Air-Gapped LLM Deployment: Running Frontier Models Offline

Published July 21, 2026 42 min read

Executive Summary

Air-gapped large language model (LLM) deployment means running frontier or near-frontier AI models entirely inside a network that has no physical or logical route to the public internet: no pip install at runtime, no hosted LLM API, no screenshot leaving the room (Source: precisionfederal.com). As of July 2026, this pattern has moved from a defense-only niche to the default deployment shape for classified government workloads, and a fast-growing option for regulated industries that cannot tolerate any outbound dependency (Source: neuralchainai.com). The market backdrop is large and defense-weighted: obligated U.S. federal AI spending rose to $7.2 billion in 2026, a 966 percent increase from 2024, and the Department of Defense alone accounted for 98.9 percent of all potential federal AI contract value in 2026, equal to $90.7 billion (Source: www.ibm.com).

The regulatory scaffolding is well defined. The Department of Defense's Cloud Computing Security Requirements Guide sorts workloads into Impact Levels 2, 4, 5, and 6, with IL6 reserved for information classified up to the SECRET level and requiring a closed, self-contained cloud environment connected only to the Secret Internet Protocol Router Network (SIPRNet) (Source: learn.microsoft.com). FedRAMP applies a parallel but distinct Low, Moderate, and High baseline, with FedRAMP High reserved for systems where a breach could cause severe or catastrophic effects on government operations (Source: www.schellman.com). Both hyperscalers and specialized platforms have built air-gapped offerings against these frameworks: Microsoft deployed GPT-4 into an isolated, air-gapped Azure Government Top Secret cloud for the Department of Defense in 2024 (Source: defensescoop.com), Amazon Web Services operates a Top Secret Region it describes as the first air-gapped commercial cloud (Source: aws.amazon.com), and Anthropic's Claude Gov models are already deployed by agencies at the highest level of U.S. national security (Source: www.anthropic.com).

For genuinely disconnected, on-premises air-gapped environments, open-weight models have become the practical default because they can be hashed, signed, and imported once rather than called over an API; API-only commercial models are simply not available air-gapped, which is why open-weight deployment dominates classified AI today (Source: precisionfederal.com). Meta's Llama 3 family ships as pretrained and instruction-tuned 8B and 70B parameter models for broad self-hosted use (Source: ai.meta.com), most of Mistral's open models carry an Apache 2.0 license (Source: help.mistral.ai), Alibaba's Qwen models are licensed under Apache 2.0 (Source: github.com), and DeepSeek-V3 is released under the MIT

license and supports commercial use (Source: github.com). Scale AI's Defense Llama, a Llama 3 derivative built specifically for national security missions, is one purpose-built example already fielded in controlled government environments (Source: scale.com), and the U.S. Army's CamoGPT, built on open-source models from Meta, Google, and Mistral, has grown to 75,000 users since its spring 2024 deployment (Source: www.army.mil).

Hardware costs anchor the build decision. A single NVIDIA H200 GPU with 141 gigabytes of HBM3e memory costs between roughly $30,000 and $40,000 to buy outright (Source: jarvislabs.ai), and a fully configured 8-GPU DGX H200 system runs approximately $467,000 before support contracts, rising toward $584,000 at list price (Source: www.broadberry.com). A representative site-level redundant deployment for a 70B-class model, including networking and storage, runs several hundred thousand dollars for four to eight H100-class GPUs (Source: www.broadberry.com). Because no package registry is reachable at runtime, the hardest engineering problem is not the first import but the update pipeline: models, container images, and dependencies move across the boundary as signed bundles on approved physical media or through a one-way data diode, verified by hash and promoted through staging on a quarterly or monthly cadence rather than a continuous one (Source: neuralchainai.com). This report walks through the taxonomy, model choices, hardware architecture, software stack, security controls, and real deployments that define air-gapped LLM practice today.

Introduction and Background

Every large language model deployment that runs in a public cloud rests on an assumption so basic that it is rarely written down: the system can reach the internet. Vendor APIs, hosted evaluation judges, telemetry pipelines, model registries, and secret managers all quietly resolve to a domain somewhere outside the organization's own network. For most enterprise AI deployments, that assumption is harmless. For a narrow but consequential set of environments, defense and intelligence networks handling classified information, national laboratories working with Controlled Unclassified Information and export-controlled nuclear data, hospitals processing protected health information, and financial institutions bound by strict custody rules, that assumption cannot hold, and the entire deployment shape changes as a result (Source: tianpan.co).

Air-gapped LLM deployment refers to running an inference stack, and typically an entire retrieval-augmented generation pipeline, inside a network that has no route, physical or logical, to the public internet at any point during normal operation.This is a stronger claim than "isolated" or "private cloud." A private virtual cloud with a NAT gateway and an egress allowlist is segregated, not air-gapped, because controlled outbound traffic still exists; an auditor asking what the system phones home to gets a list of approved destinations rather than a definitive "nothing" (Source: www.truefoundry.com). A true air gap has no NAT gateway, no DNS server that resolves external hostnames, and no public certificate authority chain that the network trusts (Source: www.truefoundry.com). Every dependency must instead be brought in through an accredited transfer path, scanned for malware, and approved before it ever touches the enclave.

The reason this matters now, in mid-2026, is that the generative AI wave arrived at exactly the moment defense, intelligence, and regulated-industry organizations were also expanding their appetite for large language models. Precision Federal, a firm that designs classified and on-premises AI systems for federal customers, frames the design constraint bluntly: on-prem and air-gapped deployments have no vendor API, no automatic update, and no outbound telemetry, so every decision, from which model to which serving stack to which hardware, is made once and lived with for months before it can be revisited through a deliberate cross-domain transfer process (Source: precisionfederal.com). That constraint touches every layer of the stack: which open-weight model is licensed for the use case, how many GPUs and how much power the enclave needs, which serving framework can run without phoning home, how updates cross the boundary without breaking an accreditation package, and how the organization proves, to an auditor's satisfaction, that the gap has actually held. This report addresses each of those questions in turn, grounded in the classification frameworks, hardware pricing, model licenses, and named deployments that define the practice as of July 2026.

Defining the Air Gap: Taxonomy of Isolated AI Deployments

Not every "private" or "secure" AI deployment is air-gapped, and the terminology distinctions carry real legal and architectural weight. Understanding where an organization sits on this spectrum determines which model, hardware, and update strategy make sense.

Segregated or isolated deployments sit at the permissive end. These typically run in a private virtual cloud with a NAT gateway, restricted egress, and an allowlist of approved external endpoints for package registries, model APIs, and identity providers. A cloud service that has cleared FedRAMP certification, of which the FedRAMP Marketplace maintains a searchable public record of certified providers, authorizing agencies, and accredited assessors, typically sits in this tier rather than the stricter air-gapped one (Source: www.fedramp.gov).

Sovereign and government cloud regions raise the bar with dedicated infrastructure, restricted personnel, and compliance overlays, but many still permit some form of managed connectivity. Microsoft's Azure Government Secret and Azure Government Top Secret regions, for example, provide secure native connections to classified networks with options for high-bandwidth private connectivity such as ExpressRoute (Source: learn.microsoft.com), while the underlying enclave itself is treated as a closed SIPRNet environment.

True air-gapped deployments sit at the strict end: no network connection to the outside at all, or only a one-way data diode for telemetry export, with every dependency pre-staged inside the enclave before the system goes live (Source: www.truefoundry.com).

The DoD Impact Level and FedRAMP frameworks

For U.S. government and defense customers, two overlapping but distinct classification frameworks determine which posture an AI deployment must meet. FedRAMP (the Federal Risk and Authorization Management Program) categorizes cloud systems into Low, Moderate, and High baselines using FIPS 199 impact categorization, independent of the separate DoD Impact Level system (Source: www.schellman.com). The Department of Defense's Cloud Computing Security Requirements Guide (SRG) then layers its own Impact Levels (IL2, IL4, IL5, and IL6) on top for DoD-specific workloads, with IL6 requiring a dedicated, physically separate cloud infrastructure connected only to SIPRNet and staffed exclusively by U.S. citizens with SECRET clearances (Source: learn.microsoft.com). FedRAMP authorization is generally a prerequisite for serving DoD workloads, but the two frameworks are not interchangeable (Source: www.schellman.com). A cloud service offering carrying an IL5 designation covers higher-sensitivity CUI and mission-critical National Security Systems information, one level below the classified overlay that defines IL6 (Source: www.secondfront.com). Federal integrators recommend classifying data early, using FIPS 199 and CNSSI 1253 to determine how sensitive a system is before any architecture decisions are locked in, since retrofitting a lower-tier design to a higher Impact Level after the fact is far more expensive than designing to the correct level from the start (Source: www.secondfront.com). Even before a Mission Owner grants a full Authorization to Operate, a formal step required starting at IL4, vendors can make meaningful accreditation progress by aligning their environment to the target Impact Level from the outset (Source: www.secondfront.com).

Table 1 below summarizes the classification levels most relevant to LLM deployments, from public-facing systems to classified national security systems.

LEVEL	FRAMEWORK	DATA SENSITIVITY	NETWORK / HOSTING REQUIREMENT
FedRAMP Moderate	FedRAMP	Sensitive but unclassified data, often Controlled Unclassified Information (CUI); a breach could cause serious adverse effects (Source: www.schellman.com)	Commercial or government community cloud with roughly 320 controls
FedRAMP High	FedRAMP	Highly sensitive, mission-critical federal data where a breach could be severe or catastrophic (Source: www.schellman.com)	Dedicated government community cloud with 400-plus controls
DoD IL5	DoD CC SRG	Higher-sensitivity CUI and National Security Systems (Source: www.secondfront.com)	Physically separate, dedicated multi-tenant infrastructure; NIPRNet via cloud access point
DoD IL6	DoD CC SRG	Classified information up to SECRET on National Security Systems (Source: learn.microsoft.com)	Closed, self-contained SIPRNet enclave; U.S.-citizen, cleared personnel only
AWS / Azure Top Secret regions	Agency-specific accreditation	Top Secret national security information	Fully air-gapped commercial cloud region, no connectivity to unclassified networks (Source: aws.amazon.com)

The table illustrates a progression rather than a strict hierarchy: an organization moving from FedRAMP Moderate toward DoD IL6 is not simply adding controls, it is changing the fundamental question an auditor asks from "what is allowlisted" to "prove nothing leaves." For most commercial or civilian-agency LLM deployments, FedRAMP Moderate or High with a hardened egress allowlist is sufficient and considerably cheaper to operate than a true air gap. True air-gapped architecture becomes necessary specifically at IL6, in SCIF-hosted (Sensitive Compartmented Information Facility) intelligence community environments, and in comparable non-defense settings such as ITAR-controlled nuclear research, where any outbound dependency is disqualifying regardless of how well it is monitored.

Open-Weight Models for Disconnected Networks

The single most consequential decision in an air-gapped LLM program is model selection, because unlike a cloud deployment, the choice cannot be revisited with a configuration change. A frontier closed-weight model accessed only through a vendor API is architecturally incompatible with a true air gap; the model has to live inside the enclave as a set of downloadable weights the organization can hash, sign, and serve locally. That single requirement has made open-weight models the default building block for disconnected government, defense, and regulated-industry deployments.

The major open-weight families

Meta's Llama family anchors most air-gapped catalogs. Llama 3 shipped as pretrained and instruction-fine-tuned models with 8B and 70B parameters intended to support a broad range of self-hosted use cases (Source: ai.meta.com). Meta's architectural choices in that release, including adopting grouped query attention across both the 8B and 70B sizes and training on sequences of 8,192 tokens, were specifically aimed at improving inference efficiency, a detail that matters directly for enclaves sizing GPU memory against a fixed hardware budget (Source: ai.meta.com). Subsequent Llama 3.1 and 3.3 releases extended the family to a 405B-parameter flagship. Federal deployment guidance identifies **Llama 3.1 and 3.3 (8B, 70B, and 405B), Mixtral 8x7B and 8x22B, Qwen 2.5 (7B through 72B), and DeepSeek V3 derivatives** as the strongest options ready for on-premises federal deployment as of 2026, with Llama 3.1 70B or Mixtral 8x22B as the practical baseline for most mission workloads (Source: precisionfederal.com). Larger-context deployments increasingly reach further up the stack: teams sizing a general chat-plus-retrieval workload for H100 clusters frequently start from Llama 4 Maverick, Qwen3 235B-A22B, or gpt-oss-120b, stepping down to Llama 4 Scout or Qwen3 32B for smaller enclaves (Source: neuralchainai.com). Meta's own training infrastructure for the largest Llama 3 models achieved a compute utilization of over 400 TFLOPS per GPU when trained on 16,000 GPUs simultaneously, an indication of the scale gap between how frontier open-weight models are trained and how modestly they can later be served inside a single enclave (Source: ai.meta.com).

Mistral AI's open models are released predominantly under an Apache 2.0 license, which permits use for any purpose, distribution, and modification without royalty obligations (Source: help.mistral.ai); some newer models carry a modified MIT license that offers the same permissions as Apache 2.0 but requires companies with monthly revenue exceeding $20 million to obtain a commercial license or use Mistral's hosted platform (Source: help.mistral.ai). Certain models are governed by this modified MIT license specifically, giving classified programs a predictable compliance path to reason about as their internal usage scales (Source: help.mistral.ai). **Alibaba's Qwen** models, including the widely deployed Qwen 2.5 series, are released under the Apache License, Version 2.0 (Source: github.com), removing most of the licensing friction that complicates commercial and government redistribution. **DeepSeek-V3**, notable for strong benchmark performance at a fraction of the training compute of comparable frontier models, ships its code under the MIT license, and the DeepSeek-V3 model series, including both Base and Chat variants, explicitly supports commercial use (Source: github.com). **Ollama**, a widely used local runtime, packages many of these same open-weight families for single-command deployment and exposes a REST API for running and managing models locally, which is one reason it shows up so often in smaller air-gapped and edge-of-enclave installs (Source: github.com).

Purpose-built defense and government variants

Beyond the general-purpose open-weight catalog, several organizations have built fine-tuned derivatives specifically for classified or defense use. **Defense Llama**, released by Scale AI, is built on Meta's Llama 3 and specifically customized and fine-tuned to support American national security missions, trained on military doctrine, international humanitarian law, and policies aligned with DoD guidelines for armed conflict, and made available exclusively in controlled U.S. government environments (Source: scale.com). Its outputs are additionally tuned to adhere to the Office of the Director of National Intelligence's style and tone of writing to increase relevance for national security users (Source: scale.com). Scale built the model using fine-tuned data from its own Data Engine to configure Defense Llama's parameters for defense-specific scenarios, and its training corpus was designed to align with the DoD's Ethical Principles for Artificial Intelligence so that outputs stay consistent with the department's stated rules for the use of AI in armed conflict (Source: scale.com). Anthropic's **Claude Gov** models were built based on direct feedback from government customers and underwent the same rigorous safety testing as Anthropic's commercial Claude models, with the notable behavioral difference that they refuse less when engaging with classified information the operator is authorized to handle (Source: www.anthropic.com). Federal AI integrators report a working catalog for classified enclaves that spans Llama 3.1 across its 8B, 70B, and 405B sizes, Mistral Large, Mixtral 8x22B, Qwen 2.5, DeepSeek-V3, and Phi-4, with every weight imported via accredited one-way transfer and SHA-256 verified before it is registered and served (Source: precisionfederal.com).

Table 2 compares the major open-weight families most commonly deployed in air-gapped enclaves as of mid-2026.

MODEL FAMILY	REPRESENTATIVE SIZES	LICENSE	TYPICAL MINIMUM GPU MEMORY	COMMON SERVING STACK
Llama 3.1 / 3.3	8B, 70B, 405B	Meta community license (self-hostable)	~18 GB (8B) to ~810 GB (405B, bf16) (Source: precisionfederal.com)	vLLM, TensorRT-LLM
Mixtral 8x22B	141B total, 39B active (MoE)	Apache 2.0	~80 GB quantized to ~280 GB (bf16)	vLLM, TensorRT-LLM
Qwen 2.5 / Qwen3	7B to 235B (some MoE)	Apache 2.0 (Source: github.com)	Varies widely by size and MoE activation	vLLM, SGLang, Ollama
DeepSeek-V3 / V3.2	671B total, MoE architecture	MIT (code and model) (Source: github.com)	Multi-node H100/H200 cluster for full precision	SGLang, TensorRT-LLM, vLLM
Defense Llama	Llama 3 derivative	Restricted, government-environment only (Source: scale.com)	Comparable to base Llama 3 size class	Scale Donovan platform

The practical implication of this table is that license terms, not raw capability, often decide which model an air-gapped program can legally redistribute and modify without a commercial agreement. Apache 2.0 and MIT-licensed families (Qwen, most Mistral models, DeepSeek) impose the fewest downstream restrictions, which matters considerably when a defense integrator needs to fine-tune, quantize, and redistribute a modified model inside a closed ecosystem where no external legal review is easily obtainable mid-mission. Production teams still verify license terms against their specific fine-tuning and redistribution plans before locking in a base model, since a fine-tune of a fine-tune of a fine-tune can inadvertently inherit a restrictive clause from a base model several generations back that nobody on the current team explicitly agreed to (Source: tianpan.co).

Air-Gapped GPU Cluster Architecture and Hardware Requirements

Once a model is selected, hardware sizing follows directly from parameter count, precision, and expected concurrency. Because air-gapped environments cannot burst into a public cloud during demand spikes, capacity has to be provisioned for peak load from day one, which makes hardware sizing one of the largest cost drivers in the entire program.

GPU memory and throughput planning

Federal on-premises deployment guidance for 2026 hardware lays out concrete sizing targets. A **Llama 3.1 8B** model in bf16 precision needs roughly 18 GB of GPU memory and runs on a single L40S or H100 at 2,000 to 4,000 tokens per second in batched serving. A **Llama 3.1 70B** model in bf16 needs roughly 140 GB, typically served on four A100 80GB GPUs or two H100s, while the same model quantized to 4-bit AWQ needs only about 40 GB and runs on a single H100 or two L40S GPUs. The largest **Llama 3.1 405B** model in bf16 needs approximately 810 GB, requiring eight H100s or four H200s. AWQ and GPTQ 4-bit formats generally give three to four times memory reduction with modest quality loss on most tasks, while FP8 on H100 or H200 hardware approaches BF16 quality for many workloads. At the extreme budget end, hobbyist and small-enterprise reference builds have shown that even 2018-era Turing GPUs, specifically a pair of Quadro RTX 6000 cards with only 24 gigabytes of memory each behind a Dell PowerEdge R740 chassis, can serve a 30-billion-parameter mixture-of-experts coding model at 80 to 120 tokens per second once tensor parallelism and 4-bit quantization are applied (Source: github.com), and that reference architecture is itself released under an MIT license, illustrating that air-gapped serving spans a spectrum from budget hardware to full DGX clusters rather than a single fixed bill of materials (Source: github.com). That same reference build reports zero cloud dependencies, no telemetry, and no external API calls anywhere in its stack, the same criteria this report uses to distinguish a true air gap from a merely isolated deployment (Source: github.com).

The NVIDIA H200 is the current workhorse for this class of deployment. It is the first GPU to offer 141 gigabytes of HBM3e memory at 4.8 terabytes per second of bandwidth, nearly double the memory capacity of the H100 with 1.4 times the memory bandwidth (Source: www.nvidia.com). A single H200 costs between roughly $30,000 and $40,000 to purchase outright as of January 2026 (Source: jarvislabs.ai). For smaller enclaves or forward-deployed edge nodes that do not need a rack-scale cluster, NVIDIA also offers desktop-class options: the DGX Spark pairs a GB10 Superchip with 128 GB of unified system memory to let researchers work with models up to 200 billion parameters locally, while the DGX Station scales further to 748

GB of coherent memory in a tower form factor for larger single-node inference and fine-tuning workloads (Source: www.nvidia.com). NVIDIA positions the H200's added memory and bandwidth as delivering better energy efficiency and lower total cost of ownership for large language model workloads relative to the H100 (Source: www.nvidia.com), and it markets the DGX Station specifically as the ultimate desktop AI supercomputer for teams that need workstation-class hardware without a full rack deployment (Source: www.nvidia.com).

System-level pricing and total cost

Table 3 summarizes representative on-premises GPU hardware pricing relevant to air-gapped cluster planning.

HARDWARE CONFIGURATION	MEMORY	APPROXIMATE PRICE	TYPICAL DEPLOYMENT FIT
Single NVIDIA H100 (80 GB)	80 GB HBM3	$25,000 to $40,000 per GPU (Source: jarvislabs.ai)	8B to 70B (quantized) single-model serving
Single NVIDIA H200 (141 GB)	141 GB HBM3e	$30,000 to $40,000 per GPU (Source: jarvislabs.ai)	70B-class model on a single card
8-GPU DGX H200 system	8 x 141 GB HBM3e (Source: www.broadberry.com)	Approximately $467,000 configured, up to $584,000 at list price (Source: www.broadberry.com)	405B-class model, high-concurrency serving
Site-level redundant cluster (4 to 8 H100s)	320 GB to 640 GB aggregate	$400,000 to $800,000 including networking, storage, and racks (Source: precisionfederal.com)	Enclave-wide production serving with failover

The table shows that GPU cost scales roughly linearly with memory capacity at the card level but non-linearly at the system level once networking, redundant power, cooling, and enterprise support contracts are added: a single H200 at roughly $35,000 becomes a $467,000 system once assembled into an 8-GPU DGX chassis, a multiple that reflects NVLink interconnect, dual Xeon Platinum processors, and 2 terabytes of DDR5 system memory (Source: www.broadberry.com), which the vendor markets as delivering up to 45 percent more inference performance and a 50 percent reduction in energy use for LLM workloads compared with the prior H100 generation (Source: www.broadberry.com). Classified enclaves often standardize instead on NVIDIA DGX, HPE Cray, or Supermicro GPU nodes hosted on accredited infrastructure or SCIF-resident compute, running Kubernetes on a hardened control plane, with availability and waitlists varying by facility.

Export controls shape hardware access

Hardware acquisition for air-gapped clusters is also constrained by U.S. export control policy on advanced computing chips. Since the October 2023 update to Bureau of Industry and Security rules, controlled chips are identified using three criteria: total processing performance, performance density, and whether the chip is designed or marketed for datacenter use (Source: cset.georgetown.edu). Total Processing Performance measures the number of operations a chip can perform per second, while Performance Density measures a chip's compute power per unit of die area and serves as a proxy for how advanced its underlying manufacturing process is (Source: cset.georgetown.edu). Chip designers, including NVIDIA, had previously developed chips just below the earlier control thresholds to keep selling into constrained markets, prompting regulators to fold these two additional tests into the definition specifically to close that gap (Source: cset.georgetown.edu). In addition to a regular licensing requirement for the most advanced datacenter chips, BIS created a license exception for less-advanced datacenter chips, giving certain destinations an expedited or unrestricted path depending on the chip's exact performance tier (Source: cset.georgetown.edu). BIS also notably removed the interconnect speed parameter, the rate at which chips talk to each other, from the criteria used in the original 2022 rule, closing another avenue chip designers had used to keep exported parts under the threshold (Source: cset.georgetown.edu). The updated criteria were designed specifically to limit the ability of restricted destinations to use large-scale AI systems for advanced weapons applications by restricting access to the chips capable of training frontier models. While these controls target exports to specific countries rather than domestic classified deployments, they shape which GPU generations are available for allied and partner-nation air-gapped programs and reinforce why U.S. government facilities generally standardize on the highest-tier NVIDIA datacenter parts they can procure domestically rather than lower-spec export variants.

Software Stack: Serving, Quantization, and Retrieval Inside the Enclave

Hardware and model selection only enable a deployment; the serving stack determines whether it actually runs disconnected. Every component that a cloud-native AI stack takes for granted, package installation, model downloads, telemetry, and certificate renewal, has to be re-engineered to function with zero outbound calls.

Inference serving frameworks

vLLM, an open-source, Apache 2.0-licensed Python framework, has become the default choice for on-premises serving in 2026. It implements PagedAttention for efficient key-value cache management, continuous batching, speculative decoding, and broad model compatibility across NVIDIA and AMD GPUs, exposing an OpenAI-compatible HTTP API along with tensor and pipeline parallelism and native AWQ, GPTQ, and FP8 quantization support (Source: precisionfederal.com). **TensorRT-LLM**, NVIDIA's optimized inference engine, delivers 20 to 40 percent higher throughput than vLLM in many configurations for steady-state production workloads, at the cost of a per-model engine build step that can take 30 to 120 minutes and a tighter dependency on matching CUDA and driver versions. Classified engagements commonly standardize on a wider runtime bench that also includes llama.cpp for fully offline use, plus NVIDIA Triton when GPU utilization efficiency matters more than raw throughput.

NVIDIA NIM microservices explicitly support this pattern at the platform level. In an air-gapped system, an operator can run a NIM microservice with no internet connection and no connection to the NGC registry or Hugging Face Hub, using either a proxy configuration or a mirrored local model registry, the two supported patterns for reaching a model catalog without direct internet access (Source: docs.nvidia.com). **Ollama**, a lighter-weight open-source runtime, is popular in smaller and edge-of-enclave deployments where teams want a simpler operational footprint than a full Kubernetes-based vLLM cluster, exposing the same kind of local REST API for chat and management operations rather than a hosted endpoint (Source: github.com). Fine-tuning inside the enclave typically relies on LoRA and QLoRA adapters, with full-parameter supervised fine-tuning reserved for DGX-class hardware, all logged to storage inside the enclave rather than a cloud experiment tracker.

Retrieval-augmented generation without leaving the perimeter

Most production air-gapped deployments pair the LLM with a retrieval-augmented generation (RAG) pipeline over an internal document corpus, and this is where architectures most often break the air gap in practice, not through the LLM itself but through the embedding model. Sending text to a remote embedding API, even one operated by a major cloud provider, is the single most common way teams discover their supposedly air-gapped retrieval pipeline is quietly leaking data, typically noticed only when network monitoring shows outbound DNS queries originating from the retrieval component of the stack because an SDK defaulted to a cloud provider endpoint (Source: www.truefoundry.com). The fix is to run a smaller open-source embedding model, commonly a 384 or 768-dimensional model such as bge-small or nomic-embed, on the same GPUs as the serving workers or on CPU, with the vector index encrypted at rest and accessible only from inside the enclave, using self-hosted vector stores with classification labels attached at the chunk level so retrieval respects clearance boundaries automatically.

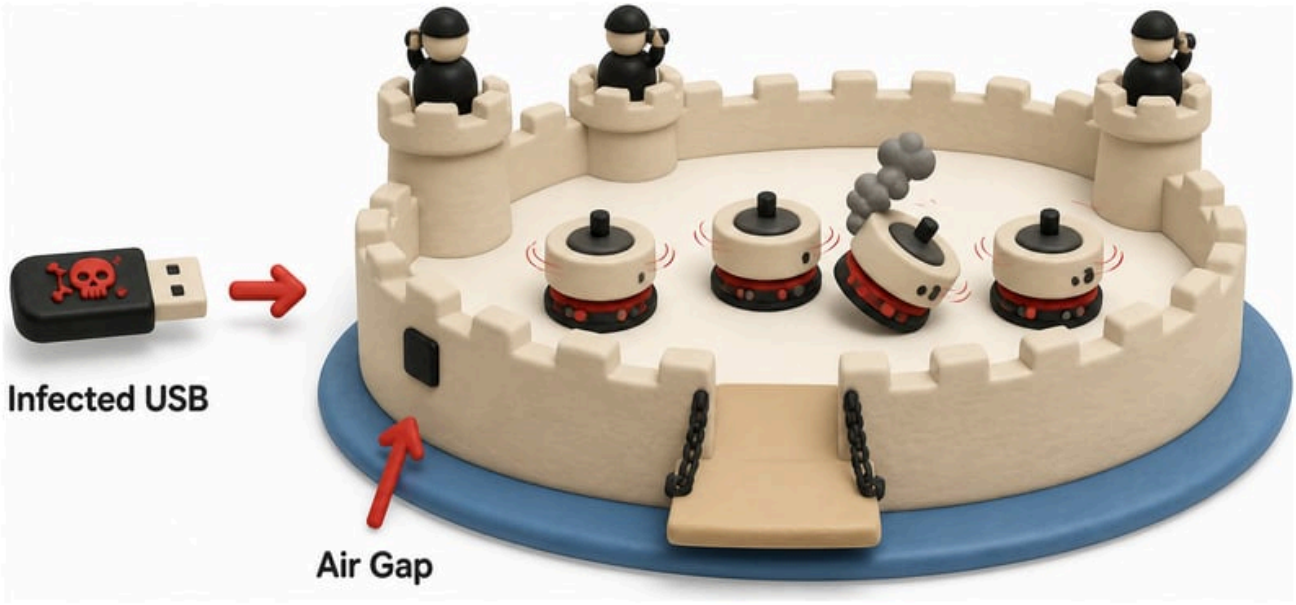
Everything else has to live inside the wall too

A complete air-gapped stack needs an internal certificate authority, since public certificate authorities such as Let's Encrypt and DigiCert are simply not reachable from inside the enclave (Source: www.truefoundry.com); an internal package and container registry mirror, since a deployment manifest that still references a public registry tag will fail at runtime in an air-gapped environment; and internal DNS and NTP servers, because default configurations that quietly reach public time-synchronization pools are exactly the kind of gap that survives a casual network review. On-premises observability with Prometheus, Grafana, and Loki replaces a SaaS monitoring vendor entirely. Because the operational load of standing all of this up correctly is nontrivial, mature programs run a staged rollout: the same gateway image and Helm chart that will eventually run in the enclave first runs in a connected sandbox with full egress logging, so any unexpected outbound connection is caught before the system moves into the regulated environment.

Security Architecture: Cross-Domain Solutions and the Update/Patching Problem

An air-gapped network is, by design, resistant to remote attack, but it is not automatically secure, and its threat model differs sharply from a connected system's. Security researchers describe the intuitive appeal directly: disconnecting a system from the internet, surrounding it with guards and gates, is supposed to make it hack-proof, but in reality it is not (Source: www.f5.com). The canonical illustration remains Stuxnet, the malicious worm

discovered in 2010 that targeted the programmable logic controllers running Iran's uranium enrichment centrifuges at the Natanz facility. The worm was first identified by the security company VirusBlokAda in mid-June 2010, months after it is believed to have begun operating inside the Natanz facility (Source: en.wikipedia.org). Stuxnet was typically introduced to the target environment via an infected USB flash drive (Source: en.wikipedia.org) (Source: www.f5.com), and once inside the isolated industrial network it caused the number of enrichment centrifuges operational in Iran to mysteriously decline from about 4,700 to about 3,900 over a period of months (Source: en.wikipedia.org). At its broadest, the worm eventually infected over 200,000 computers worldwide and caused roughly 1,000 machines to physically degrade, a scale that illustrates how far a single air-gap breach can propagate once it escapes its original target (Source: en.wikipedia.org). Beyond direct infection, attackers can also pre-penetrate a target before it ever reaches the air gap by sabotaging its supply chain, hiding malware in hardware or software components long before the equipment is installed and accredited (Source: www.f5.com). Insider physical transfer poses an equally serious risk on the exfiltration side: in one widely documented case, a U.S. Army intelligence analyst physically carried gigabytes of classified files out of a Sensitive Compartmented Information Facility (SCIF) on a CD, the same kind of removable-media pathway that today legitimately carries signed model updates into a classified enclave (Source: www.f5.com). The lesson for AI programs is not abstract: an air gap defends against network-borne compromise, but every controlled physical transfer, whether removable media carrying model weights or a technician's laptop, is a potential vector unless it is subject to the same rigor as a network boundary.



Cross-domain solutions and one-way transfer

Formal defense and intelligence programs manage this risk with accredited **cross-domain solutions (CDS)**, a category that includes hardware data diodes engineered to physically enforce one-directional data flow between networks of different classification levels. Federal AI delivery teams that handle multi-level security typically integrate with accredited cross-domain solution vendors such as Forcepoint, Owl, or Arbit for these cross-enclave flows rather than building a custom guard from scratch (Source: precisionfederal.com). In LLM programs, this hardware sits at the boundary the model weights, container images, and telemetry cross: telemetry can leave a classified enclave over a diode for monitoring, but nothing new enters the enclave except through the accredited import process on the other side. Cloud regions built for this tier of assurance are assessed against the same government standards: AWS's Secret Region, for example, is assessed and accredited under the Director of National Intelligence's Intelligence Community Directive 503 and NIST Special Publication 800-53 Revision 4 (Source: aws.amazon.com).

The update pipeline is the hardest engineering problem

Model bumps are trivial in a cloud deployment, a configuration change that swaps an API endpoint. Inside an air gap, a model bump is an artifact transfer that has to clear a signed-bundle release process, a security re-review, and a chain-of-custody log, and model weights themselves can range from tens to hundreds of gigabytes, which have to move on encrypted media or through a one-way data diode (Source: tianpan.co). Because updates

are not idempotent inside an enclave the way a cloud config change is, the discipline that has emerged across production programs is a signed bundle workflow: a vendor or internal release-engineering team publishes a release as a signed tarball, containing manifests, container images, a Helm chart, and any weight artifacts, all under an integrity hash. That bundle is loaded onto physical media, typically a USB stick or removable hard drive depending on the regime's media policy, and walked across the air gap to a holding server inside the enclave (Source: neuralchainai.com). On the high side, the process re-computes every hash, validates the signature against pre-loaded public keys, and runs an evaluation suite before promoting the new version, following the standard federal pattern of downloading the model on the low side, hashing it, scanning for malware, having it signed by an authorizing official, then burning it to approved removable media or pushing it through a cross-domain solution for verification on the high side. Cadence for this cycle is typically quarterly rather than weekly, a deliberate tradeoff between staying current and keeping the accreditation burden manageable (Source: neuralchainai.com).

Supply-chain integrity for model weights

Open-weight model repositories have become an attack surface in their own right. There is still no widely adopted mechanism for cryptographically signing model weights and verifying that signature at load time by default, so mature programs build that gate themselves, hash-pinning every model, signing the artifact against an internal key during ingestion, and refusing to load anything whose signature does not verify (Source: tianpan.co). Open-weight models are generally treated as unclassified on import, but any fine-tune trained on classified data becomes classified at the derivation level, with the model registry itself enforcing that reclassification. An AI/ML bill of materials, increasingly published in CycloneDX format, extends this discipline beyond the base model to track the tokenizer, safety classifier, LoRA adapters, merged checkpoints, quantization tooling, and evaluation suite that gated the release, since a fine-tune of a fine-tune can inadvertently drag in a license clause from a base model the team never explicitly agreed to.

Data Analysis and Evidence

The quantitative picture around air-gapped and government AI deployment shows a market growing far faster than the broader AI economy, concentrated overwhelmingly in defense spending, and increasingly anchored to open-weight models rather than proprietary API access.

Federal AI spending growth

Worldwide AI spending is forecast to total $2.52 trillion in 2026, a 44 percent year-over-year increase according to Gartner (Source: www.ibm.com), but the U.S. federal segment has grown even faster off a smaller base. Obligated federal AI spending increased from $675 million in 2024 to $7.2 billion in 2026, a 966 percent increase, while potential contract value rose from $4.6 billion to $91.8 billion over the same period, a 1,912 percent increase (Source: www.ibm.com). This spending concentrates heavily in two North American Industry Classification System categories, which together account for 98 percent of all federal AI spending (Source: www.ibm.com), indicating that the market remains focused on technical and advisory services rather than broad-based deployment across every federal function.

The most consequential concentration, however, is by agency. In 2026, the Department of Defense accounted for 1,319 of the 1,743 federal AI contracts and 98.9 percent of all potential contract value, equal to $90.7 billion and a 1,605 percent increase from 2024 (Source: www.ibm.com). Other agencies, including Health and Human Services, NASA, and Commerce, saw only modest activity by comparison, which underscores why air-gapped LLM deployment is, in practical terms, primarily a defense and intelligence community discipline rather than a broad civilian-agency one, even as the underlying techniques increasingly transfer to healthcare and financial-services use cases.

Adoption at scale inside the DoD

Beyond dollar figures, adoption numbers within specific programs illustrate how quickly air-gapped and semi-air-gapped LLM tools have scaled inside the Army specifically. CamoGPT, the Army's internally hosted chatbot built on open-source models from vendors including Meta, Google, and Mistral, grew from a fall 2023 development start to 75,000 users by the time of reporting, having first deployed in spring 2024 (Source: www.army.mil). The Army's AI Integration Center subsequently built a broader Army Enterprise Large Language Model Workspace that gives personnel access to more than 23 approved industry AI frontier models, while retaining CamoGPT as a specialized research and development platform for testing agentic AI capabilities before they transition to production (Source: defensescoop.com).

Hardware and quantization economics

On the hardware side, the throughput and memory tradeoffs quantified in the previous section translate directly into deployment cost. Quantizing a Llama 3.1 70B model from bf16 to 4-bit AWQ reduces the memory footprint from roughly 140 GB to roughly 40 GB, a 3.5x reduction that moves the model from a 4-GPU requirement to a single-GPU deployment, at the cost of modest quality loss on most tasks but a measurable collapse on narrow structured-extraction tasks that aggregate evaluation suites can miss entirely. This is one reason mature air-gapped programs run task-level evaluation, not just aggregate benchmark scores, before promoting a quantized model to a mission-critical role, since an aggregate score can look acceptable while masking a specific failure mode that only appears in a narrow but operationally important task category. Federal AI teams also increasingly run adversarial red-team evaluation as a distinct step, using open-source probes such as Garak and PyRIT alongside general-purpose evaluation harnesses and custom mission-specific evals, with findings routed to the enclave's information system security manager before a model reaches production (Source: precisionfederal.com).

Case Studies and Real-World Examples

The following named deployments illustrate how the frameworks, hardware, and software patterns described above have actually been fielded across defense, intelligence, and national laboratory environments as of 2026.

Microsoft Azure Government Top Secret Cloud and GPT-4

In May 2024, Microsoft announced it had deployed the GPT-4 large language model into an isolated, air-gapped Azure Government Top Secret cloud for use by the Department of Defense, pending accreditation for operational use (Source: defensescoop.com). William Chappell, Microsoft's chief technology officer for strategic missions and technologies, described the shift plainly: connected systems require constant defense against probing, and an isolated and air-gapped environment now brought the same commercial-grade AI capability available on the unclassified side into a Top Secret environment for the first time (Source: defensescoop.com). This deployment used Azure OpenAI Service specifically inside the Azure Government Secret and Top Secret regions, the same environments carrying Microsoft's DoD Impact Level 6 provisional authorization, the highest classified cloud service offering authorization Microsoft has attained (Source: learn.microsoft.com).

Palantir and Microsoft's classified analytics partnership

Building on that infrastructure, Palantir Technologies and Microsoft announced in August 2024 an expanded partnership to bring Palantir's Foundry, Gotham, Apollo, and AIP products into Microsoft's classified cloud environments, with Palantir deploying into Azure Government Secret (DoD Impact Level 6) and Top Secret clouds (Source: news.microsoft.com). Palantir was positioned as the first industry partner to deploy Microsoft's Azure OpenAI Service into classified environments (Source: news.microsoft.com). Palantir's chief technology officer framed the announcement in similar terms, saying Palantir AIP has pioneered the approach to operationalizing AI value across the enterprise (Source: news.microsoft.com), and Microsoft's public sector leadership added that Palantir, a leader in delivering actionable insights to government, would now leverage Microsoft's government and classified clouds to further develop AI innovations for national security missions (Source: news.microsoft.com), integrating Palantir's data ontology and AI Platform (AIP) capabilities with Microsoft's cloud compute and language models to build operational workloads spanning logistics, contracting, and prioritization for defense and intelligence missions.

Los Alamos National Laboratory's Venado supercomputer

Los Alamos National Laboratory installed its Venado supercomputer beginning in April 2024, an installation the laboratory describes as centered on NVIDIA's proprietary Grace Hopper processing technology and associated Hewlett Packard Enterprise hardware, explicitly to introduce AI-powered computation into scientific modeling and simulation alongside the lab's existing high-performance computing portfolio (Source: www.nextgov.com). HPC program manager Jim Lujan described the acquisition as a statement of ambition, saying the laboratory wanted to make a significant statement in that it is leading in computing as a national laboratory (Source: www.nextgov.com). Laboratory staff described the acquisition explicitly as a bridge to classified work: part of the acquisition goal was to eventually cleave off a portion of the system and transition it into a classified environment to support the laboratory's nuclear stockpile stewardship mission alongside the Department of Energy's National Nuclear Security Administration, which broadly characterizes research within any discipline as having the potential to be deemed classified whenever it relates to a laboratory's national security mission (Source: www.nextgov.com). As one Los Alamos scientist put it, unclassified does not mean that a project lacks national security relevance or importance, a reminder of how thin the line can run between a laboratory's open and classified AI work (Source: www.nextgov.com).

Separately, industry analysis indicates Los Alamos moved its broader LLM stack fully on-premises in early 2025 specifically to handle Controlled Unclassified Information, ITAR-flagged data, and Unclassified Controlled Nuclear Information, categories of sensitive but unclassified data that cannot legally transit a vendor's cloud inference endpoint (Source: tianpan.co).

The U.S. Army's CamoGPT and Enterprise LLM Workspace

Sources describe the program's timeline with slightly different emphasis: the Army's own reporting places CamoGPT's development start in the fall of 2023 with a spring 2024 first deployment (Source: www.army.mil), while later trade coverage describes the Army's AI Integration Center (AI2C) as having begun prototyping CamoGPT in 2024 (Source: defensescoop.com). Either way, the platform is designed to reliably support administrative and operational tasks across a range of use cases while maintaining security for Controlled Unclassified Information and certain classified information (Source: defensescoop.com). Notably, the small AI2C engineering team does not train its own foundation models; a data engineer on the team described the approach directly: the team simply hosts published, open-source models that have been trained by companies like Meta, Google, and Mistral (Source: www.army.mil), a distillation of the entire open-weight-plus-air-gap strategy this report describes. CamoGPT can process classified data on SIPRNet and unclassified information on NIPRNet (Source: www.army.mil), demonstrating a single program operating across two distinct classification enclaves with separate model deployments in each.

Anthropic's Claude Gov and the AWS Secret and Top Secret Regions

Anthropic's Claude Gov models represent a proprietary-model counterpoint to the open-weight pattern dominant elsewhere in this report: rather than distributing weights for local hosting, Anthropic operates the models within accredited government cloud environments, with the models already deployed by agencies at the highest level of U.S. national security and access restricted to operators working in classified environments (Source: www.anthropic.com). Beyond classified-material handling, Anthropic describes Claude Gov as offering enhanced proficiency in languages and dialects critical to national security operations and improved interpretation of complex cybersecurity data for intelligence analysis (Source: www.anthropic.com). This deployment model runs on infrastructure comparable to Amazon's AWS Secret Region, which the company describes as extending the same tools and workflows already available for Top Secret workloads to customers with Secret-level datasets (Source: aws.amazon.com), built on the earlier AWS Top Secret Region that Amazon describes as the first air-gapped commercial cloud when it launched roughly three years prior (Source: aws.amazon.com).

Implications and Future Directions

Several trends are shaping how air-gapped LLM deployment will evolve over the next several years. First, the licensing landscape for open-weight models continues to consolidate around Apache 2.0 and MIT terms of the kind DeepSeek and Qwen already use (Source: github.com), which is quietly becoming the deciding factor in which models classified programs standardize on, since a restrictively licensed model that cannot be freely fine-tuned and redistributed inside a closed ecosystem creates legal risk that most defense integrators are unwilling to accept mid-mission. Meta itself has framed this openness as a deliberate strategy, stating that an open approach is intended to kickstart the next wave of AI innovation across the stack from applications to developer tools to inference optimizations (Source: ai.meta.com). Programs increasingly treat license compatibility as a gating criterion on par with raw benchmark performance when selecting a base model.

Second, the update pipeline problem described in this report, the fact that model bumps require a signed-bundle release process rather than a configuration change, is likely to remain the primary operational bottleneck for years rather than a temporary inconvenience solved by better tooling. As frontier open-weight models continue to release new checkpoints on a roughly quarterly cadence, and as classified programs settle into equally quarterly update cycles for their own accreditation reasons, the two timelines are converging toward a stable rhythm, but the underlying friction of moving tens to hundreds of gigabytes of weights across a controlled boundary will not disappear even as diode bandwidth and automation improve.

Third, hardware economics will keep pushing toward quantization and smaller, more efficient models rather than ever-larger dense architectures, particularly inside classified enclaves where every GPU has to be procured, powered, and cooled inside a facility with fixed physical constraints rather than rented elastically from a hyperscaler. Mixture-of-experts architectures such as Mixtral, which activates only a fraction of its total parameters per forward pass, are likely to gain further share specifically because they map well onto memory-constrained, power-constrained on-premises hardware, including legacy GPU generations that predate modern data formats but can still serve production traffic when paired with an efficient MoE model and modern inference kernels.

Fourth, the boundary between "air-gapped" and merely "isolated" deployment is likely to sharpen rather than blur as more organizations outside pure defense contexts, healthcare systems handling protected health information and financial institutions under stricter custody regimes, adopt the same architectural discipline that classified programs pioneered, mapping onto frameworks such as CMMC, HIPAA with HITRUST attestation, and financial-

sector rules like SR 11-7 and NYDFS Part 500. As those industries build their own accreditation frameworks around AI-specific risk, the vocabulary and tooling developed for DoD IL6 and FedRAMP High enclaves, signed model bundles, internal certificate authorities, local embedding models, and staged egress-monitoring sandboxes, will likely propagate into commercial regulated-industry deployments largely unchanged, since the underlying engineering problem, proving that nothing leaves the perimeter, is identical regardless of which regulator is asking.

Frequently Asked Questions (FAQs)

How do you run AI models without internet access? The model weights, tokenizer, and serving software are downloaded once on a connected "low side" system, hashed and scanned for malware, signed by an authorizing official, and transferred across the boundary on approved physical media or through a one-way data diode. Inside the disconnected network, an inference server such as vLLM, TensorRT-LLM, or NVIDIA NIM loads the local weights and serves requests with zero outbound network calls at runtime (Source: docs.nvidia.com).

What are the best open-weight models for offline inference? Federal deployment guidance for 2026 points to Llama 3.1 and 3.3 (8B, 70B, and 405B), Mixtral 8x7B and 8x22B, Qwen 2.5, and DeepSeek V3 derivatives as the strongest ready-to-deploy options, with Llama 3.1 70B or Mixtral 8x22B serving as a practical baseline for most mission workloads.

What security requirements apply to on-premises LLM deployment? Requirements depend heavily on data classification. Unclassified but sensitive federal data typically requires FedRAMP Moderate or High compliance, CUI and non-classified National Security Systems generally require DoD IL4 or IL5 (Source: www.secondfront.com), and classified information up to SECRET requires DoD IL6 with a dedicated SIPRNet-connected enclave staffed by cleared U.S. citizens. Non-defense regulated sectors typically map to comparable frameworks such as CMMC Level 2 or 3, HIPAA with HITRUST attestation, or financial-sector rules like SR 11-7 and NYDFS Part 500.

What GPU hardware is required for a disconnected data center AI deployment? Sizing depends on model size and precision. An 8B model needs roughly 18 GB of GPU memory and runs on a single GPU; a 70B model needs 40 GB quantized or 140 GB at full precision; a 405B model needs approximately 810 GB, typically eight H100s or four H200s. Classified enclaves typically standardize on DGX-class systems, HPE Cray, or Supermicro GPU nodes on accredited hosting, with a single H200 running roughly $30,000 to $40,000 (Source: jarvislabs.ai).

How are models updated or patched in an air-gapped environment? Through a controlled transfer pipeline rather than an automatic update: the new model is downloaded and verified on a connected system, signed, burned to approved removable media or pushed through a cross-domain solution, verified again inside the enclave, and promoted through staging before reaching production, typically on a quarterly rather than weekly cadence.

Can classified networks use commercial frontier models like GPT-4 or Claude? Yes, but only through accredited government cloud regions rather than the public API. GPT-4 has been deployed in Microsoft's air-gapped Azure Government Top Secret cloud (Source: defensescoop.com), and Anthropic offers purpose-built Claude Gov models restricted to operators in classified environments (Source: www.anthropic.com).

Conclusion

Air-gapped LLM deployment has moved from an edge case to a well-defined engineering discipline with its own hardware economics, licensing logic, and accreditation frameworks. The core architectural claim, no NAT gateway, no external DNS resolution, no trusted external certificate chain, is a strict standard that most "isolated" enterprise deployments do not actually meet, and organizations evaluating this path should be honest with themselves and their auditors about which side of that line they are actually operating on. For the classified, national-laboratory, and increasingly regulated-industry environments profiled in this report, open-weight models under permissive Apache 2.0 or MIT licenses, served through vLLM, TensorRT-LLM, or NVIDIA NIM on H100 or H200-class hardware, have become the default building block, precisely because they are the only option compatible with an environment that has no vendor API to call.

The evidence assembled here points to three durable conclusions. First, the market is real and growing fast, with U.S. federal AI spending rising 966 percent in obligated dollars between 2024 and 2026, concentrated almost entirely in defense. Second, the technology stack has matured well beyond experimental status: Microsoft, Amazon, Palantir, Scale AI, Anthropic, and the U.S. Army have all fielded production or near-production air-gapped and classified-cloud AI systems, several already serving tens of thousands of users. Third, the hardest unsolved problem is not inference performance but the update pipeline: moving signed, hashed, quarterly model releases across a controlled boundary without breaking an accreditation package remains the operational bottleneck that determines whether an air-gapped AI program succeeds or stalls. Organizations planning a deployment in this space should budget as much engineering effort for the transfer and verification pipeline as for the GPU cluster itself, since a fast inference server behind a boundary nobody can reliably update is, in practice, no better than no deployment at all.

Tags: air-gapped llm deployment, offline ai models, open-weight models, classified ai infrastructure, dod impact levels, fedramp high, on-premise llm, gpu cluster architecture, defense ai, cross-domain solutions

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will GPUSmith or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.