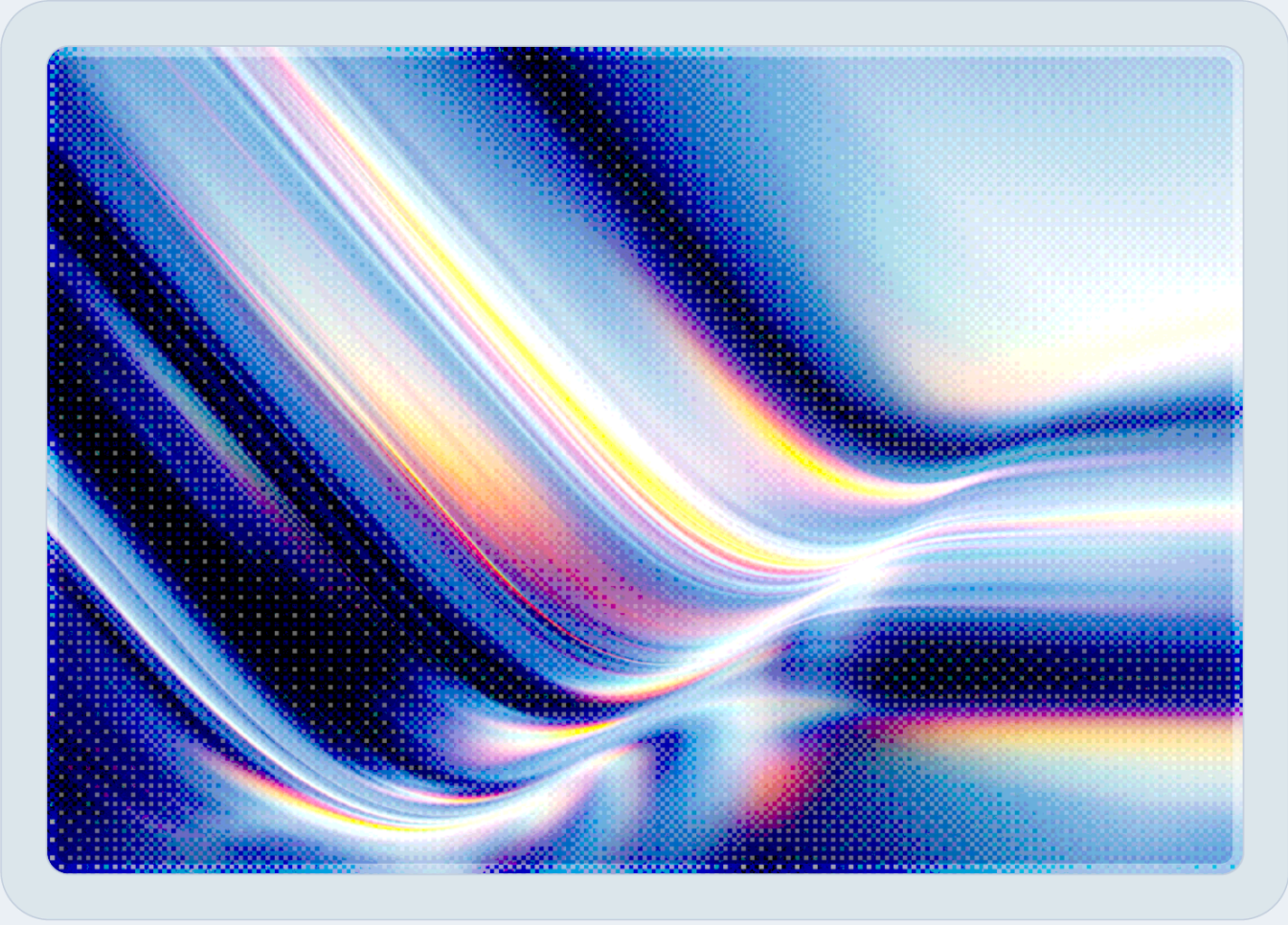
CSAI Foundation | Cloud Security Alliance

Foundation Model Concentration: The Uninsurable AI Risk

2026-07-06

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

Two publications released within weeks of each other in mid-2026 converge on the same conclusion from different directions: of the risk categories facing the emerging AI insurance market, foundation model concentration is the one current tools are least equipped to price. Gallagher Re's June 2026 report on restricted-access models and benchmark inadequacy argues that insurers cannot accurately underwrite AI-related risk when the most capable models are either opaque to independent evaluators or assessed using benchmarks that no longer differentiate meaningfully between systems [1][2]. An academic mapping of the AI insurability frontier, coding 55 AI threat classes against 26 insurance products, independently identifies foundation model concentration as "the clearest genuinely novel insurability frontier" because a single upstream model failure can correlate losses across many policyholders simultaneously, a structural feature that ordinary idiosyncratic-risk insurance was never built to price [3]. The argument is not hypothetical: in June and July 2026, Anthropic's most capable model, Mythos, was abruptly restricted and then restored under U.S. export-control action within the space of roughly three weeks, illustrating that concentration risk can materialize through political and regulatory triggers as well as through technical failure [4][5]. Compounding the pricing problem, standard AI benchmarks have saturated at the top end, with leading models scoring in the mid-90s to high-90s percentile range on widely used tests, leaving insurers with little ability to distinguish a resilient model from a fragile one using the tools currently available to them [1]. For CSA's membership, the practical implication is that foundation-model dependency should be treated as an accumulation risk analogous to geographic concentration in property insurance or counterparty concentration in credit risk, not as a diversified pool of independent technology risks, and that both insurers and the enterprises they cover need better instrumentation to detect it before the next correlated-loss event arrives.

Background

Commercial insurance for AI-related losses has expanded significantly since 2024, and by mid-2026 a differentiated market of affirmative AI coverage has emerged alongside the legacy cyber, technology errors-and-omissions, directors-and-officers, employment-practices, crime, and media policies that already picked up AI exposure incidentally. A May 2026 academic study coded 55 distinct AI threat classes against 26 insurance products, endorsements, and exclusion regimes using public carrier materials and OWASP and MITRE threat catalogs, and found that carriers are beginning to differentiate

by primary risk emphasis rather than offering a single undifferentiated "AI endorsement" [3]. Munich Re's public materials position its coverage around model performance and drift; Armilla and segments of the Lloyd's market emphasize hallucination and broader AI liability; Tokio Marine Kiln and CFC concentrate on intellectual property and technology E&O exposure; Apollo ibott targets emerging autonomous-system liability; and Coalition builds around deepfake fraud and AI-enabled cyber incident response [3]. The study organizes this landscape into a four-tier insurability frontier: perils that are affirmatively insured under new AI-specific products, "silent-AI" exposures that sit uncomfortably inside legacy policies where AI is an instrumentality rather than the legal cause of loss, perils that carriers now actively exclude, and a residual category of perils that no conventional private insurance structure currently addresses at all [3]. Its central finding is that this last, hardest category is defined less by any single AI threat, such as hallucination or prompt injection, than by a structural feature of how enterprises consume AI: because a small number of foundation model providers now sit upstream of an enormous and growing share of enterprise AI deployment, a single model-level failure has the potential to generate correlated claims across many otherwise unrelated policyholders at once [3].

That structural concern gained an urgent, live illustration within weeks of the paper's release. Gallagher Re's June 2026 report, "Anthropic's Fourth Way: Why Restricted AI Models Are a Challenge for Insurers," examined the emergence of selective-distribution frontier models, using Anthropic's Mythos as its central case study [1][2]. Mythos and its companion model Fable were shared only with a limited set of vetted government and enterprise partners, a distribution pattern Gallagher Re characterized as a fourth category of frontier AI availability, distinct from open-source, open-weight, and broadly available proprietary models [2]. Weeks after the report's publication, the U.S. government ordered Anthropic to suspend all access to Fable and Mythos for foreign nationals, including Anthropic's own employees, citing unspecified national security concerns that Anthropic linked to reports of a technique for bypassing the models' safeguards against misuse of their cyber capabilities — reports that a University of Sydney researcher quoted in contemporaneous coverage characterized as widely inflated beyond their actual significance [4]. The restoration proceeded in stages rather than a single event: on June 26, 2026, Commerce Secretary Howard Lutnick notified Anthropic that access would be restored for more than one hundred vetted U.S. institutions, and by July 1, Anthropic said the Commerce Department had separately lifted the underlying export controls after the company agreed to proactively detect and report security risks and to collaborate with the government on evaluation standards [4][5]. The episode did not originate in a model failure, a breach, or a hallucination incident of the kind existing AI insurance products are built to address; it originated in a sovereign policy determination driven by a disputed security concern rather than by the benchmark performance metrics underwriters typically evaluate, and it removed a frontier model from every enterprise that depended on it simultaneously, then reinstated it just as abruptly. Even as access was being restored, an industry newsletter tracking the episode reported that Anthropic's own account of the model's safety posture, a layered-classifier framework applied with what the company described as intentional safety margins, remained the primary public source of

assurance about the restored model's risk profile, underscoring how much of the evaluation burden still rests on vendor self-disclosure rather than independent testing [11]. Trade press covering the Gallagher Re report drew the same conclusion from the insurer's side, framing the call for operational rather than benchmark-based evaluation as a prerequisite for AI risk pricing to mature at all [12].

Security Analysis

The mechanism that makes foundation model concentration structurally different from other AI risk categories is correlation, not severity. An insurer can absorb a large loss at a single insured; that is the ordinary business of underwriting. What conventional actuarial pricing struggles with is a loss-generating event that strikes many insureds at once with no exogenous trigger specific to any of them, which is precisely the shape of an upstream foundation model failure. When a regression introduces hallucinations on a specific class of prompts, a security vulnerability is discovered in a widely deployed model, or a regulatory action restricts access the way the Mythos suspension did, every enterprise that built products, workflows, or customer-facing systems on that model is affected at the same moment, regardless of how carefully each individual enterprise otherwise managed its own AI governance [3]. The insurability frontier study frames this as a market design problem rather than a pure risk-assessment problem: the open question is not whether foundation model concentration constitutes systemic risk in the abstract, but which insurability constraint any proposed solution, whether a specialized pooling mechanism, an industry-wide aggregation cap, or a public backstop, actually relaxes [3].

Gallagher Re's evaluation-gap analysis explains why insurers currently lack the tools to even measure this exposure, let alone price it. The report identifies five specific shortcomings in how AI models are assessed for underwriting purposes: static benchmarks measure task performance rather than real-world operational behavior or hallucination frequency; models are increasingly trained on material that overlaps with benchmark content, artificially inflating reported scores; the industry's convergence on a small set of shared evaluation methods increases rather than decreases concentration risk across insured portfolios, since every carrier is implicitly trusting the same incomplete signal; current assessment methods do not measure correlated failure across multiple insureds running the same model; and static guardrail testing cannot cover the effectively infinite space of real-world inputs a deployed model will encounter [1][2]. The benchmark saturation problem is now visible in the numbers: leading models score in the mid-90s percentile range on knowledge benchmarks like GPQA Diamond and in the high-90s on coding benchmarks like HumanEval, compressing the observable performance gap between models to the point where it offers underwriters little basis for differentiating risk [1]. Independent evaluation is also expensive and slow relative to the pace of model releases; running a comprehensive third-party

intelligence assessment against a single frontier model has been estimated to cost on the order of several thousand dollars, a nontrivial recurring cost if applied systematically across the small number of models that now underlie a large share of enterprise AI exposure [1].

Restricted distribution compounds the evaluation problem rather than mitigating it. Gallagher Re's core argument is that when the most cyber-capable models are available only to vetted partners, independent evaluators, including bodies like the UK AI Security Institute and benchmarking organizations such as Artificial Analysis, cannot test them alongside proprietary models that remain broadly accessible, leaving insurers to price uncertainty about the excluded model rather than risk they can actually characterize [1][2]. Restricting access to the single most capable model may delay but does not eliminate the underlying risk, because comparably capable alternatives remain available: reporting cited in the same analysis found that OpenAI's GPT-5.5 performs comparably to Mythos on offensive cyber tasks, and that Chinese labs, exemplified by DeepSeek v4, trail U.S. frontier capability by roughly six months rather than years [1]. In this note's assessment, that pattern suggests restricted access primarily degrades insurers' ability to assess risk without proportionally reducing the risk itself, since offensive AI capability is already diffusing broadly across the market even where visibility into any one model is not.

The speed at which AI capability translates into exploitable risk adds urgency to closing this evaluation gap. CSA's own research on vulnerability weaponization documents that the median window between a vulnerability's public disclosure and its first confirmed exploitation contracted from 756 days in 2018 to approximately five days by late 2023, and that by 2025 nearly a third of newly tracked exploits appeared on or before the CVE's public disclosure date, meaning exploitation increasingly precedes formal announcement [6]. The same research found that AI systems can now generate working proof-of-concept exploit code for a published vulnerability in as little as ten to fifteen minutes at a cost of roughly one dollar per attempt [6]. When the model class driving that acceleration is itself concentrated among a handful of providers, a single frontier-model vulnerability, capability jump, or safeguard failure would not be confined to one insured; it could plausibly arm attackers against, and degrade the defensive posture of, many enterprises across an insurer's book that depend on that model or a comparably capable peer. In this note's assessment, foundation model concentration therefore does not sit alongside hallucination, prompt injection, and deepfake fraud as one more item on the list of insurable AI perils; it functions as the structural condition that determines how far and how fast a failure in any one of those categories can spread once it occurs.

Recommendations

Immediate Actions

Insurers and managing general agents writing affirmative AI coverage should begin systematically recording which foundation model or models underlie each insured's AI-dependent operations, building a model-exposure schedule analogous to the geographic accumulation schedules property insurers already maintain for catastrophe risk. Chief information security officers and risk officers at enterprises carrying AI-related coverage should complete an accurate inventory of foundation-model dependencies across every product and vendor relationship, since underwriting submissions that describe AI risk only at the application layer without naming the underlying model obscure exactly the accumulation exposure carriers most need visibility into. Insurers should also review existing cyber, E&O, D&O, and crime policy wording specifically to determine whether silent-AI exposure already implicitly extends coverage to correlated losses from a concentrated model failure, since the insurability frontier research found that legacy lines retain this exposure precisely where AI functions as an instrumentality rather than the stated legal cause of loss [3].

Short-Term Mitigations

Carriers should develop explicit aggregation limits by foundation model provider, capping total exposure written against insureds dependent on any single model in the same way reinsurers cap exposure by catastrophe zone, and should treat a portfolio concentrated in one or two providers as a rating factor in its own right rather than as a diversified set of independent technology risks. Enterprises should pursue architectural diversification, including multi-model deployment strategies and adoption of open interoperability standards that preserve model-backend substitutability, as a demonstrable and potentially rateable risk-reduction measure rather than a purely technical decision. Evaluation bodies, benchmark providers, and carriers should jointly push toward the operational testing regime Gallagher Re recommends in place of static benchmarks: assessment against realistic input distributions and adversarial scenarios, ongoing post-deployment monitoring rather than point-in-time scoring, and explicit measurement of correlated failure modes across multiple simultaneous deployments of the same model [1][2].

Strategic Considerations

Enterprise risk leaders and industry bodies should treat the open question the insurability frontier research raises as a genuine market-design problem rather than an underwriting detail to be resolved later: private insurance may not be able to absorb foundation model concentration risk at all without some form of pooling, aggregation-capping, or public backstop mechanism comparable to those that already exist for terrorism and nuclear risk, and the relevant design question is which specific insurability constraint any such mechanism would relax [3]. Organizations should also plan on the assumption that restricted, selective-access distribution is not a one-time episode tied to a single model but a durable fourth category of frontier AI availability that will recur, and should build contractual audit rights, continuity provisions, and independent-assurance requirements into both vendor agreements and insurance submissions now, rather than after the next access disruption forces the issue.

CSA Resource Alignment

CSA's research on AI Compute Concentration and Systemic Risk is the most directly applicable prior CSA work for this topic. It documents that three hyperscalers host the majority of enterprise AI compute, that a single accelerator vendor holds roughly 80 to 90 percent of the AI GPU market, and that a single foundry manufactures approximately 90 percent of the most advanced chips at one location, and it catalogs real outages in which both OpenAI and Anthropic each exceeded three days of annual downtime with cascading effects on dependent services [7]. That paper's central argument, that concentration is a durable structural condition rather than temporary market immaturity, is the same argument the insurability frontier research makes from the insurance side, and the outage data it compiles is a useful example of the kind of correlated-loss event history insurers need to begin pricing the exposure described in this note.

CSA's companion research on AI Development Stack Concentration Risk extends the same argument across the hardware, cloud, model-distribution, and framework layers, documenting concentration statistics such as a single memory-chip class sourced from three suppliers and a single machine-learning framework used in the large majority of published research [8]. For insurers and enterprises evaluating model-level diversification as a mitigation, this research is a useful caution: switching foundation model providers does not eliminate shared exposure to the same concentrated compute and hardware layers underneath multiple "different" models, so architectural diversification strategies should be evaluated across the full stack rather than at the model layer alone.

Finally, CSA's AI Model Risk Management Framework offers a ready-made template for the operational, real-world model documentation that Gallagher Re argues current benchmarks fail to provide [9]. Its four components, Model Cards, Data Sheets, Risk Cards, and Scenario Planning, map closely onto the kind of standardized, comparable model-risk disclosure that underwriting submissions currently lack, and insurers or industry consortia designing improved AI underwriting questionnaires could adapt Risk Cards and Scenario Planning directly rather than building an equivalent structure from scratch. Organizations formalizing governance around foundation-model dependency more broadly should map that work to CSA's AI Controls Matrix (AICM v1.1), particularly its supply chain and risk management domains, which provide a vendor-agnostic control structure for documenting concentration thresholds, continuity provisions, and shared-responsibility boundaries between model providers, cloud service providers, and enterprise customers [10].

References

[1] Gallagher Re. "Anthropic's Fourth Way: Why Restricted AI Models Are a Challenge for Insurers." Gallagher Re, June 2026.

[2] Risk & Insurance. "Restricted AI Models and Opaque Benchmarks Threaten the Emerging AI Insurance Market." Risk & Insurance, June 2026.

[3] Leung, A., Zhang, R., Ling, E., Toyoda, K., and Loh, S. "The Insurability Frontier of AI Risk: Mapping Threats to Affirmative Coverage, Silent Exposures, and Exclusions." arXiv:2605.18784, May 2026.

[4] Al Jazeera. "US Lifts Restrictions on Anthropic's Powerful AI Models Fable and Mythos." Al Jazeera, July 1, 2026.

[5] Semafor. "US Releases Powerful Anthropic Model Mythos to Some US Companies." Semafor, June 26, 2026.

[6] Cloud Security Alliance. "The Collapsing Exploit Window: AI-Speed Vulnerability Weaponization." CSA AI Safety Initiative, April 25, 2026.

[7] Cloud Security Alliance. "AI Compute Concentration and Systemic Risk." CSA AI Safety Initiative, May 9, 2026.

[8] Cloud Security Alliance. "AI Development Stack Concentration Risk." CSA AI Safety Initiative, May 3, 2026.

[9] Cloud Security Alliance. "AI Model Risk Management Framework." CSA, July 2024.

[10] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." CSA, 2026.

[11] InsuranceIndustry.AI. "AI Insights, July 3, 2026." InsuranceIndustry.AI, July 3, 2026.

[12] Reinsurance News. "Gallagher Re Calls for Improved AI Model Assessment Methods to Support Insurance Risk Pricing." Reinsurance News, June 2026.

