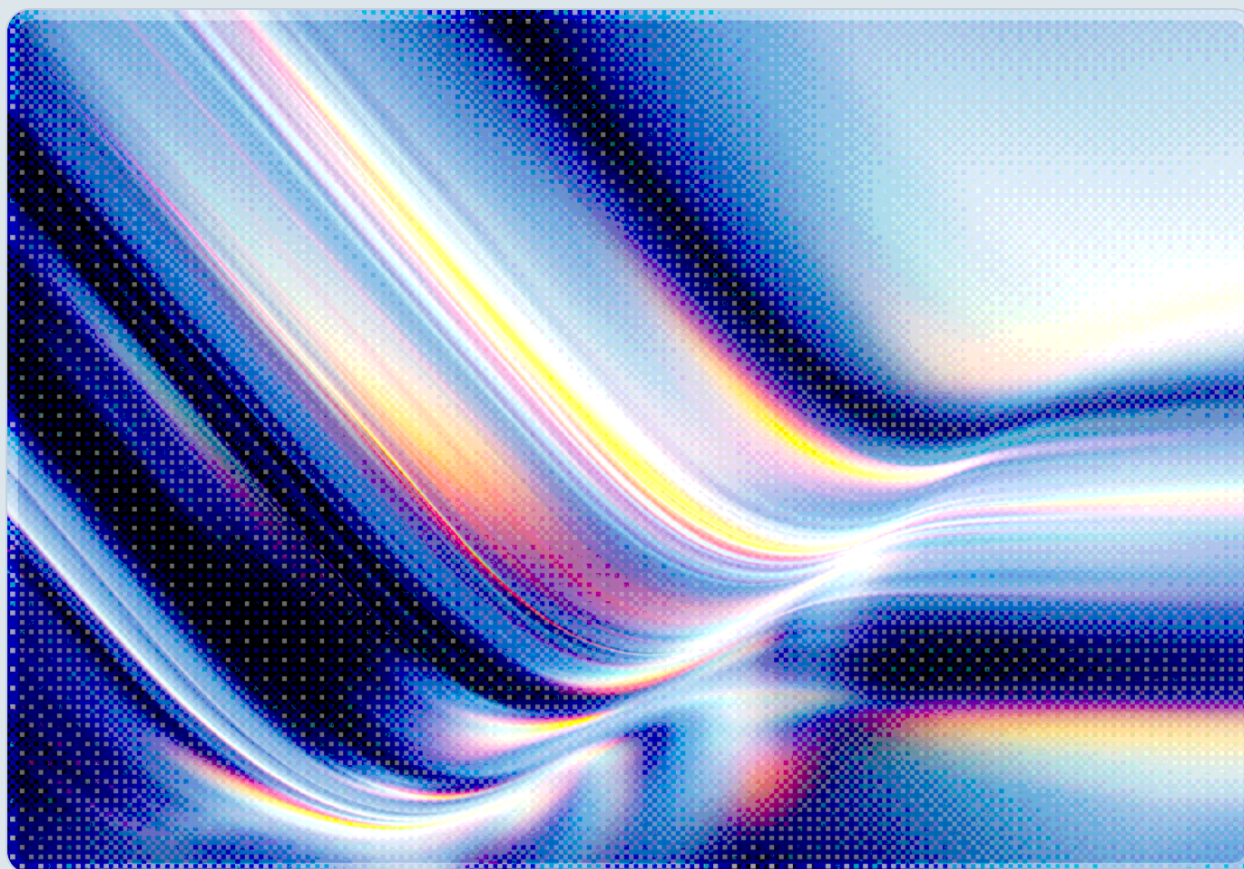


Same Model, Many Clouds: Cross-Hosting and Model Diversity

Why multi-cloud access to frontier models does not provide the resilience enterprises assume

2026-10-25

 AI-assisted Rapid Research



© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

During 2026 the major frontier model families largely stopped being tied to a single cloud. OpenAI models now run on Amazon Bedrock, and Anthropic's Claude models are generally available on AWS, Google Cloud, and Microsoft Foundry. Enterprises that adopted multi-cloud AI architectures before 2026 may therefore be treating "multi-cloud" and "multi-model" as a single resilience control when they are really two. Cross-hosting spreads *infrastructure* risk across clouds, but it does not spread *model-level* risk. The model weights, the safety training, the vendor's usage policy, and the vendor's regulatory exposure are the same on every cloud that hosts the model.

The June 2026 suspension of Anthropic's Fable 5 and Mythos 5 illustrates the gap. According to Anthropic, a U.S. export-control directive followed a report of a method for bypassing Fable 5's safeguards, and the resulting suspension removed both models from Anthropic's own platform and from AWS, Google Cloud, and Microsoft Foundry at once. Using a second model family helps, but it does not fully solve the problem either. Peer-reviewed research shows that errors from different large language models are substantially correlated, and that the most accurate models tend to be the most correlated with one another.

In response, enterprises should map their AI dependencies on three separate axes: hosting infrastructure, model lineage, and governing jurisdiction or policy authority. They should then test failover against model-level withdrawal, not only against regional cloud outages, and measure how often their primary and fallback models fail on the same inputs.

Background

Through most of 2023 to 2025, choosing a frontier model largely meant choosing a cloud. OpenAI's frontier models reached enterprises mainly through Microsoft Azure or OpenAI's own API. Google's Gemini models were offered through Google Cloud. Anthropic's Claude was the partial exception, with distribution through Amazon Bedrock and Google Vertex AI. Under these arrangements, an enterprise that split its AI workloads across two clouds usually ended up running two different model families as well. Infrastructure diversity and model diversity came as a pair, and resilience programs could treat a multi-cloud AI architecture as a reasonable proxy for a multi-model one.

That pairing broke apart during 2026. On April 27, 2026, Microsoft and OpenAI announced an amended partnership that keeps Azure as OpenAI's primary cloud and first launch venue. Microsoft's license to OpenAI intellectual property became non-exclusive, and OpenAI gained the right to serve its products to customers on any cloud provider [1]. Amazon then brought OpenAI models to Bedrock: a limited preview opened on April 28, the GPT-5.6 family reached general availability on July 9, and GPT-6 Astra was added on September 8, 2026. Amazon states that pricing matches OpenAI's first-party rates [2]. Microsoft announced on June 29, 2026 that Claude in Microsoft Foundry was generally available. The service runs on Azure infrastructure, but Anthropic operates the inference service and acts as data processor and SLA provider [3], which means the hosting cloud does not independently control the model's availability. Google, for its part, has extended Gemini beyond its public cloud regions to on-premises and air-gapped deployments through Google Distributed Cloud [4].

The result is a market in which, for at least two of the three leading frontier model families, the same model can be reached through several hyperscalers. For procurement teams this is attractive, because the model can be bought against existing cloud commitments and governed with familiar identity and network tooling. This note looks at a less obvious consequence. When the same model is available on many clouds, the cloud choice reveals little about which model failures an enterprise is actually exposed to, and resilience strategies built on the old pairing may now protect less than their owners believe.

Security Analysis

Two Kinds of Failure, One Control

Enterprise AI dependencies can fail in two broadly different ways. Infrastructure failures stop a hosting environment from serving requests: regional outages, control-plane faults, capacity exhaustion, or network partitions. A clear example is the AWS us-east-1 event of October 19–20, 2025. AWS attributes it to a latent race condition in DynamoDB's DNS management automation that left an empty DNS record for the service's regional endpoint, and the impact spread to EC2 instance launches, Network Load Balancer health checks, Lambda, and other services for more than fourteen hours [5]. Cloud-level redundancy addresses this kind of failure directly. If the same model is available in another region or on another cloud, traffic can be rerouted and the application keeps working with identical behavior.

Model-level failures are a different category. They include withdrawal or suspension of a model by its developer or by a government, a vendor usage-policy change that blocks a use case, a safety-classifier update that starts refusing previously accepted requests, a jailbreak or prompt-injection technique that works against the model's shared training, and a behavioral regression introduced by a model update. The most consequential of these do not depend on where the model runs. Weights, alignment training,

system-level safeguards, and the developer's legal obligations go with the model to every cloud that hosts it. Moving traffic from one cloud to another therefore gives little or no protection against them. At most it buys time where a cloud provider lags in rolling out a model update or applies its own content-filtering layer and acceptable-use policy, and that lag is incidental rather than a designed control. Cross-hosting thus makes it easy to believe one is protected because the failover path *appears* diverse.

The table below summarizes how each resilience posture holds up against these failure classes. The assessments are qualitative and follow from the architectural reasoning above.

Failure class	Example	Same model, second region	Same model, second cloud	Different model family	Different family, different jurisdiction
Regional infrastructure outage	DNS or control-plane fault in one region	Mitigates	Mitigates	Mitigates if hosted in another region	Mitigates
Whole-provider cloud incident	Global control-plane or identity failure	Does not mitigate	Mitigates	Mitigates if hosted elsewhere	Mitigates
Developer-initiated withdrawal or policy change	Model deprecation, usage-policy tightening	Does not mitigate	Does not mitigate	Mitigates	Mitigates
Government-directed suspension	Export-control directive against a model	Does not mitigate	Does not mitigate	Partially (same-jurisdiction vendors may face similar orders)	Mitigates, with legal and data-residency trade-offs
Shared vulnerability or jailbreak	Technique that exploits	Does not mitigate	Does not mitigate	Partially (attacks often transfer	Partially

Failure class	Example	Same model, second region	Same model, second cloud	Different model family	Different family, different jurisdiction
	the model's training			across families)	
Correlated reasoning error	Same wrong answer on a hard case	Does not mitigate	Does not mitigate	Partially	Partially

A fallback to a different model from the *same* developer sits between these columns. It gives partial model-level redundancy: as the next section shows, it survived the June 2026 model-specific suspension, but it would not survive a developer-wide policy change or an order directed at the developer itself. Only a fallback from a different developer, and ideally one subject to a different jurisdiction, offers redundancy against all of the model-level failure classes in the table.

The June 2026 Suspension as a Cross-Hosting Test Case

The suspension of Anthropic's Fable 5 and Mythos 5 in June 2026 is the clearest real-world example of a model-level failure that cloud diversity could not absorb. On June 12, 2026, Anthropic received a U.S. government export-control directive barring access to the two models by any foreign national, inside or outside the United States, including Anthropic's own foreign-national employees. Because it had no reliable way to verify nationality in real time, the company said the net effect required it to "abruptly disable Fable 5 and Mythos 5 for all our customers to ensure compliance" [6], while all other Anthropic models stayed available [6][7]. According to Anthropic, the directive came after the government became aware of a report in which Amazon researchers had found a method of bypassing Fable 5's safeguards, and access was restored only after Anthropic deployed a new safety classifier [8]. Anthropic also states that Fable 5 and Mythos 5 share the same underlying model [8], so two separately branded products failed together.

The suspension covered the cloud channels as well as Anthropic's own platform. In announcing redeployment on June 30, Anthropic said it would "re-enable access on AWS, Google Cloud, and Microsoft Foundry as quickly as possible," after first restoring Fable 5 on its own Claude Platform [8]. A field report from Classmethod documents that access to the two models on Amazon Bedrock was suspended on June 12 because of the export controls and resumed on July 1 after the controls were lifted on June 30 [9], and Amazon's own announcement describes Fable 5 as "back on Amazon Bedrock" [10].

Restoration on cloud channels therefore followed first-party restoration, which means the cross-hosted path was not the first to return. Restoration also happened in stages and on terms set by the regulator. On June 26 the Commerce Department cleared Mythos 5 for roughly 100 U.S. companies and federal agencies [11], and Anthropic reported that expansion to its broader set of domestic and international partners remained subject to ongoing coordination with the government [8].

The event combined two model-level failure classes from the table above, a shared safeguard weakness and a government-directed suspension, and changing clouds could absorb neither of them. The hosting cloud had no independent ability to keep the model running, because the binding constraint was the model's safeguards and the legal status of the model and its developer, not any property of the infrastructure. An enterprise whose primary path was the Anthropic API and whose "resilient" fallback was the same model on Bedrock, Vertex AI, or Foundry lost every path at the same moment. Organizations positioned to keep operating would have been those able to fall back to a different model, whether an earlier Anthropic model that was unaffected or a different vendor's model, and whose prompts, evaluations, and output parsers were already validated against that fallback.

It would be a mistake to treat this as unique to one vendor. Any frontier model whose capabilities fall within a government's definition of dual-use technology, or whose developer is subject to a single jurisdiction's directives, carries a comparable exposure, though the trigger and scope of any restriction will depend on the jurisdiction and circumstances. Cross-hosting widens the set of channels that a single directive affects.

Model Diversity Is Weaker Than It Looks

The obvious response is to diversify across model families instead of clouds. That is necessary, but research suggests it gives less independence than intuition would predict. In a study of more than 350 large language models presented at ICML 2025, Kim and colleagues found substantial correlation in model errors. On the HELM leaderboard, pairs of models gave the same answer about 60% of the time when both were wrong. Shared developer and shared base architecture increased correlation, and more accurate models had more correlated errors even when their architectures and providers differed [12]. The finding builds on earlier work on algorithmic monoculture. Kleinberg and Raghavan showed theoretically that independent decision-makers converging on the same algorithm can lower overall decision quality [13]. Bommasani and colleagues then found empirically that sharing training data reliably increases outcome homogenization, so that the same individuals are more likely to be failed across systems, while sharing a model produced mixed results [14].

For security architects, the implication is that an ensemble or fallback built from two frontier models is not equivalent to two independent controls. Kim and colleagues studied answer correlation on benchmarks rather than security weaknesses, but separate research on adversarial prompts has shown

that attacks developed against one model often transfer to others, including commercial models from different developers [15]. If a primary model and its fallback share training-data sources, alignment techniques, or architecture, they may therefore share blind spots: the same susceptibility to a class of prompt injection, the same failure on an unusual input distribution, or the same mistaken classification in an LLM-as-judge pipeline. As frontier models converge in capability, this correlation may increase. If the benchmark pattern holds in practice, the models enterprises most want as fallbacks may be the ones whose errors most resemble those of the primary.

Consolidating Channels, Concentrating Exposure

Cross-hosting also changes the concentration picture at the infrastructure layer. When one hyperscaler hosts most of the major frontier model families an enterprise uses (Amazon Bedrock and Microsoft Foundry now each offer both OpenAI and Anthropic models), the enterprise can consolidate its separate model vendors into a single cloud contract, identity plane, and network path. That simplifies governance, but it means one cloud control-plane failure, one compromised cloud identity, or one misconfigured cloud-level guardrail service can disable or degrade every model the enterprise uses simultaneously. In practice, many organizations may be swapping vendor diversity for procurement convenience without registering the change as a risk decision. Financial-sector regulation already treats this dynamic as material. The EU Digital Operational Resilience Act requires financial entities to assess ICT concentration risk and to account for dependencies on a single provider or a small number of closely connected providers [16].

Recommendations

The analysis above points to a sequence of actions, starting with how AI dependencies are inventoried and ending with how diversification is governed at the portfolio level. The common thread is that each workload's resilience should be judged by what its fallback path shares with its primary path, not by how many clouds it touches.

Immediate Actions

Organizations should start by rebuilding their AI dependency inventory around three separate attributes for each production workload: the hosting channel (the cloud, region, and service through which inference is consumed), the model lineage (the developer and specific model whose weights are executed), and the governing authority (the jurisdictions and vendor policies that can lawfully restrict access). Any workload whose primary and fallback paths run the same model should be reclassified as

having no model-level redundancy, however many clouds it spans. A workload whose fallback is a different model from the same developer should be recorded as having only partial redundancy, since it remains exposed to developer-wide policy changes and to orders directed at the developer. Business continuity plans should also add a "model withdrawn everywhere" scenario next to the regional-outage scenarios most teams already exercise, using the June 2026 suspension as the reference case.

Short-Term Mitigations

Over the next one to two quarters, teams should validate at least one fallback model from a different developer for each business-critical AI workload. Validation should cover the same evaluation suite used for the primary model, not just a functional smoke test. Because prompts, tool-calling schemas, and output parsers are often tuned to one model's behavior, this work belongs in the engineering backlog rather than in a contract clause. An abstraction layer or AI gateway that can route between model families helps only if the fallback route is exercised regularly in production traffic or in scheduled game days. A route that is never exercised should be treated as unverified.

Teams should also measure correlation directly instead of assuming independence. Where two models are used together, for example as generator and verifier, as an LLM-as-judge pair, or as primary and failover, the organization should run both against a held-out set of hard cases from its own domain and record how often they fail on the same items. High agreement on failures signals that the second model adds less assurance than the architecture diagram suggests, and that a non-LLM control (deterministic validation, human review, or a rules-based check) may be needed for the highest-risk decisions.

Strategic Considerations

At the portfolio level, procurement and risk functions should treat model lineage and governing jurisdiction as diversification dimensions in their own right, alongside cloud provider. For some regulated or mission-critical workloads, that may justify keeping an open-weight model that the enterprise can host itself as a last-resort path. An open-weight model cannot be technically withdrawn by the vendor once the weights are held, although license terms and export controls may still constrain its use, and it carries its own security and maintenance burden. Contract negotiations with cloud providers and model developers should ask specifically what notice, transition period, and alternative access are provided if a model is suspended for regulatory reasons, and whether the cloud provider has any independent continuity commitment or is purely a pass-through. Finally, boards and risk committees should be told plainly that "available on every cloud" describes distribution reach and infrastructure resilience. It is not model-level resilience.

CSA Resource Alignment

CSA's published guidance gives a foundation for putting these recommendations into practice, though cross-hosting raises a model-lineage question that existing control language does not yet address directly. The **AI Controls Matrix (AICM) v1.1** [17] provides control objectives for supply chain management, transparency and accountability, and business continuity and operational resilience. Organizations can apply these to AI dependencies by recording model lineage and governing authority, not just the contracting party, in supplier inventories, and by requiring that continuity controls be tested against model-level withdrawal. A natural next step for AICM implementers is to treat "same model, different host" explicitly as a single point of failure in supply chain assessments.

CSA's **AI Model Risk Management Framework** [18] organizes model risk around Model Cards, Data Sheets, Risk Cards, and Scenario Planning. Model Cards and Data Sheets give a structured place to record a model's developer, training lineage, and known limitations, which is the information needed to judge whether a fallback model is meaningfully independent. The framework's Scenario Planning pillar is a suitable vehicle for the "model withdrawn everywhere" and "correlated failure across model families" scenarios described in this note.

CSA's **Zero Trust Guidance for Achieving Operational Resilience** [19] connects Zero Trust architecture to operational resilience. It covers business impact analysis, resilience maturity models, and regulatory obligations including DORA. Its approach of mapping critical business services to their dependencies, then testing recovery against realistic disruption, applies directly to AI workloads. The finding here is that the dependency map has to go down to the model rather than stopping at the cloud service.

This note also extends earlier CSA research on frontier-model concentration. CSA's analyses of the June 2026 suspension, *Sovereign AI Access Controls and Enterprise Concentration Risk* [20] and *The Fable 5 Precedent: AI Models Under Export Control* [21], examine the incident from the supply-chain and export-control perspectives, while *Sovereign AI Access Controls and Frontier Model Dependency Risk* [22], *AI Provider Concentration Risk: Enterprise Resilience* [23], and *The Triple AI Outage: A Wake-Up Call for AI Continuity* [24] address enterprise dependency on frontier model providers and continuity planning. The present note adds the cross-hosting dimension to that body of work. Organizations that rely on cloud-hosted frontier models should also include these questions in third-party assurance reviews conducted through CSA's [STAR program](#), asking providers to describe how they would handle a regulatory suspension of a hosted third-party model.

References

- [1] Microsoft. "[The next phase of the Microsoft-OpenAI partnership.](#)" Official Microsoft Blog, April 27, 2026.
- [2] Amazon. "[OpenAI models on Amazon Bedrock.](#)" About Amazon, updated September 8, 2026.
- [3] Microsoft Azure. "[Claude in Microsoft Foundry is now generally available.](#)" Microsoft Azure Blog, June 29, 2026.
- [4] Shirodkar, V. "[Run Gemini anywhere, including on-premises, with Google Distributed Cloud.](#)" Google Cloud Blog, August 28, 2025.
- [5] Amazon Web Services. "[Summary of the Amazon DynamoDB Service Disruption in the Northern Virginia \(US-EAST-1\) Region.](#)" AWS Post-Event Summaries, October 2025.
- [6] Anthropic. "[Statement on the directive to suspend Fable 5 access.](#)" Anthropic, June 12, 2026.
- [7] CNN Business. "[Anthropic suspends all access to Mythos model after US government bans foreign nationals use.](#)" CNN, June 13, 2026.
- [8] Anthropic. "[Redeploying Fable 5.](#)" Anthropic, June 30, 2026.
- [9] Classmethod (suzuki.ryo). "[I tested the re-released 'Claude Fable 5' on Amazon Bedrock.](#)" DevelopersIO, July 1, 2026.
- [10] Amazon. "[Claude Fable 5 is back on Amazon Bedrock.](#)" About Amazon, July 2026.
- [11] CNBC. "[Trump admin allows Anthropic to release Mythos AI model to some companies, government agencies.](#)" CNBC, June 26, 2026.
- [12] Kim, E., Garg, A., Peng, K., and Garg, N. "[Correlated Errors in Large Language Models.](#)" Proceedings of ICML 2025 (arXiv:2506.07962), June 2025.
- [13] Kleinberg, J., and Raghavan, M. "[Algorithmic monoculture and social welfare.](#)" Proceedings of the National Academy of Sciences, Vol. 118, No. 22, May 2021.
- [14] Bommasani, R., Creel, K., Kumar, A., Jurafsky, D., and Liang, P. "[Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?](#)" NeurIPS 2022 (arXiv:2211.13972), November 2022.

- [15] Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. "[Universal and Transferable Adversarial Attacks on Aligned Language Models](#)." arXiv:2307.15043, July 2023.
- [16] European Union. "[Regulation \(EU\) 2022/2554 on digital operational resilience for the financial sector \(DORA\)](#)." Official Journal of the European Union, December 2022.
- [17] Cloud Security Alliance. "[AI Controls Matrix \(AICM\) v1.1](#)." CSA, June 2026.
- [18] Cloud Security Alliance. "[AI Model Risk Management Framework](#)." CSA, July 2024.
- [19] Cloud Security Alliance. "[Zero Trust Guidance for Achieving Operational Resilience](#)." CSA, April 2026.
- [20] Cloud Security Alliance. "[Sovereign AI Access Controls and Enterprise Concentration Risk](#)." CSA Lab Space, 2026.
- [21] Cloud Security Alliance. "[The Fable 5 Precedent: AI Models Under Export Control](#)." CSA Lab Space, 2026.
- [22] Cloud Security Alliance. "[Sovereign AI Access Controls and Frontier Model Dependency Risk](#)." CSA Lab Space, 2026.
- [23] Cloud Security Alliance. "[AI Provider Concentration Risk: Enterprise Resilience](#)." CSA Lab Space, 2026.
- [24] Cloud Security Alliance. "[The Triple AI Outage: A Wake-Up Call for AI Continuity](#)." CSA Lab Space, September 2026.