TEST / CONTROL / AI / MULTIPLIED

SUMMARY REPORT · AZURE GOVERNMENT AI

Running AI in Azure Government

Five regions, two of which have the AI. Why the deployment type you pick changes which models exist.

5 REGIONS / 2 WITH FOUNDRY / 4 DEPLOYMENT TYPES / 11 MODELS

AZURE GOVERNMENT · DoD CC SRG IL2/IL4/IL5 · PREPARED 5 SEPTEMBER 2026

Alexander's Test & Control Solutions

Huntsville, Alabama · (256) 341-7917 · Billy@AlexanderTestControl.com

alexandertestcontrol.com

SECTION 1

Five regions, and they are not interchangeable

Azure Government is a physically separated sovereign cloud with its own regions, its own portal at portal.azure.us, and its own service catalogue. Microsoft holds a FedRAMP High provisional authorization from the Joint Authorization Board and DoD provisional authorizations at Impact Levels 2, 4 and 5 across it.

Region	Class	Authorization	AI available
US Gov Virginia	Azure Government	FedRAMP High, IL2, IL4, IL5	**Yes**
US Gov Arizona	Azure Government	FedRAMP High, IL2, IL4, IL5	**Yes**
US Gov Texas	Azure Government	FedRAMP High, IL2, IL4, IL5	**No**
US DoD East	Azure Government DoD	Separate IL5 PA, DoD exclusive use	**No**
US DoD Central	Azure Government DoD	Separate IL5 PA, DoD exclusive use	**No**

> **The finding that changes architectures**
> Microsoft Foundry, which is where Azure OpenAI now lives, is available in usgovvirginia and usgovarizona only. It is not in US Gov Texas, and it is not in either of the two Azure Government DoD regions. If your programme requires the workload to sit in US DoD East or US DoD Central, there is no Azure OpenAI there to call. That is not a capacity constraint you can escalate around. It is a catalogue fact, and it needs to be established in week one rather than discovered in month four.

The practical consequence: an organisation with a DoD region requirement and an AI requirement has to either place the AI workload in the commercial Azure Government regions and manage the boundary between them, or use a different path entirely. Both are legitimate. Neither is free, and the second is usually cheaper than people expect.

SECTION 2

The deployment type changes which models exist

This is the part that is genuinely non obvious. In Azure Government the same region offers different model catalogues depending on which deployment type you choose. Two teams in the same subscription, in the same region, can have entirely different models available, and neither will understand why.

Model availability by deployment type

Model	Data Zone Standard	Data Zone Provisioned	Standard	Provisioned
gpt-5.6-sol	Both regions	Both regions	No	No
gpt-5.6-terra	Both regions	Both regions	No	No
gpt-5.6-luna	Both regions	Both regions	No	No

Model	Data Zone Standard	Data Zone Provisioned	Standard	Provisioned
gpt-5.1	Both regions	No	No	No
gpt-4.1	Both regions	No	Both regions	No
gpt-4.1-mini	Both regions	Both regions	Both regions	No
o3-mini	Both regions	Both regions	No	No
gpt-4o	Both regions	Both regions	Both regions	Both regions
text-embedding-3-large	Both regions	No	Arizona only	No
text-embedding-3-small	Both regions	No	Arizona only	No
text-embedding-ada-002	No	No	Arizona only	No

Both regions means usgovvirginia and usgovarizona. Read from the Microsoft Foundry models in Azure Government documentation, 5 September 2026.

What to take from that table

Observation	So
Data Zone Standard is the only type with the full catalogue	Unless you have a specific reason not to, this is the default. It routes traffic within Azure Government to provide higher throughput and it is the only way to reach gpt-5.1
Plain Standard has no gpt-5.6, no gpt-5.1 and no o3-mini	A team that provisions Standard by habit gets a materially older model set and will conclude Azure Government is behind. It is not. They picked the wrong deployment type
Provisioned managed carries gpt-4o only	If you need reserved throughput with a guaranteed latency profile, gpt-4o is your model. Everything else means accepting shared capacity
Embeddings are the messiest row in the table	Under plain Standard they exist in Arizona and not Virginia. Under Data Zone Standard both regions have the v3 models and neither has ada-002. A retrieval build that assumed parity will fail on deployment

SECTION 3

Two traps in the model specifications

Both of these are documented by Microsoft and both routinely get missed because they sit in a note rather than in the headline number.

The GPT-4.1 context window is not the one you read about

GPT-4.1 is marketed on a 1,047,576 token context window. In Azure Government that figure is explicitly marked as not available. What you actually get is 128,000 tokens on standard and provisioned managed deployments, and 300,000 on batch deployments. Maximum output is 32,768.

If your design assumed you could put a million tokens of program documentation into a single prompt, it will not work, and the failure arrives at integration rather than at design review. Build retrieval, not a giant prompt. You should be doing that anyway for cost and accuracy reasons, but here it is not optional.

The tool definition failure, which is worse because it is silent about the cause
There is a known issue affecting all GPT-4.1 series models where tool or function definitions exceeding 300,000 tokens fail, even though the stated context limit was never reached. The Chat Completions API returns a 400 with a context_length_exceeded message that counts the function definitions rather than the messages. The Responses API returns a generic 500 server error with no useful detail at all. If you are building an agentic workflow with a large tool surface, you will hit this, and the 500 will send you looking in the wrong place for a day. Keep tool definitions lean and count their tokens as part of your budget.

Model retirement diverges from commercial

Microsoft states plainly that in some cases models are retired in Azure Government earlier or later than in the commercial cloud. This matters more than it sounds. A retirement date you tracked from the commercial documentation is not the date that applies to your government deployment, in either direction.

Practical handling: put the Azure Government model retirement page in your change management process with a named owner and a quarterly review, and pin the model version explicitly in your deployment rather than tracking a floating alias. A model that disappears underneath a production workload inside an assessed boundary is an availability incident and, depending on your contract, a reportable one.

SECTION 4

IL5 isolation is your job, not Microsoft's

Microsoft holds the IL5 provisional authorization. That authorization comes with conditions, and in the Azure Government commercial regions those conditions land on you.

The DoD Cloud Computing SRG requires, for IL5, physical separation from non-DoD and non-federal tenants, and at minimum virtual or logical separation between tenant and mission systems. The two Azure Government DoD regions deliver the physical separation by being DoD exclusive. The three commercial Azure Government regions do not, which is why Microsoft publishes a separate isolation guidance document and tells you to review it before deploying IL5 workloads there.

Requirement	Source	Who satisfies it
Physical separation from non-federal tenants	SRG 5.2.2.3	Microsoft in the DoD regions. In US Gov Virginia and Arizona you satisfy it through the documented isolation configuration
Virtual or logical separation between missions	SRG 5.2.2.3	You. Subscription and network design
CSP personnel are US citizens, nationals or persons	SRG 5.6.2	Microsoft
FedRAMP High baseline plus DoD FedRAMP+ controls	SRG 5.1.1	Microsoft holds it. You inherit it
Everything in NIST SP 800-171 that is not infrastructure	DFARS 252.204-7012	You, entirely

Microsoft publishes an Azure Policy regulatory compliance built-in initiative for DoD IL5 in Azure Government. Turn it on early. It will not make you compliant, and Microsoft says as much, but it gives you a control-by-control dashboard that is far better evidence than a spreadsheet somebody maintains by hand.

SECTION 5

FedRAMP changed in June 2026

On 25 June 2026 FedRAMP published its Consolidated Rules for 2026, replacing three impact levels with four certification classes. Your contract language and your SSP are written in the old vocabulary.

Old	New	Note
Low	Class B	
Moderate	Class C	What DFARS 252.204-7012 points at when it requires FedRAMP Moderate equivalency for external cloud services holding covered defense information
High	Class D	Azure Government 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

Rev5 sunsets on 31 December 2028. Note also the SRG's own position, which predates all of this and still holds: FedRAMP is the minimum baseline at every impact level, and an IL5 provisional authorization is built on a FedRAMP High authorization supplemented by DoD FedRAMP+ controls. The class rename does not change that structure, it renames one layer of it.

Do not run a find and replace across your documents. DFARS 7012 still says Moderate because the clause text has not been amended. Write the mapping down once, in the SSP, so that 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

Microsoft does not publish a single government premium percentage, and anyone quoting you one has made it up. In Azure Government the dominant cost decision is not the region, it is the deployment type, and the two options behave completely differently.

Deployment type	Billing model	Right when
Data Zone Standard	Per token consumed, shared capacity, traffic routed within Azure Government	Variable or unknown load, pilots, and anything where you are still learning the usage pattern. This is where almost everyone should start
Data Zone Provisioned managed	Provisioned throughput units purchased in advance across Azure Government infrastructure	Sustained predictable load with a latency requirement you have to guarantee. You are buying capacity, not tokens, so idle time is wasted money
Standard	Per token, single region	Rarely the right pick now, because the catalogue is materially smaller. See section 2
Provisioned managed	PTUs, single region	gpt-4o only. Reserved capacity for one model

How to actually budget it

Price your own workload rather than trusting a rule of thumb. Take the number of users, a realistic queries-per-user-per-day figure, and your actual retrieved context size, then multiply. A team of forty engineers at ten

queries a day with 8,000 tokens of retrieved context and 700 tokens of output runs roughly 640,000 input and 56,000 output tokens per day. On a mid tier model that is a small monthly number. The environment around it, meaning the search index, the storage, the networking, the logging and the identity, routinely costs more than the inference does.

The provisioned throughput trap is worth naming. PTUs are bought in blocks and billed whether or not you use them. Teams commit to reserved capacity before they know their load curve, then run at thirty percent utilisation for a year. Start on Data Zone Standard, measure for a quarter, then convert the portion of load that is genuinely steady.

SECTION 7

Standing one up

Elapsed calendar time for an organisation that already has an Azure commercial presence and an Entra tenant. Azure Government requires a separate tenant and separate identities, which is the step people underestimate.

#	Phase	Elapsed	What actually happens
0	**Eligibility and tenant**	Wk 1-4	Azure Government eligibility validation, separate tenant, separate Entra directory, separate subscriptions. Your commercial identities do not come with you
1	**Boundary and region decision**	Wk 1-3	Which data, at which impact level, in which region. If a DoD region is required, resolve the Foundry gap in section 1 now
2	**Landing zone**	Wk 3-7	Management groups, policy, the DoD IL5 built-in initiative, private endpoints, customer managed keys, log analytics in a separate subscription
3	**Foundry and model deployment**	Wk 5-8	Resource creation, deployment type decision from section 2, model version pinning, content filter configuration, quota request. Quota is not instant in Azure Government
4	**Workload build**	Wk 7-15	Retrieval over your own documents, prompts, evaluation set, human in the loop wherever output is contractual or safety related
5	**Evidence and SSP**	Wk 11-17	Architecture, data flow, control mapping, SSP language, POA&M; entries. Alongside the build, not after
6	**Assessment readiness**	Wk 15-19	Internal walk through against 800-171A, then remediate

Seventeen to nineteen weeks from nothing to assessment ready. Longer than AWS GovCloud in practice, almost entirely because of the separate tenant and identity work in phase 0.

SECTION 8

What Azure Government does not buy you

It does not	What that means in practice
Make you CMMC compliant	Microsoft satisfies the cloud provider side. Most of the 110 requirements in NIST SP 800-171 remain yours, and the IL5 isolation conditions in section 4 are yours as well
Put AI in every region	Foundry is in two of five. US Gov Texas and both DoD regions have no Azure OpenAI. This is the single most consequential fact in this report
Give you commercial model parity	The government catalogue trails commercial, and within government the catalogue depends on deployment type. Do not promise a model you have not confirmed for your region and your deployment type
Give you the advertised GPT-4.1 context window	128,000 tokens, not 1,047,576. See section 3
Follow commercial retirement dates	Models retire earlier or later in Azure Government. Track the government page, not the commercial one
Bring your identities across	Separate tenant, separate directory, separate credentials, separate conditional access policy. Plan the identity work as its own project
Handle classified data	IL6 and above is a different environment and a different conversation

SECTION 9

Azure Government, AWS GovCloud, Ask Sage, or on premises

Path	Fastest when	Slow or wrong when
Azure Government	You are a Microsoft shop, identity is already Entra, and the documents you want to reason over already live in M365	You need AI in US Gov Texas or a DoD region, or you need a model the government catalogue does not carry
AWS GovCloud	You already run AWS and you want the AI next to your own data inside your own boundary	You have no cloud team, or you need the capability in weeks. See the companion report for the region asymmetry there
Ask Sage	You need a compliant capability in days and inheriting a FedRAMP High and IL5 authorization is worth more than controlling the stack	You need deep integration with internal systems
On premises inference	No cloud is acceptable at any impact level	You want frontier model quality, or you have nobody to run it

The one line worth remembering

Choose Azure Government when your identity and your documents are already Microsoft, because that integration is worth more than any model benchmark. Choose AWS GovCloud when your data and your engineers are already there. Choose Ask Sage when you need the capability before you can staff a platform team. Choose on premises when the programme leaves you no choice, and price it honestly, because it is the most expensive of the four to run over three years.

SECTION 10

Sources, and how to check this yourself

Model catalogues move. Everything here was read on 5 September 2026 from Microsoft's own documentation, and the deployment type table in section 2 in particular will change. Check it before you design against it.

Check availability yourself

From the Azure CLI, authenticated against your Azure Government subscription:

```
az cloud set --name AzureUSGovernment
```

```
az cognitiveservices account list-models --name --resource-group --query "[].{model:name, version:version, sku:skus[].name}" -o table
```

Sources

What	Where
Model availability by region and deployment type, GPT-4.1 context limits, the tool definition known issue, retirement divergence	learn.microsoft.com · Foundry Models sold by Azure in Azure Government
Azure Government region list, FedRAMP High P-ATO, DoD IL2/IL4/IL5 provisional authorizations, DoD region separation	learn.microsoft.com/azure/compliance/offerings/offering-dod-il5
IL5 isolation requirements for the commercial Azure Government regions	learn.microsoft.com/azure/azure-government/documentation-government-impact-level-5
DoD IL5 Azure Policy regulatory compliance initiative	learn.microsoft.com/azure/governance/policy/samples/gov-dod-impact-level-5
DoD Cloud Computing SRG, sections 3.1.3, 5.1.1, 5.2.2.3 and 5.6.2	public.cyber.mil/dccs/dccs-documents
FedRAMP Consolidated Rules 2026, classes A to D, Rev5 sunset 31 December 2028	fedramp.gov, plus published analyses from Crowell & Moring and Davis Wright Tremaine, July 2026
DFARS 252.204-7012 FedRAMP Moderate equivalency	acquisition.gov
Deployment types and provisioned throughput in Azure Government	learn.microsoft.com · Azure OpenAI deployment types, government variant

What is measured here and what is judgement

Measured, and traceable to Microsoft's own documentation: the region list, the authorizations, the model availability matrix by deployment type, the GPT-4.1 context limits and tool definition issue, the retirement divergence statement, and the SRG isolation requirements. Judgement, and worth arguing with: the phase durations in section 7, the advice to start on Data Zone Standard rather than provisioned, and the platform ordering in section 9. If you disagree with the judgement, the measured parts still stand.