

Running AI in AWS GovCloud

What is actually available, what is actually authorized,
and why those are not the same list.

2 REGIONS / 8 MODELS WEST / 2 MODELS EAST / ~20% BEDROCK PREMIUM
AWS GOVCLOUD (US) · DoD CC SRG IL4/IL5 · PREPARED 5 SEPTEMBER 2026

SECTION 1

What GovCloud is, and what it is not

AWS GovCloud (US) is two physically and logically isolated sovereign regions with their own account namespace, their own IAM, their own credentials and their own console. It is not a setting you enable on your existing AWS account. It is a separate cloud that happens to run the same APIs.

The isolation is the product. Root account holders must pass a screening process validating US persons status, defined as a citizen or green card holder under US Department of State rules. AWS operations personnel with access to the environment are subject to the same restriction. That is what makes GovCloud a defensible home for ITAR technical data and for CUI at DoD Impact Levels 4 and 5, and it is the reason the environment exists.

What it is not: a compliance certificate. Standing up a GovCloud account does not make you compliant with anything. It gives you an environment in which compliance is achievable. The controls you inherit are the ones AWS operates. Everything above the hypervisor is still yours, and under DFARS 252.204-7012 it is still your assessed environment.

The authorizations that come with the environment

Authorization	Status	What it actually means for you
FedRAMP High	Held	Now expressed as Class D under the 2026 rules. See section 5
DoD CC SRG IL2	Held	Public and non-CUI DoD workloads
DoD CC SRG IL4	Held	CUI
DoD CC SRG IL5	Held	Higher sensitivity CUI and National Security Systems. Requires physical separation from non-federal tenants, which GovCloud provides by design
ITAR	Compliant	Export controlled technical data. US persons access enforced at the root account and at AWS operations level
DFARS	Compliant	Satisfies the cloud service provider side of 7012. It does not satisfy your side
FIPS 140-2	Validated endpoints	Use the FIPS endpoints. They are not the default in every SDK

The line that gets missed in every kickoff meeting

IL5 is an authorization held by AWS for the infrastructure. It is not an authorization for your application. Your system still needs its own ATO or its own place inside your customer's boundary, and your SSP still has to describe what you built on top. Inheriting the platform authorization removes roughly the bottom third of the control set. It does not remove the other two thirds.

SECTION 2

Two regions, and they are not the same cloud

Everyone plans for GovCloud as though it is one environment with a failover. For general compute that is roughly true. For AI it is not true at all, and the gap is large enough to dictate architecture.

	AWS GovCloud (US-West)	AWS GovCloud (US-East)
Region code	us-gov-west-1	us-gov-east-1
Bedrock foundation models	8	2
Anthropic	Claude Opus 5, Claude Sonnet 5, Claude Fable 5.1	Claude Fable 5.1
Amazon	Nova Pro, Nova Lite, Nova Micro, Nova Multimodal Embeddings, Titan Text Embeddings V2	Titan Text Embeddings V2
Practical read	The region you build in	Embeddings and a small model. Not a failover for a Bedrock workload

Source: AWS Bedrock model regional compatibility documentation, read 5 September 2026. This list moves. Section 10 gives you the command to check it yourself on the day you need it.

What this does to your design

If your continuity plan says the AI workload fails over from West to East, it does not. Two models cannot carry an eight model application. You have three honest options, and picking one early is cheaper than discovering the problem during an assessment.

Option	What it costs you	When it is right
Build West only, accept the single region exposure	Regional outage takes the AI capability down. Everything else stays up	Almost always, for a first deployment. Say it out loud in the risk register rather than discovering it later
Constrain the design to what East also carries	You give up the frontier models and build on embeddings plus a small model	Retrieval heavy workloads where the model is doing extraction rather than reasoning
Degrade gracefully to the East model set	Two prompt paths, two evaluation sets, real engineering effort	Only when the workload is genuinely mission critical and somebody is funding the second path

SECTION 3

Available is not authorized

This is the distinction the whole report exists for. A model being available in a GovCloud region and a model being FedRAMP High and IL4/IL5 approved are two different facts, published in two different places, on two different schedules.

Availability means you can call it. Authorization means the model has been assessed within the boundary and may lawfully process CUI at that impact level. AWS announces authorization separately, model by model, in the What's New feed. Availability is in the Bedrock documentation. The two lists do not update together and they are not the same length.

Authorization announcements to date

Announced	Models	Authorization granted
May 2025	Anthropic Claude 3.5 Sonnet v1, Claude 3 Haiku; Meta Llama 3 8B and 70B	FedRAMP High and DoD CC SRG IL4 and IL5
25 June 2026	OpenAI GPT, OpenAI GPT OSS, NVIDIA Nemotron	FedRAMP High and DoD CC SRG IL4 and IL5

The trap, stated plainly

The newest and most capable models in the West region are not necessarily on the authorization list. As of this writing the frontier Anthropic models are callable in us-gov-west-1, while the models named in the AWS authorization announcements are an older and partly different set. Calling an available but unannounced model with CUI in the prompt is the kind of finding that shows up in an assessment and is very hard to explain afterwards. Before any model touches CUI, get the authorization statement in writing for that specific model, in that specific region, on that specific date, and keep it in your evidence package.

What to do about it

Maintain a two column register: model, and the authorization evidence for it. One row per model you have enabled. If the second column is empty, that model is for non-CUI work only and your access control has to actually enforce that, not just document it. Bedrock guardrails and IAM policy scoped per model identifier are the mechanism.

SECTION 4

The commercial account you cannot avoid

To enable a Bedrock foundation model in GovCloud, you must first accept that model's end user licence agreement in a standard commercial region, us-east-1 or us-west-2, and only then enable the model in your GovCloud account through the console or CLI.

This surprises people, and it surprises security teams more than it surprises engineers. An organisation that has deliberately built a GovCloud only posture, with no commercial AWS footprint, discovers at model enablement time that it needs a commercial account after all.

Nothing controlled crosses. The commercial account is used to accept a licence, not to process data. But it is an account that has to exist, be owned by someone, be paid for, be secured, and be explained. Handle it deliberately:

Do	Why
Create the commercial account before you need it, not during the build	Enablement blocks on it, and account creation is not instant
Keep it empty. No workloads, no data, no VPCs	An empty account is trivially easy to explain to an assessor
Document its purpose in one paragraph in the SSP	An undocumented AWS account discovered during an assessment is a finding. A documented licence-acceptance account is a footnote
Apply an SCP that denies everything except the Bedrock model subscription actions	Makes the constraint technical rather than procedural, which is what an assessor wants to see
Put it in the same AWS Organization if your structure allows	Consolidated billing and one place to prove the deny policy is applied

SECTION 5

FedRAMP changed in June 2026

On 25 June 2026 FedRAMP published its Consolidated Rules for 2026. The three impact levels became four certification classes. This matters because your contract language, your SSP and every platform comparison you have written are phrased in the old vocabulary.

Old	New	Note
Low	Class B	
Moderate	Class C	This is the one DFARS 252.204-7012 points at when it requires FedRAMP Moderate equivalency for external cloud services holding covered defense information
High	Class D	AWS GovCloud sits here. Class D has no 20x path and no Program path. Agency authorization under Rev5 remains the only route, with a Class D pilot targeted for FY27
n/a	Class A	New transitional class for providers intending to certify higher later

Two dates to put in your compliance calendar

Rev5 sunsets on 31 December 2028. Everything currently authorized under it has to move. And the Class D pilot is targeted for the first half of FY27, which means the High equivalent tier is the last one to get an automated path. If your platform selection assumed FedRAMP was getting faster across the board, it is getting faster for Classes A through C and not yet for D. Plan the GovCloud path on Rev5 timelines.

What to change in your documents

Do not do a find and replace. DFARS 7012 still says Moderate, because the clause text has not been amended. Write the mapping down once, in your SSP, in a sentence that says which class you are relying on and which clause language it satisfies. Then when an assessor arrives with the old vocabulary and your evidence uses the new one, the bridge is already on the page.

SECTION 6

What it costs

GovCloud is more expensive than commercial and the premium is not uniform. Budget it as two separate numbers, because they behave differently.

Component	Premium over commercial	Basis
Core compute, storage, networking	10 to 15 percent	Published GovCloud pricing versus the equivalent commercial region
Bedrock inference	About 20 percent	Observed on Claude Opus 4.8: \$6.00 and \$30.00 per million input and output tokens in GovCloud against \$5.00 and \$25.00 on commercial Bedrock and on Anthropic's own API
Regional versus global endpoints	About 10 percent	From Claude Sonnet 4.5 onward, regional and multi-region endpoints carry a premium over global endpoints. GovCloud is regional by definition, so this is already inside the number above
Support plan	Same tiers	Business or Enterprise support is priced the same way but applies to a second account, so it is a second bill

A worked number, so this is not hand waving

A team of forty engineers using a retrieval assistant over internal procedures, at a realistic ten queries per person per working day, with roughly 8,000 input tokens of retrieved context and 700 output tokens per query, runs about 640,000 input tokens and 56,000 output tokens per day. Priced at the GovCloud Opus rate above, that is roughly \$5.52 per day in input and \$1.68 in output, so about \$150 a month in inference. Move it to a mid tier model and it is under \$40. Add the embeddings for the corpus, the vector store, the compute and the logging, and a workload of that shape lands between \$600 and \$1,400 a month all in.

The point of that arithmetic is not the number. It is that inference is rarely the expensive part. The environment around it is. Teams routinely over-model, paying frontier prices for extraction work a small model does correctly, and then wonder why the bill looks the way it does.

SECTION 7

Standing one up

Elapsed calendar time, not effort. Assumes an organisation with an existing commercial AWS presence and someone who has done this before. Add four to six weeks if the account structure has to be designed from nothing.

#	Phase	Elapsed	What actually happens
0	Account and screening	Wk 1-3	Request the GovCloud account against your commercial payer. Root account holder completes US persons screening. This is the longest pole and it is entirely outside your control, so start it first
1	Boundary decision	Wk 1-4	Which data goes in, at which impact level, and what stays out. Written determination per data category before anything is built. If you skip this you will rebuild the network later
2	Landing zone	Wk 3-7	Organization structure, SCPs, identity federation, logging to a separate account, FIPS endpoints enforced, KMS with customer managed keys
3	Model enablement	Wk 5-7	Licence acceptance in the commercial account, model enablement in GovCloud, per-model IAM scoping, Bedrock guardrails, and the authorization register from section 3
4	Workload build	Wk 6-14	The actual application. Retrieval, prompts, evaluation set, human in the loop where output is contractual or safety related
5	Evidence and SSP	Wk 10-16	Architecture diagram, data flow, control mapping, SSP language, POA&M; entries for anything open. Do this alongside the build, not after it
6	Assessment readiness	Wk 14-18	Internal walk through against 800-171A. Fix what it finds. Then the conversation with your assessor is short

Sixteen to eighteen weeks from a standing start to something an assessor will accept. Six to eight if the landing zone already exists and you are only adding the AI workload to it.

SECTION 8

What GovCloud does not buy you

Every platform document in this market is a list of what the platform does. This is the other list, and it is the more useful one.

It does not	What that means in practice
Make you CMMC compliant	GovCloud satisfies the cloud provider side. Your access control, audit, configuration management, incident response and awareness training are unchanged. Most of the 110 requirements in 800-171 are still yours
Give you the newest models on day one	GovCloud lags commercial on model availability, and authorization lags availability again. Two lags, compounding. If your business case needs a model within a month of its release, GovCloud is the wrong plan
Cover every AWS service	Service parity with commercial is incomplete and always has been. Check every service in your architecture, individually, before you commit to it. Assume nothing carries over
Give you a failover for AI	See section 2. Two models in the East region is not a disaster recovery posture for an eight model application

It does not	What that means in practice
Handle classified data	IL6 and above is the AWS Secret Region and above. Different environment, different contract, different conversation
Remove the need for a written AI use policy	A compliant platform with no acceptable use policy still produces shadow AI, still produces ungoverned output in contract deliverables, and still leaves you unable to answer what your people are doing with it

SECTION 9

GovCloud, Azure Government, Ask Sage, or on premises

Four compliant paths. They are not competing for the same job and the decision is usually made on constraints rather than preference.

Path	Fastest when	Slow or wrong when
AWS GovCloud	You already run AWS, you have engineers, and you want the AI inside your own boundary next to your own data	You have no cloud team, or you need a capability in weeks rather than a quarter
Azure Government	You are a Microsoft shop, your identity is already Entra, and your documents are already in M365	You need the AI in US Gov Texas or a DoD region. Foundry is not in those. See the companion report
Ask Sage	You need a compliant capability in days, you have no platform team, and inheriting the authorization is worth more than controlling the stack	You need deep integration with your own systems, or the workload is unusual enough that a platform will not fit it
On premises inference	No cloud is acceptable at any impact level, or the programme forbids it outright	You want frontier model quality. Local models are good and getting better, and they are still not the same thing

The honest ordering for most of the defense industrial base in North Alabama

Start with Ask Sage if you need capability now and have no platform team, because inheriting a FedRAMP High and IL5 authorization in days is a real advantage and most firms do not have a better use of a quarter. Move to GovCloud or Azure Government when the integration requirements outgrow what a platform will do, which is usually the second year rather than the first. Go on premises only when the programme leaves you no choice, and go in knowing it is the most expensive of the four to run. Anyone who tells you there is one right answer is selling one of them.

SECTION 10

Sources, and how to check this yourself

Model availability and authorization both move. Everything in this report was read on 5 September 2026 and some of it will be stale by the time you read it. The verification method matters more than the snapshot, so here is both.

Check availability yourself

From a GovCloud account with the CLI configured, this returns the live list for the region and takes about two seconds:

```
aws bedrock list-foundation-models --region us-gov-west-1 --query "modelSummaries[].modelId" --output table
```

Run it against us-gov-east-1 as well. If the two lists match, something has changed since this report and the asymmetry in section 2 is no longer the constraint it is today.

Sources

What	Where
GovCloud regions, authorizations, US persons requirement, IAM separation	aws.amazon.com/govcloud-us
Bedrock model availability by region, and the commercial EULA prerequisite	docs.aws.amazon.com/bedrock/latest/userguide/models-region-compatibility.html
Claude 3.5 Sonnet v1, Claude 3 Haiku, Llama 3 8B and 70B FedRAMP High and IL4/IL5 approval, May 2025	aws.amazon.com/about-aws/whats-new/2025/05/amazon-bedrock-models-fedramp-high-dod-il-4-5-govcloud/
OpenAI GPT, GPT OSS and NVIDIA Nemotron FedRAMP High and IL4/IL5 approval, 25 June 2026	aws.amazon.com/about-aws/whats-new/2026/06/addl-bedrock-model-fedramp-il-5-govcloud/
Bedrock in GovCloud US-East general availability	aws.amazon.com/about-aws/whats-new/2024/11/amazon-bedrock-available-aws-govcloud-east-region/
FedRAMP Consolidated Rules 2026, certification classes A to D, Rev5 sunset 31 December 2028, Class D pilot targeted FY27	fedramp.gov , plus published analyses from Crowell & Moring and Davis Wright Tremaine, July 2026
DoD Cloud Computing SRG, impact level definitions and separation requirements	public.cyber.mil/dccs/dccs-documents
DFARS 252.204-7012 FedRAMP Moderate equivalency requirement	acquisition.gov
Anthropic pricing, and the regional versus global endpoint premium	anthropic.com pricing documentation

What is measured here and what is judgement

Measured, and traceable to a primary source: the region codes, the authorization list, the model availability lists, the commercial EULA prerequisite, the US persons requirement, the FedRAMP class mapping and the published token prices. Judgement, and worth arguing with: the phase durations in section 7, the cost worked example in section 6, the ordering in section 9, and the view that most firms should start on a platform rather than a landing zone. If you disagree with the judgement, the measured parts still stand on their own.