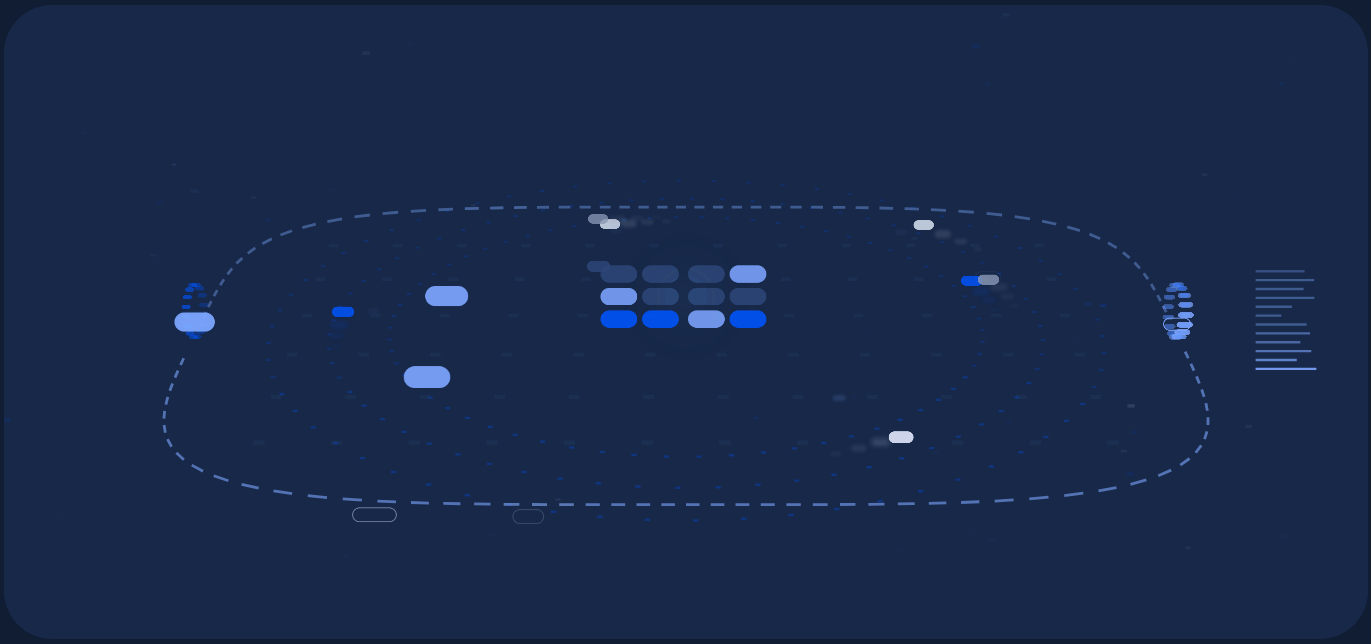
Rillor

EXPLAINER

Private LLM deployment: who controls what

On-premises LLM deployment, colocation, dedicated GPU capacity and managed APIs each place the hardware, facility, data, model weights, logs and operations with different parties.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Each way of running a large language model places the hardware, facility, data, model weights, logs and operations with a different mix of parties.

- NIST's 2011 cloud definition calls a cloud private when its infrastructure is provisioned for one organization, whether that organization, a third party or both own, manage and operate it, and whether it sits on or off premises [1]. NIST's 2012 guidance splits private clouds into on-site and outsourced variants, which differ in who builds the security perimeter [2].
- On a managed model API, the provider runs the hardware, the facility and the model, and the customer chooses among documented settings for processing location, retention and its own request logs [6, 9, 10, 13]. On Azure, the Global and DataZone deployment types change only where prompts are processed, and data at rest stays in the customer's chosen geography [6].
- By default, four managed model services may keep prompts to detect abuse: Azure, Amazon Bedrock, Google Cloud's managed model platform and one model developer's API [6, 9, 13, 14]. That developer's API keeps abuse monitoring logs for up to 30 days unless longer retention is required by law or needed to prevent harm [14]. On Amazon Bedrock, some model families require human review by AWS, and prompts and completions for those are kept for up to 30 days [9]. Keeping prompts out of the abuse monitoring store, called modified abuse monitoring, takes an approval on Azure and on that developer's API [6, 7, 14].
- On premises does not settle who owns the hardware. AWS Outposts racks sit in the customer's building, are owned and managed by AWS, and AWS operates, monitors and manages their capacity as part of an AWS Region [17].
- An organization that runs the serving software itself takes on its defaults. The documentation of vLLM, an open-source serving engine, says traffic between the nodes of a multi-node deployment is insecure by default, and the engine logs prompt inputs only when request logging is on and the log level is DEBUG [28, 29].
- Colocation and dedicated GPU capacity add parties. One colocation operator typically bills for space and power and runs an on-site service in which its technicians install and troubleshoot customer equipment [18]. One GPU cloud states that it leases its data center space and does not control the operation of those facilities [19].
- Published guidance turns the choice into questions about roles, weights, keys, logs and updates. Joint guidance from April 2024 asks for two-person control over access to model weights, encryption of weights and logs at rest with keys in a hardware security module, and a full evaluation before each model update is redeployed [24].

Background

What "private" means in NIST's definitions

NIST Special Publication 800-145, published in September 2011, defines four deployment models [1]. A private cloud is "provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units)", and it "may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises" [1]. A public cloud is provisioned for open use by the general public and exists on the premises of the cloud provider [1].

Under this definition, private describes whom the infrastructure serves. It leaves open who owns the servers, who runs them and where they sit [1].

The same publication's service models set how much of the stack the customer controls [1]. Under software as a service, the consumer does not manage or control the underlying infrastructure "or even individual application capabilities" [1]. Under platform as a service, it controls the deployed applications. Under infrastructure as a service, it "has control over operating systems, storage, and deployed applications" while the provider keeps the infrastructure beneath them [1]. In a pooled cloud the customer "generally has no control or knowledge over the exact location of the provided resources" and may be able to specify location at a higher level of abstraction, such as a country, state or datacenter [1].

On-site and outsourced private clouds

NIST SP 800-146, from May 2012, splits private clouds into on-site and outsourced variants "because they affect the security perimeter" [2]. An on-site private cloud has one perimeter, which exists only if the consumer implements it. An outsourced one has two, the consumer's and the provider's, joined by a protected communications link, and the provider "accepts a responsibility to enforce the provider-implemented security perimeter and to prevent mingling of private cloud resources with other cloud resources" [2].

Who the parties are

NIST's reference architecture, SP 500-292 from September 2011, names the cloud consumer, the cloud provider, the cloud carrier that carries traffic between them and the cloud auditor among its actors [3]. Provider and consumer share control, and an IaaS consumer "should assume full responsibility for the guest OS's, while the IaaS provider controls the host OS" [3].

Cloud providers call the same split shared responsibility. AWS protects the hardware, software, networking and facilities that run its cloud services, and a customer's share depends on the services it selects [4]. Microsoft's guidance for model workloads says responsibilities vary with whether the integration is SaaS, PaaS or IaaS, and at the platform layer the customer must "build and safeguard the infrastructure that runs

the AI model, training data, and specific configurations that change the behavior of the model, such as weights and biases" [5]. Microsoft calls the page illustrative guidance, not intended to convey legal conclusions [5].

This report is for engineering leads, security reviewers and buyers choosing where a model runs.

Method

What was examined. Seven ways to run an LLM: a managed model API, dedicated tenancy inside a public cloud, provider-owned hardware in the customer's building, the organization's own servers in its own facility, its servers in colocation space, dedicated GPU capacity contracted from a GPU cloud, and a vendor-managed deployment with no link to the vendor's cloud.

The control questions. For each option, six questions: who owns the hardware, who runs the facility, where the data and prompts are processed and stored, who holds the model weights, who writes and keeps the logs, and who operates the system day to day.

Sources. NIST publications, provider documentation, two model developers' data pages, two annual reports filed with the SEC, two vendor blog posts, joint guidance and its announcement, a RAND report, serving engine and model hub documentation, one model license, a GPU maker's page and one preprint, 33 in all.

How. Each claim was read in its source, and quotes are verbatim. Provider pages are dated by their viewed date, 8 November 2026, or 9 November 2026 for [33], and by their own update date where they show one. Model families are not named in the text. The matrix in Finding 7 is a reading of these documents. Nothing was built or measured for this report, and no option is ranked.

Findings

1. The managed API baseline: the provider holds the stack, the customer holds settings

On a managed model API, the provider owns the hardware, runs the facility and hosts the model. The documents differ on what the customer can set.

Who sees prompts. Azure's documentation states that prompts, completions, embeddings and training data on Foundry Models sold by Azure are not available to other customers or to the model providers, and are not used by those providers to improve their models [6]. On Amazon Bedrock, each model provider has one Model Deployment Account per Region, owned and operated by the Bedrock service team. Model providers have no access to these accounts, so they have no access to Bedrock logs or to customer prompts and completions [8]. Google Cloud's documentation states that customer data on its managed model platform is not used to train or fine-tune models without the customer's prior permission or instruction [13].

Where prompts are processed. On Azure, prompts and responses are processed within the customer-specified geography unless the deployment type is Global or DataZone [6]. A Global deployment may process them in any geography where the model is deployed, and a DataZone deployment in the United States anywhere within the United States [6]. For both types, data stored at rest, including the abuse monitoring data store, stays in the customer-designated geography: "Only the location of processing is affected" [6]. On Bedrock, global inference profiles let the service pick any supported commercial AWS Region, geographic profiles stay within boundaries such as the US, the EU and APAC, and AWS lists the global option at approximately 10% savings [11]. On Google Cloud's global endpoint, requests and their cached data can be served from any region [13].

What is retained, and for how long. Table 1 sets out the documented defaults.

Service	Default retention for abuse or safety	Options the customer can set	Sources
Foundry Models sold by Azure	When abuse indicators are detected, a sample of prompts and completions may be reviewed by automated means, including LLMs, and by authorized employees as necessary	Modified abuse monitoring, after approval under Limited Access criteria. Stored data encrypted with AES-256 by default, with a customer managed key as an option	[6, 7]
Amazon Bedrock	Default mode: AWS may retain the data for safety and abuse prevention, and the model provider does not receive it. Under the human review mode, prompts and completions are kept up to 30 days	Retention modes, zero data retention among them. Some model families require human review by AWS as a condition of access	[9]

Service	Default retention for abuse or safety	Options the customer can set	Sources
Google Cloud's managed model platform	Prompts may be logged to detect abuse. The default in-memory cache is deleted no later than 24 hours after it is written	Request-response logging, off by default. Grounding with web search keeps query logs up to three days, and that storage cannot be turned off	[13]
One model developer's API	Abuse monitoring logs for all API usage, kept up to 30 days unless longer retention is required by law or needed to prevent harm	Modified abuse monitoring and zero data retention, each after prior approval	[14]
Another model developer's API	Inputs and outputs deleted within 30 days, unless kept longer to enforce the Usage Policy or by law. Flagged Usage Policy violations: inputs and outputs up to 2 years, trust and safety classification scores up to 7 years. Covered Models require limited retention and review	Zero data retention by agreement	[15]

Table 1. Documented retention on managed model services, as viewed on 8 November 2026.

On Azure, human reviewers can reach prompts and completions only when the abuse monitoring system has flagged them or when they form part of a potentially abusive pattern of use [6]. Under modified abuse monitoring, storage and human review are not performed, though automated review may still run [6]. AWS adds that "Most models currently do not require human review" [9].

Logs the customer keeps. Bedrock's model invocation logging is disabled by default. Once enabled, its logs can hold the full request data, response data and metadata, kept until the logging configuration is deleted, in the same account and Region [10]. Request-response logging on Google Cloud's managed model platform is also off by default, and the documentation tells customers who want zero data retention not to enable it [13]. AWS documents a private network path as well: a VPC interface endpoint with AWS PrivateLink gives a private connection, so the data is not available over the internet [12].

On a managed API, then, the provider holds the hardware, the facility, the weights and its own service logs. The customer holds settings: the deployment type or endpoint, the retention mode where offered, its own invocation logs, keys for stored data where offered, and private network paths (derived from [6, 9, 10, 12, 13, 14]).

2. Dedicated tenancy inside a public cloud

Dedicated tenancy keeps the provider's facility and operations and gives one customer its own physical hosts. AWS describes Dedicated Instances as EC2 instances that run on hardware dedicated to a single AWS account, physically isolated at the host hardware level from instances of other accounts [16]. One limit matters for data: an EBS volume attached to a Dedicated Instance does not run on single-tenant hardware [16].

The EC2 customer manages the guest operating system, its updates and its security patches [4], and in NIST's terms the provider controls the host operating system [3]. The customer holds the serving software, the weights it loads and the logs it writes inside its instances, and the provider holds the hardware, the facility and the host layer (derived from [3, 4, 16]).

NIST's 2012 guidance offers a test: an outsourced private arrangement that relies only on the mechanisms that separate customers in a public cloud, such as hardware virtualization and VLANs, "would essentially become the public cloud scenario" [2]. Dedicated Instances separate customers at the host hardware level, and their EBS storage sits outside that separation, so storage is a question of its own (derived from [2, 16]).

3. Your servers in your own facility

What NIST says about on-site. An on-site private cloud lets the organization build its own perimeter. For higher-impact data, NIST lists more restrictive firewall policies, multi-factor authentication, encryption, intrusion detection and prevention, "and even physical isolation" [2]. NIST also lists the costs: significant-to-high up-front costs, equipment paid for up front and "a fixed computing and storage capacity that has been sized to correspond to anticipated workloads and cost restrictions" [2]. In this option the organization holds the hardware, the facility, the data, the weights, the logs and the operations, or contracts parts of operations to a party it chooses (derived from [1, 2]).

The serving software's defaults become the owner's job. vLLM's security documentation states that "All communications between nodes in a multi-node vLLM deployment are insecure by default" and must be protected by placing the nodes on an isolated network [28]. Its --api-key option covers only the endpoints under four path prefixes, /v1 and /v2 among them, and the documentation advises against relying on it alone [28]. It recommends a reverse proxy that explicitly allowlists only the endpoints to be exposed [28]. Whether prompts reach the logs is also a setting. The engine's request logging is off by default, and when enabled it logs request IDs and parameters at INFO level and prompt inputs at DEBUG level [29].

Weights arrive as files that need checks. Hugging Face's Hub documentation warns that loading a pickle file can allow arbitrary code execution attacks, and pickle is the default format for PyTorch model weights [27]. It suggests loading models from trusted sources and relying on signed commits, and says its scanner "is not 100% foolproof" [27]. The joint guidance "Deploying AI Systems Securely" asks deployers to "Use cryptographic methods, digital signatures, and checksums to confirm each artifact's origin and integrity" [24]. It also asks them to inspect imported pre-trained models in a secure development zone first [24].

The license travels with the weights. One open-weight model family's community license requires a licensee whose products had more than 700 million monthly active users in the calendar month before the licensed version's 23 July 2024 release date to request a license, which the licensor may grant in its sole discretion [32]. The same license incorporates an acceptable use policy by reference [32]. How much GPU memory the weights and the key-value cache need is a separate calculation, set out in How much GPU memory LLM inference needs (https://rillor.com/insights/gpu-memory-for-llm-inference).

Provider hardware in the customer's building. On premises does not settle ownership. AWS describes an Outpost as "a pool of AWS compute and storage capacity deployed at a customer site", and states that it "operates, monitors, and manages this capacity as part of an AWS Region" [17]. The Outpost site is the customer-managed building, which must meet the facility, networking and power requirements. The Outpost equipment, its racks, servers, switches and cabling, is "owned and managed by AWS" [17]. The customer holds the building, and the provider holds the hardware and its operation (derived from [17]).

4. Your colocation space

In colocation the organization rents space and power in an operator's building and brings its own equipment. One colocation operator's annual report for 2025 describes its offerings as providing "all the components required by a customer to house its IT infrastructure or equipment", typically billed on the space and power a customer consumes and delivered under a fixed duration contract [18]. Its private cages "are typically designed and built to order for a single customer", with steel mesh walls and a locking door [18]. The filing calls its colocation data centers carrier-neutral [18], so the network carrier, NIST's cloud carrier role, is a separate choice (derived from [3, 18]).

The operator's staff can handle the equipment. The same report describes an around-the-clock, on-site support service "for remote management, installation and troubleshooting of customer data center equipment", using the operator's technicians [18]. In colocation the organization holds the hardware, the data, the weights and the logs. The operator runs the building and its power, and who may handle the servers on the organization's behalf is a contract term (derived from [18]). The joint guidance asks deployers to consider deployment environment security requirements when developing contracts for model products or services [24].

5. Dedicated GPU capacity contracted for you

Dedicated GPU capacity is hardware a GPU cloud operates and commits to a customer under contract. One GPU cloud's annual report for 2025 states that customers "generally access our platform through multi-year committed contracts", under which they buy a specified amount of capacity on a take-or-pay basis over the term [19]. Its committed contracts had a weighted-average duration of approximately five years at 31 December 2025 [19].

The same filing shows a third party behind the hardware. The company leases or licenses space in third-party data centers in the United States, Europe and the United Kingdom, and states that "we do not control the operation of these third-party facilities" [19]. Dedicated capacity of this kind involves at least three parties: the customer, the GPU cloud that operates the hardware and the owner of the facility (derived from [19]).

NIST's outsourced private cloud is the closest of its models to this arrangement (derived from [2, 19]): two perimeters joined by a protected link, resources provisioned by the provider, and main startup costs of negotiating the service level agreement, possibly upgrading the consumer's network, moving applications and operations to the cloud, and training, without server-side equipment [2].

6. Vendor-managed air-gapped deployments

A disconnected deployment removes the network link to the vendor's cloud and keeps the vendor's hardware and software. Google Cloud's documentation describes Google Distributed Cloud air-gapped as not requiring connectivity to Google Cloud and as letting customers "host, control, and manage infrastructure and services directly on your premises", in on-premises data centers [20]. Its hardware "comes fully integrated into racks and securely delivered to your premises" [20]. Google Cloud runs a site survey and gives a planning document to the hardware provider that ships the hardware [20]. Software crosses the gap on a portable storage device, checked with a SHA256 or MD5 checksum [20].

Managed models run inside such installs. In August 2025, Google Cloud announced general availability of a commercial model family on the air-gapped platform, as a fully managed endpoint within a customer or partner data center with what the post calls "a seamless, zero-touch update experience" [22]. The post also describes audit logging and access control with which customers can monitor all data traffic in and out of the environment [22]. On 20 October 2026, Google Cloud announced that a newer model of the same family and NVIDIA HGX B300 systems were generally available on the air-gapped platform, and that it supports open-weight model families [21].

The control question changes shape here. The customer holds the facility and the data, and the vendor plans the install and delivers integrated racks [20]. Neither page says who carries each update across the gap or who verifies it (derived from [20, 22]).

7. Control matrix: hardware, facility, data, weights, logs, operations

Table 2 brings Findings 1 to 6 together.

Option	Hardware	Facility	Data and prompts	Model weights	Logs	Operations	Sources
Managed model API	Provider	Provider	Processed by the provider, in a location set by deployment type or endpoint	Provider	Provider's abuse and service logs, plus the customer's invocation logs if enabled	Provider	[6, 8, 9, 10, 13, 14]
Dedicated tenancy in a public cloud	Provider's hosts, dedicated to one account	Provider	In the customer's instances, with EBS volumes on shared hardware	Customer, in its instances	Customer, in its instances	Provider below the guest OS, customer from the guest OS up	[3, 4, 16]
Provider hardware on your premises	Owned and managed by the provider	Customer	On provider-managed capacity at the customer's site	Set by the services run on it	Set by the services run on it	Provider operates, monitors and manages the capacity	[17]
Your servers, your facility	Customer	Customer	Customer	Customer	Customer, as set in its serving software	Customer or its contractor	[2, 28, 29]
Your servers in colocation	Customer	Colocation operator, whose technicians can handle equipment	Customer	Customer	Customer	Customer or its contractor	[18]
Dedicated GPU capacity	GPU cloud, under a committed contract	Third-party facility the GPU cloud leases	Set by contract	Set by contract	Set by contract	GPU cloud for the hardware, the contract above it	[2, 19]

Option	Hardware	Facility	Data and prompts	Model weights	Logs	Operations	Sources
Vendor-managed air-gapped	Integrated racks delivered to the customer	Customer or partner data center	On the customer's premises, with no link to the vendor's cloud	Vendor's managed endpoint or open-weight models	Audit logging and access control for the customer	Customer hosts, controls and manages the platform. Managed endpoint updated by the vendor, software carried in on portable media	[20, 21, 22]

Table 2. Who holds each part of an LLM deployment, read from the documents cited. "Set by contract" and "Set by the services run on it" mark cells the documents leave to the agreement or the service.

Three patterns run through the table. Location and ownership are separate questions, as Outposts shows [17]. Logs follow whoever runs the serving software: the provider keeps abuse logs on a managed API [9, 13, 14], and prompt logging is the operator's setting on self-hosted serving [29]. And every option past the managed API adds parties whose access rests on a contract [18, 19, 20].

8. Questions to settle before choosing, traced to published guidance

The joint guidance is for "organizations deploying and operating AI systems designed and developed by another entity", and it is most applicable to those doing so on premises or in private cloud environments [24]. That covers an organization serving a model it did not train. It does not apply to organizations that use systems deployed by others [24], so on a managed API its questions go to the provider. Its April 2024 announcement presents it as practices for deploying and operating externally developed systems [23].

Who is accountable for each part? The guidance asks deployers to "Identify the roles and responsibilities for each stakeholder along with how they are accountable for fulfilling them", and to identify the environment's security boundaries [24]. Table 2's columns are one way to list those roles.

Who can reach the weights? The guidance asks for an access control system for model weights, with access limited to a set of privileged users under two-person control and two-person integrity [24]. It asks deployers to isolate weight storage, for example "in a protected storage vault, in a highly restricted zone (HRZ) (i.e., a separate dedicated enclave), or using an HSM" [24]. RAND's May 2024 report on securing frontier model weights identifies 38 meaningfully distinct attack vectors and defines five security levels [26]. It recommends centralizing all copies of weights on a limited number of access-controlled and monitored systems and reducing the number of people authorized to access them. RAND wrote for developers of frontier models [26].

Who holds the keys? The guidance asks deployers to encrypt model weights, outputs and logs at rest and to store the keys in a hardware security module [24]. On Azure's managed service, stored data is encrypted with AES-256 by default, with the option of a customer managed key [6].

What is logged, by whom, and for how long? The guidance asks for logs that cover inputs, outputs, intermediate states and errors, with automated alerts and triggers [24]. On a managed API, the provider's abuse logs sit beside the customer's own [9, 14]. On self-hosted serving, the operator sets what the engine logs [29].

Who applies updates, and how are they checked? The guidance asks deployers to run a full evaluation of accuracy, performance and security before redeploying a new or different model version [24]. It also asks them to apply hardware vendors' patches and to prepare for automated rollbacks [24]. In an air-gapped install, software crosses on portable media and is checked against a checksum [20].

Which third parties have access? NIST AI 600-1, the generative profile of NIST's risk management framework from July 2024, suggests that organizations "Inventory all third-party entities with access to organizational content" [25]. It also suggests vendor contracts with clear assignment of liability and responsibility for incidents, notification of serious incidents, and service level agreements on incident response [25].

What if the hardware is not yours? NVIDIA states that confidential computing on its GPUs encrypts and protects data during execution, including inference prompts and model weights, and that its bare-metal mode keeps the infrastructure provider outside the trusted boundary [30]. A May 2025 preprint measured the cost in one setup, a single virtual machine with one H100 GPU running relaxed batch inference with model swapping [31]. Without confidential computing, latency was 20 to 30% lower, throughput 45 to 70% higher and GPU utilization about 50% higher than in confidential mode [31]. The authors attribute the difference mainly to the encryption and decryption needed to load models onto the GPU [31].

Implications

Matching the option to the workload. Where prompts may not be kept by another party, the managed API's abuse logs and review process are the first facts to check, along with the approvals that change them [7, 9, 14]. Where the requirement is a processing location, the deployment type or endpoint is the setting to read [6, 11, 13]. Where no other party may hold the weights or the logs, the options narrow to the organization's own servers, in its facility or in colocation, and the questions move to who operates them and who can touch them [2, 18, 24]. Where the organization has no hardware, dedicated GPU capacity commits capacity to it, and the contract answers what ownership answers in the other options [2, 19].

For buyers and reviewers. "On premises" and "private" each settle part of the matrix. Who owns the hardware, who runs the building and who operates the stack are separate questions [1, 17, 19]. Provider pages change: the two Foundry pages were updated on 5 June 2026 and 19 May 2026 [6, 7], and Google Cloud's on 7 November 2026 [13] and again on 8 November 2026 [33].

Cost and sizing. This report covers control. The break-even between self-hosted serving and an API is worked through in Self-hosted LLM vs API cost: a break-even method (https://rillor.com/insights/self-hosted-llm-vs-api-cost).

Private deployment. Rillor deploys and operates a private superintelligence (SI) environment, and the same environment runs on your servers, in your colocation space or on dedicated GPU capacity Rillor sources and contracts for you. In every option your data, model weights and logs stay in your environment, on hardware that is yours or dedicated to you.

Limits of this analysis

- This is a reading of public documents. No deployment option was built, tested or measured for it, and the matrix ranks none.
- Provider documentation changes without version numbers, so each fact is dated by its viewed date in the Sources.
- Approvals and settings can change the defaults described here: Azure's modified abuse monitoring and the model developers' zero data retention terms are approved per customer, and Bedrock's retention mode is set by the customer for each Region, at account or project level, with zero data retention for models that require retention granted per account and model [7, 9, 14, 15].
- The pages read give no retention period for Azure's abuse monitoring data store or for Google Cloud's abuse-monitoring prompt logs on its own models [6, 7, 13], and Google Cloud's overview of the air-gapped platform does not say who carries each update across the gap [20].
- NIST SP 800-145, 800-146 and 500-292 date from 2011 and 2012 and predate LLM serving [1, 2, 3]. Their definitions frame the options, and their statements about the state of the art are dated.
- The confidential computing figures come from one preprint, one GPU and one workload pattern [31]. NVIDIA's claims are the vendor's own [30].
- Two annual reports stand for colocation and dedicated GPU capacity [18, 19]. Each describes one operator.

Sources

1. The NIST Definition of Cloud Computing, NIST Special Publication 800-145. Peter Mell and Timothy Grance, National Institute of Standards and Technology, September 2011, viewed 8 November 2026. https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

2. Cloud Computing Synopsis and Recommendations, NIST Special Publication 800-146. Lee Badger, Tim Grance, Robert Patt-Corner and Jeff Voas, National Institute of Standards and Technology, May 2012, viewed 8 November 2026. https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-146.pdf

3. NIST Cloud Computing Reference Architecture, NIST Special Publication 500-292. Fang Liu and others, National Institute of Standards and Technology, September 2011, viewed 8 November 2026. https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication500-292.pdf

4. Shared Responsibility Model. Amazon Web Services, viewed 8 November 2026. https://aws.amazon.com/compliance/shared-responsibility-model/

5. Artificial intelligence shared responsibility model. Microsoft Learn, updated 24 August 2026, viewed 8 November 2026. https://learn.microsoft.com/en-us/azure/security/fundamentals/shared-responsibility-ai

6. Data, privacy, and security for Foundry Models sold by Azure in Microsoft Foundry. Microsoft Learn, updated 5 June 2026, viewed 8 November 2026. https://learn.microsoft.com/en-us/azure/foundry/responsible-ai/openai/data-privacy

7. Foundry Models sold by Azure abuse monitoring. Microsoft Learn, updated 19 May 2026, viewed 8 November 2026. https://learn.microsoft.com/en-us/azure/foundry/openai/concepts/abuse-monitoring

8. Data protection, Amazon Bedrock User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/bedrock/latest/userguide/data-protection.html

9. Data retention, Amazon Bedrock User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/bedrock/latest/userguide/data-retention.html

10. Monitor model invocation using CloudWatch Logs and Amazon S3, Amazon Bedrock User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/bedrock/latest/userguide/model-invocation-logging.html

11. Route model inference requests across AWS Regions with cross-Region inference, Amazon Bedrock User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/bedrock/latest/userguide/cross-region-inference.html

12. Protect your data using Amazon VPC and AWS PrivateLink, Amazon Bedrock User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/bedrock/latest/userguide/usingVPC.html

13. Gemini Enterprise Agent Platform and zero data retention. Google Cloud Documentation, last updated 7 November 2026, viewed 8 November 2026. https://docs.cloud.google.com/gemini-enterprise-agent-platform/resources/zero-data-retention

14. Data controls in the OpenAI platform. OpenAI, viewed 8 November 2026. https://developers.openai.com/api/docs/guides/your-data

15. How long do you store my organization's data? Anthropic Privacy Center, 1 July 2026, viewed 8 November 2026. https://privacy.claude.com/en/articles/7996866-how-long-do-you-store-my-organization-s-data

16. Amazon EC2 Dedicated Instances, Amazon EC2 User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/dedicated-instance.html

17. What is AWS Outposts? AWS Outposts User Guide. Amazon Web Services, viewed 8 November 2026. https://docs.aws.amazon.com/outposts/latest/userguide/what-is-outposts.html

18. Equinix, Inc. Annual Report on Form 10-K for the fiscal year ended 31 December 2025. U.S. Securities and Exchange Commission, filed 11 February 2026, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/1101239/000110123926000032/eqix-20251231.htm

19. CoreWeave, Inc. Annual Report on Form 10-K for the fiscal year ended 31 December 2025. U.S. Securities and Exchange Commission, filed 2 March 2026, viewed 8 November 2026. https://www.sec.gov/Archives/edgar/data/1769628/000176962826000104/crwv-20251231.htm

20. About Google Distributed Cloud air-gapped. Google Cloud Documentation, last updated 26 October 2026, viewed 8 November 2026. https://docs.cloud.google.com/distributed-cloud/hosted/docs/latest/gdcag/overview

21. Enabling the next era of sovereign AI: The latest Google Distributed Cloud innovations. Muninder Sambi, Google Cloud Blog, 20 October 2026, viewed 8 November 2026. https://cloud.google.com/blog/topics/hybrid-cloud/new-innovations-in-google-distributed-cloud

22. Run Gemini anywhere, including on-premises, with Google Distributed Cloud. Vithal Shirodkar, Google Cloud Blog, 28 August 2025, viewed 8 November 2026. https://cloud.google.com/blog/topics/hybrid-cloud/gemini-is-now-available-anywhere

23. Joint Guidance on Deploying AI Systems Securely. Cybersecurity and Infrastructure Security Agency, 15 April 2024, viewed 8 November 2026. https://www.cisa.gov/news-events/alerts/2024/04/15/joint-guidance-deploying-ai-systems-securely

24. Deploying AI Systems Securely: Best Practices for Deploying Secure and Resilient AI Systems, U/OO/143395-24, PP-24-1538, version 1.0. NSA Artificial Intelligence Security Center, CISA, FBI, ASD's ACSC, CCCS, NCSC-NZ and NCSC-UK, April 2024 (copy hosted by the FBI's Internet Crime Complaint Center), viewed 8 November 2026. https://www.ic3.gov/CSA/2024/240415.pdf

25. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1. National Institute of Standards and Technology, July 2024, viewed 8 November 2026. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

26. Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models, RR-A2849-1. Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley and Jeff Alstott, RAND, 30 May 2024 (revised June 2024), viewed 8

November 2026. https://www.rand.org/pubs/research_reports/RRA2849-1.html

27. Pickle Scanning, Hub documentation. Hugging Face, viewed 8 November 2026. https://huggingface.co/docs/hub/security-pickle

28. Security, vLLM documentation. vLLM project, viewed 8 November 2026. https://docs.vllm.ai/en/latest/usage/security.html

29. Engine Arguments, vLLM documentation. vLLM project, viewed 8 November 2026. https://docs.vllm.ai/en/latest/configuration/engine_args.html

30. AI Security with Confidential Computing. NVIDIA, viewed 8 November 2026. https://www.nvidia.com/en-us/data-center/solutions/confidential-computing/

31. Performance of Confidential Computing GPUs. Antonio Martínez Ibarra, Julian James Stephen, Aurora González Vidal, K. R. Jayaram and Antonio Fernando Skarmeta Gómez, arXiv:2505.16501, 22 May 2025, viewed 8 November 2026. https://arxiv.org/abs/2505.16501

32. Llama 3.1 Community License Agreement. Meta, version release date 23 July 2024, viewed 8 November 2026. https://www.llama.com/llama3_1/license/

33. Gemini Enterprise Agent Platform and zero data retention. Google Cloud Documentation, last updated 8 November 2026, viewed 9 November 2026. https://docs.cloud.google.com/gemini-enterprise-agent-platform/resources/zero-data-retention

About this report

Published by Rillor on 10 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. Private LLM deployment: who controls what. 10 November 2026. https://rillor.com/insights/private-llm-deployment-models

Notices

Rillor is not a registered investment adviser or commodity trading advisor and does not provide investment or trading advice. Research, forecasts, data and software described on this site are for research and engineering use. Nothing here is an offer or recommendation to buy or sell any security, commodity interest or digital asset. Past or simulated results do not indicate future results.

NVIDIA, AMD, Ceph, Kubernetes and related product names are trademarks of their owners and are used only to identify products. No affiliation or endorsement is implied.

Superintelligence (SI) outputs can be wrong. Important decisions should include human review.

Work with Rillor

Send us the workloads you want to run, the data they use, who uses them and where your hardware sits today. Contact us at rillor.com/contact?type=private-si or contact@rillor.com. More on private SI cloud: rillor.com/compute/private-si