Keys.” Proceedings of the 17th USENIX Security Symposium.

Markettos, A.T., Rothwell, C., Gutstein, B.F., Pearce, A., Neumann, P.G., Moore, S.W., and Watson, R.N.M. (2019). “Thunderclap: Exploring Vulnerabilities in Operating System IOMMU Protection via DMA from Untrustworthy Peripherals.” Network and Distributed System Security Symposium (NDSS).

Yitbarek, S.F., Aga, M.T., Das, R., and Austin, T. (2017). “Cold Boot Attacks are Still Hot: Security Analysis of Memory Scramblers in Modern Processors.” IEEE International Symposium on High Performance Computer Architecture (HPCA).

Zou, A., Wang, Z., Kolter, J.Z., and Fredrikson, M. (2023). “Universal and Transferable Adversarial Attacks on Aligned Language Models.” arXiv:2307.15043.

B. STANDARDS AND REGULATORY DOCUMENTS

Defense Information Systems Agency (current revision). “Cloud Computing Security Requirements Guide.”

European Commission (2024). “Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).” Official Journal of the European Union.

Executive Order 14179 (2025). “Removing Barriers to American Leadership in Artificial Intelligence.” The White House, signed 23 January 2025; 90 FR 8741.

FY2026 National Defense Authorization Act, Section 1513 (2025). “Physical and Cybersecurity Procurement Requirements for Artificial Intelligence Systems.” Public Law 119-60.

Committee on National Security Systems (current). “CNSS Instruction 1253: Security Categorization and Control Selection for National Security Systems.”

Executive Order 14365 (2025). “Ensuring a National Policy Framework for Artificial Intelligence.” The White House, signed 11 December 2025; 90 FR 58499.

General Services Administration FedRAMP (2024). “Emerging Technology Prioritization Framework,” June 2024.

International Organization for Standardization / International Electrotechnical Commission (2026, Final Draft International Standard). “ISO/IEC 27090, Cybersecurity, Artificial intelligence, Addressing security threats and compromises to artificial intelligence systems.”

Internet Engineering Task Force (2023). “RFC 9334: Remote ATtestation procedureS (RATS) Architecture.”

National Security Agency (2022). “Commercial National Security Algorithm Suite 2.0 (CNSA 2.0).”

National Security Agency and Cybersecurity and Infrastructure Security Agency (2024). “Deploying AI Systems Securely: Best Practices for Deploying Secure and Resilient AI Systems,” joint guidance with international partners, April 2024.

Office of the Director of National Intelligence (current). “Intelligence Community Directive 503: Intelligence Community Information Technology Systems Security Risk Management, Certification and Accreditation.”

Office of Management and Budget (2025). “M-25-21: Accelerating Federal Use of AI through Innovation, Governance, and Public Trust” and “M-25-22: Driving Efficient Acquisition of Artificial Intelligence in Government,” April 2025.

 © 2026 Ziru Labs · CC BY 4.0

North Atlantic Treaty Organization (current). “NATO Standardization Agreement (STANAG) framework activities on AI trust.”

National Institute of Standards and Technology (2018). “NIST SP 800-193: Platform Firmware Resiliency Guidelines.”

Trusted Computing Group (2003 to present). “TCG TPM Specifications” and successor specifications.

United Kingdom National Cyber Security Centre (current). “Government Security Classifications Policy.”

C. INDUSTRY AND VENDOR TECHNICAL DOCUMENTATION

Amazon Web Services (current). “AWS Secret Region and AWS Top Secret Region” technical and accreditation documentation.

Anthropic (2025). “Claude Gov Models for U.S. National Security Customers,” June 2025.

Confidential Computing Consortium (current). “Technical Documentation and Member Working Group Output.”

CXL Consortium (2022). “Compute Express Link (CXL) Specification, Revision 3.0.”

DMTF (current). “DSP0274: Security Protocol and Data Model (SPDM) Specification.”

PCI-SIG (current). “PCI Express Integrity and Data Encryption (IDE)” and “TEE Device Interface Security Protocol (TDISP)” specifications.

International Organization for Standardization / International Electrotechnical Commission (current). “ISO/IEC JTC 1/SC 42 working group output on AI trustworthiness,” including ISO/IEC 23894:2023 (AI risk management guidance).

Microsoft Corporation (2025). “Unlocking new possibilities: Microsoft AI, database, and identity products receive authorization for Top Secret workloads,” Azure Government blog, 16 January 2025; and “Azure OpenAI Service authorized at DoD Impact Level 6,” Azure Government blog, 16 April 2025.

NVIDIA Corporation (2024). “NVIDIA Confidential Computing Deployment Guide.”

NVIDIA Corporation (2025). “AI Factory for Government Announcement,” October 2025.

NVIDIA Corporation (2026). “NVIDIA Launches Open Agent Safety Platform to Secure Agents From Testing to Deployment,” press release, 28 September 2026.

OpenAI (2025). “Introducing OpenAI for Government,” June 2025.

U.S. Department of Defense (2022). “Joint Warfighting Cloud Capability (JWCC) Contract Award Announcement,” December 2022.

U.S. Army (2025). “U.S. Army Awards Enterprise Service Agreement to Enhance Military Readiness and Drive Operational Efficiency,” announcement of 31 July 2025.

D. HISTORICAL AND ANALYTICAL REPORTING

United Kingdom Government (2025). “UK-Palantir Strategic Partnership Announcement,” September 2025.

United Kingdom Ministry of Defence (2025). “MOD Palantir Enterprise Agreement,” Find a Tender contract notice 083874-2025, published 17 December 2025; contract signed 30 December 2025.

Ziru Labs (2026). *The Trust Layer for AI: A Reference Framework for the Category, v1.0.* Published at zirulabs.com/research.

Ziru Labs (2026). *The Physics-Layer Threat Taxonomy for AI Infrastructure: A Reference Framework, v1.0.* Published at zirulabs.com/research.

Ziru Labs (2026). *The AI Infrastructure Stack and the Trust Layer Position: A Reference Primer, v1.0.* Published at zirulabs.com/research.

Ziru Labs (2026). *The Trust Layer Category Map: A Reference Primer, v1.0.* Published at zirulabs.com/research.

Ziru Labs (2026). *Ziru Labs Capability Posture: A Reference Primer, v1.0.* Published at zirulabs.com/research.

E. FRAMEWORK INTEGRATION REFERENCES

National Institute of Standards and Technology (2023). “AI Risk Management Framework (AI RMF 1.0),” NIST AI 100-1.

National Institute of Standards and Technology (2024). “The NIST Cybersecurity Framework (CSF) 2.0,” NIST CSWP 29.

MITRE Corporation (2024). “ATT&CK Framework, Enterprise Matrix,” Version 14.

 © 2026 Ziru Labs · CC BY 4.0

Acknowledgments

This primer draws on the publicly disclosed deployment programs cited in Section 3, the academic security literature cited in Section 2, and the federal and allied standards-and-regulatory work cited in Section 5. Specific acknowledgments to the DISA Cloud Computing Security Requirements Guide working community, the Office of Management and Budget AI procurement guidance development team, the FY2026 NDAA Section 1513 legislative drafting community, the EU AI Act technical development process under Articles 40 and 43, the NATO Allied Command Transformation AI capability gap working community, and the founding team's prior working experience across U.S. federal classified AI, signals intelligence, cryptographic warfare, financial services, and AI infrastructure. Initial draft review provided by the Ziru Labs founding team. Any errors in characterization are solely Ziru Labs' responsibility.

Citation

Ziru Labs. *The Runtime Verification Gap in Federal AI Deployment: A Reference Primer, v1.0.* Published at zirulabs.com/research.

 © 2026 Ziru Labs · CC BY 4.0