

Operationalizing AI Risk Governance for Multi-Tenant SaaS: A Deployment-Specific Translation of NIST CSF v2.0 and ISO/IEC 42001

Nazarii Romankiv 

*Kharkiv National University of Radio Electronics
Kharkiv, Ukraine*

ayzrian@gmail.com

Dmytro Sytnikov 

*Kharkiv National University of Radio Electronics
Kharkiv, Ukraine*

Corresponding Author: Nazarii Romankiv

Copyright © 2026 N. Romankiv and D. Sytnikov. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

The 2024 wave of AI risk instruments (NIST AI 600-1, ISO/IEC 42001, the EU AI Act) and the CSF v2.0 update establish what an organization must govern, but each is deployment-agnostic and leaves the multi-tenant SaaS engineering layer to the implementer. This paper builds and tests a deployment-context profile for it. Seventeen technical vulnerability classes are recut into five Risk Domains carrying accountable owners, board-visible indicators and regulatory touchpoints, and the lifecycle is decomposed into twenty control families. A multi-tenancy amplification factor, split into an architectural bound and a realisation coefficient, supports a Monte Carlo treatment of platform loss. Evaluation uses 30 publicly disclosed incidents from MITRE ATLAS and vendor advisories, January 2023 to January 2026, with Wilson intervals on proportions and a paired bootstrap over incidents on coverage scores; what each construct claims and what would refute it is fixed before any result is reported. The headline comparison ranks eight existing instruments and excludes our own profile, since we scored all of them: the CSA AI Controls Matrix leads at 4.00 of 4 and COBIT 2019 trails at 0.87. Twenty-nine of the 30 incidents map onto the underlying taxonomy, and vendor and supply-chain control is the most frequently decisive control family at 53%, ahead of output sanitisation at 50%. Our profile reaches 3.87 on the same basis, and third place on the externally sampled subset; what separates it is the governance layer, at 6 of 6 on owner assignment, board indicators and cross-tenant amplification against a best comparator score of 4. Two constructs are reported as untested rather than defended. The assessment questionnaire, scoring code and coded dataset are released so every result can be recomputed or contested.

Keywords: AI risk governance, Cyber-insurance, Empirical validation, ISO/IEC 42001, LLM security, Maturity model, Multi-tenant SaaS, NIST CSF v2.0, Tenant isolation

1. INTRODUCTION

Cyber risk on a corporate risk register inherited from the 2010s has a recognizable shape: network, application, third-party and personnel sub-categories, each with an owner, an appetite statement, a score and a residual-risk acceptance line reviewed on a fixed cycle [1–4]. Large language models do not sit in any of those boxes. A retrieval-augmented copilot touches application security through the orchestration layer, third-party risk through the model vendor and data privacy through whatever the model is shown, while adding a failure with no historical analogue: natural-language reasoning subverted into outputs that look authoritative and are adversarially controlled. AI risk therefore tends to appear on a Chief Risk Officer’s worksheet as one line under emerging technology, or not at all.

Multi-tenancy, in which one application instance and its infrastructure serve many customer organizations [5, 6], sharpens the problem. A failure of input filtering in a single-tenant deployment produces one organization’s incident; the same failure on a shared platform can reach every customer at once, through a poisoned document in a shared vector store [7, 8], a malicious tool definition in a shared Model Context Protocol (MCP) server [9], a cache-timing side channel in shared inference infrastructure [10], or an isolation failure in shared serving, session and logging components [11]. Our prior work [12], catalogues 17 such vulnerability classes and finds most of them worse under multi-tenancy, and our work on agents over SaaS codebases [13], and on context engineering for SaaS agents [14], shows how widely such agents reach into tenant data. The unit of risk becomes the platform rather than the customer engagement: one domain failure can breach service commitments to hundreds of tenants, trigger notification duties in several jurisdictions, and exhaust an aggregate insurance limit in a single event [15, 16].

Authoritative guidance exists: NIST AI 600-1 enumerates twelve generative-AI risk categories [17], ISO/IEC 42001 specifies an organization-level AI management system [18], the EU AI Act sets a horizontal regulatory floor tiered by use case [19], and CSF v2.0 adds a *Govern* function while inviting contextual profiles [20]. Each is scoped deliberately, and each leaves the same question open: how a particular deployment architecture should put them into practice together.

This paper answers that question for multi-tenant LLM-SaaS and then tests the answer. Seventeen technical vulnerability classes from [12], are recut into five Risk Domains, each with an accountable-owner archetype, board-visible indicators and regulatory touchpoints, and the CSF v2.0 lifecycle is regrouped into five ERM-facing pillars holding twenty control families calibrated for shared LLM infrastructure. A multi-tenancy amplification factor separates what the architecture permits from what an attacker realises and feeds a Monte Carlo model of platform-level loss; a five-level maturity instrument, three derived indicators, an assessment questionnaire and scoring code complete the set.

Evaluation uses 30 publicly disclosed incidents from MITRE ATLAS and vendor advisories, coded against the artefact and against eight comparator instruments. It is retrospective and observational rather than a controlled trial: it does not show that adopting the profile reduces loss, only which instrument articulates a given real failure and supplies a control for it. Because the quantitative constructs carry very different evidential weight, Section 3.4 states for each what is tested, against what data, with which statistic, and what would refute it; two of them rest on assumption rather than evidence, and the paper reports them as such. Section 3 gives the derivation and coding procedure,

Section 4 the artefact and its mathematics, Section 5 the results, and Section 6 what the numbers support and what they do not.

2. RELATED WORK AND POSITIONING

2.1 Source Instruments

COSO ERM [1], and ISO 31000 [2], supply principles-based scaffolding into which technology-specific content is plugged, Aven's treatment of risk as consequence plus uncertainty [21], and Renn's separation of governance from management [22], give the Govern pillar its theoretical footing, and on the cybersecurity side NIST SP 800-37 [4], contributes a seven-step control-application discipline and SP 800-30 [23], the likelihood-times-impact rubric adapted in our Assess pillar. CSF v2.0 [20], is the lifecycle backbone profiled here.

Among AI-specific instruments, NIST AI RMF 1.0 [24], organizes guidance around Govern, Map, Measure and Manage, and AI 600-1 [17], refines it for generative systems. ISO/IEC 42001 [18], shadows rather than displaces engineering controls: the AIMS is a management shell inside which control families are applied, and whose operation is verified by independent audit [25], layered for LLMs in [26]. Above these sit the OECD AI Principles [27], and the G7 Hiroshima guiding principles [28], which set expectations without prescribing controls. The EU AI Act [19], which layers onto existing sectoral law as [29], shows for medical AI, reaches non-EU vendors selling into the Union, and composite uses such as LLM-assisted credit decisioning inside a CRM can attract obligations the same feature would not attract alone. At the application layer, the OWASP Top 10 for LLM Applications [30], names deployment vulnerabilities directly. Two 2025 instruments complete the list. The CSA AI Controls Matrix [31], released in July 2025, carries 243 control objectives across 18 domains for cloud-based AI systems and allocates them across the roles of the AI supply chain, which makes it the closest control-level comparator to this work. ISACA's 2025 guidance on applying COBIT to AI systems [32], places AI risk under EDM03 and APO12 and board-level oversight under EDM05, giving COBIT a board-reporting path that the AI-specific instruments lack.

2.2 Critical Reading of 2025–2026 Governance Research

Governance research from 2025 and 2026 clusters into five groups, and the profile developed here sits differently against each.

Operationalization frameworks. Gupta's AI TIPS 2.0 [33], argues that ISO 42001 and NIST AI RMF stay conceptual and answers with use-case-level controls; Koch (2026), [34], assigns standards-derived objectives to control layers through a control tuple and an enforceability rubric, reserving runtime guardrails for the enforceable subset; Morgan (2026) [35], stacks six governance layers for operational stability. All three share our diagnosis and differ in target, organising by use case, control layer and traceability layer respectively, none by deployment architecture. A shared vector namespace and a per-tenant one are therefore indistinguishable in each, although they differ by two orders of magnitude in how many customers one failure reaches; that distinction is the axis

this paper adds. Koch's rubric is nonetheless the sharper instrument for deciding *where* a control belongs.

Maturity models. Acharya's AAGMM (2026) [36], is the closest competitor: five levels, twelve governance domains, the same two standards, aimed at agent sprawl. Its validation rests on 750 simulation runs over synthetic scenarios rather than on observed cases, which is the evidential gap Section 5 addresses here by scoring against a disclosed-incident corpus instead. Surve et al.'s CEDAR-42001 [37], converts ISO 42001 audit evidence into an architecture-aware posture with a five-dimensional maturity profile; their strata are those of cyber-physical systems and ours those of shared SaaS infrastructure, so the two complement rather than compete. CEDAR-42001 also derives a risk-proportionate target maturity instead of asserting one, a discipline Section 4.4 adopts through the Residual Amplification Exposure.

Empirical incident analysis. Mumtaz et al. (2026) [38], code AI Incident Database entries from 2020–2026 against nine post-deployment provisions of the EU AI Act, NIST AI RMF and GDPR, reporting that 77.1% of incidents show no evidence of post-market monitoring. Theirs is the methodological precedent for Section 5: a public corpus, an explicit coding frame, coverage as a proportion. The unit differs, since they ask whether regulatory obligations were discharged and we ask whether an instrument names the failure and supplies a control. Huwyler (2025) [39], organises AI risks into nine domains and 53 sub-threats for quantitative assessment, benchmarked against MITRE ATLAS and the OWASP list; those domains are cut by harm type and ours by accountable owner.

Agentic and infrastructure security. Errico et al. (2025) [40], set out MCP-specific controls and state that NIST AI RMF and ISO 42001 do not yet cover this surface, and Tallam (2026) [41], argues that authorization propagation across delegating agents is distinct from prompt injection, which our RD3 under-specifies. SpliceLeak [42], demonstrates that non-prefix KV-cache fusion leaks across tenants in a vLLM serving stack with LMCache and PrefixWall [43], treats the prefix-caching channel as warranting a dedicated mitigation, so together they support treating shared inference as an amplifying resource; AgentRedBench [44] reports unguarded attack success between 32% and 81% across 24 enterprise SaaS integrations.

Insurance. Zhu surveys the emerging agentic-AI insurance market in [45] and in [46], formalises it: a deployment's risk state includes governance maturity and dependency concentration, from which an insurability region and governance certification thresholds are derived. This matters here because it is the only source that ties a governance maturity score to an underwriting decision through a stated model rather than through market commentary, and Section 6.2 accordingly frames the insurance argument against [45] and [46], rather than against industry reports alone.

2.3 Where This Paper Sits

None of the instruments above was written for the operational engineering of multi-tenant SaaS that wraps LLM components, where shared inference, shared vector storage and shared tool integrations produce failure modes the source documents leave to implementers. TABLE 5, in Section 5 reports what that gap costs in measured terms rather than asserting it.

3. MATERIALS AND METHODS

The artefact was built with a design-science process in the tradition of Hevner et al. (2004) [47]. Design science is only as good as its audit trail, so this section states how each component was derived, which rules constrained the derivation, and how each was checked, before Section 3.4 fixes what the empirical part of the paper is entitled to conclude.

3.1 Derivation of the Risk Domains

The input was the 17-class technical taxonomy of [12], cross-checked against the OWASP LLM Top 10 [30] and the twelve AI 600-1 categories [17]. Three rules governed consolidation: every class lands in exactly one domain; each domain admits one accountable-owner archetype that exists on a typical SaaS organization chart; each domain attaches to at least one regulatory touchpoint so the owner's obligations are legible outside the firm. Seven-, six-, five- and four-domain partitions were drafted, and five is the smallest that satisfies all three. The seven-domain variant put three supply-chain owners inside one procurement function; the six-domain variant gave output handling and content integrity the same owner. Supplementary S1 records the discarded partitions, the rule each violated, and the two class assignments (B1 and B2) where a different coder could reasonably disagree.

3.2 Derivation of the Control Families and the Threat–Control Matrix

The five pillars are CSF v2.0 functions regrouped for ERM vocabulary, with no novelty claimed for the regrouping: Protect sits under Mitigate, Detect, Respond and Recover collapse under Monitor, and Assess is split out of Identify to expose quantitative scoring as its own step. Four control families per pillar were derived by asking, for each pillar and each Risk Domain, what a SaaS provider must do that a single-tenant deployer need not, then merging until every pillar held four.

Cell strengths in the threat-to-control matrix (FIGURE 2), follow a four-level rubric fixed before coding and stated in Supplementary S2: *High* where the control changes likelihood or impact directly and is auditable, *Medium* where it leaves a residual path open, *Low* for indirect coverage, *None* where no defensible relationship exists. The first author coded all 340 cells and the second reviewed them. Both are authors of the source taxonomy, so this establishes internal consistency and not independence; no inter-rater statistic is reported and none should be inferred. The matrix is released as machine-readable data so that a third party can recode it, and Section 4.4 reports a test showing how little the coded weights actually affect the indices derived from them.

3.3 Derivation of the Maturity Criteria

Level labels follow the CMMI [48] and CSF tier vocabulary. Per-pillar criteria for the lower tiers were set by the authors against the AI exposures that insurance-market reports identify [16, 49], and the upper tiers follow the auditable controls of ISO/IEC 42001 [18]. Anchoring means the criteria

were written so that an organization at level 3 holds the evidence those questionnaires request; the levels were not fitted to observed vendor behaviour.

3.4 Validation Design

The artefact is a stack of constructs of very different evidential standing, and treating them uniformly would be the main way to mislead a reader. TABLE 1, therefore fixes, before any result is reported, what each construct claims, what evidence and statistic bear on it, and what observation would refute it. Its upper block is testable against the corpus and Sections 5.2 to 5.4 report it with intervals; its lower block is not tested by anything available here, and where a construct sits there the paper says so in the sentence introducing it, so that no result in Section 5 depends on it.

3.5 Retrospective Validation Corpus

Evaluation needs cases the authors did not construct. The primary sampling frame is the MITRE ATLAS case-study collection [50], a third-party register of 57 documented AI security cases with dated public sources, retrieved on 3 August 2026. Three criteria were applied, all of which must hold: the target places a language or generative model in its request path, excluding classical-ML evasion, biometric spoofing and cases where the model is the attacker's instrument; the target is a shared service or shared artefact consumed by multiple independent organizations, excluding single-user desktop and self-hosted single-organization deployments; and disclosure is dated 1 January 2023 or later. Twenty-five cases qualify. Section 6.3 records where our application of the second criterion is not fully consistent.

A second set of five covers a known gap in ATLAS, namely production multi-tenant SaaS incidents disclosed by vendors rather than researchers, which are the cases carrying information about how many customers were reached: the 2023 ChatGPT shared-cache exposure [51], the Asana MCP cross-tenant exposure [52], EchoLeak [53], ForcedLeak [54], and the ServiceNow Now Assist case [55]. These five were selected by us after the rule table was written, so Section 5.3 reports the comparison with and without them, and the corpus is 30 incidents spanning January 2023 to January 2026. Each incident was coded for its technical mechanism, the classes of [12], it instantiates, the Risk Domains it touches, its amplification group, and the control families whose operation would most plausibly have prevented or bounded it; the last is a counterfactual judgement about systems we did not operate, made with hindsight from the vendor's own account, and is reported with that status.

3.6 Comparative Coverage Scoring

Eight comparator instruments were scored against the same 30 incidents: NIST AI RMF 1.0 with AI 600-1, ISO/IEC 42001, ISO/IEC 27001:2022 Annex A, COBIT 2019 with ISACA's AI guidance, the OWASP LLM Top 10 (2025), MITRE ATLAS, the CSA AI Controls Matrix v1.0, and NIST CSF v2.0.

Table 1: Validation design. Each construct is paired with the claim it is entitled to make, the evidence and statistic that support it, and the observation that would refute it. The last three rows are the constructs that the available evidence does not test.

Construct	Claim under test	Evidence and statistic	Refuted by
Taxonomy and Risk Domains (IV.A)	Every disclosed multi-tenant LLM failure instantiates at least one of the 17 classes	30 coded incidents; proportion with Wilson 95% interval	An incident outside all classes; one occurred (LLMjacking), so the claim holds at 29/30
Comparative coverage (V.C)	Instruments differ in how far they name a disclosed failure and supply a control for it	9×30 ordinal scores; paired bootstrap over incidents, 2×10^4 resamples	A resampled difference interval containing zero; of the eight pairings with this profile, two are not separated
Governance translation (V.C)	The profile assigns owners, board indicators and an amplification construct that the comparators do not	Three document properties per instrument, read once from published text	A comparator text supplying them; determinate, not sampled, so $n = 1$ per cell and no interval exists
Control priority (V.D)	Some control families would have been decisive more often than others	Counterfactual coding of 30 incidents at atomic granularity; Wilson intervals	Overlapping intervals; at $n = 30$ the two leading families are not separated from each other, so only the ordering is reported
Amplification decomposition (IV.C)	Observed reach factors as $A_{\text{obs}} = \varrho \bar{A}$	2 of 30 incidents disclose an affected fraction, both with $\bar{A} = 1$	Nothing available. Falsification needs a case with $\bar{A} < 1$ and a disclosed denominator
Platform loss model (V.F)	No claim; the model propagates declared priors, it does not estimate loss	Monte Carlo, 2×10^5 draws, fixed seed; Spearman rank sensitivity	Not applicable. Priors are declared rather than fitted, so no observation bears on the output
Governance indices (IV.D)	The indices measure governance quality independently of mean maturity	Enumeration of all $5^5 = 3125$ profiles; correlation against mean capability	$r \approx 1$ with mean capability. Observed $r = 0.968$, so this claim is refuted and withdrawn

Two ordinal dimensions vary by incident. *Risk articulation* scores 0 if the instrument names no category the incident instantiates, 1 if the category is so general that the reader must supply the LLM-SaaS mechanism, and 2 if it names the mechanism at a level a practitioner would act on. *Control specificity* scores 0 if no published control applies, 1 if one applies generically, and 2 if a control names the mechanism or artefact to change. To limit per-cell discretion, scores are assigned once per (instrument, mechanism) pair and applied uniformly, an incident taking the maximum over its mechanisms. Three further dimensions are properties of the instrument rather than the incident and are scored once each on the same scale: accountable-owner assignment, board-visible indicator,

and explicit treatment of cross-tenant amplification; their sum is a governance-translation score out of six. Supplementary S3 gives both rule tables with a justification for every cell.

One hazard needs stating in advance rather than in the limitations. We scored our own instrument alongside eight we did not write, and the Risk Domains derive from a taxonomy describing many of these mechanisms, so a high score for the profile is partly definitional. The primary comparison therefore excludes it, comparators are scored generously wherever their text plausibly reaches a mechanism, and Section 5.3 reports both the externally sampled subset and the effect of removing our four self-assigned discounts. None of this substitutes for independent scoring.

3.7 Statistical and Simulation Procedures

Proportions carry Wilson score intervals at 95%, which behave better than the normal approximation at $n = 30$ and near the boundaries.

Comparative coverage needs an inference frame matched to how the data arise. Write $s_{k,i} \in \{0, \dots, 4\}$ for the coverage score of instrument k on incident i , the sum of the two ordinal dimensions above, and $\bar{s}_k = n^{-1} \sum_{i=1}^n s_{k,i}$ with $n = 30$. A repeated-measures test such as Friedman blocks on incidents and assumes the blocks carry independent information; ours do not, because $s_{k,i}$ is a deterministic function of incident i 's mechanism set, so the 30 incidents contain only 11 distinct score vectors and blocking on all 30 inflates significance. The frame is sampling rather than blocking: the unit drawn from the world is the incident, scores are fixed given an incident, so all inferential uncertainty concerns which incidents happen to be in the corpus. A paired bootstrap over incidents estimates exactly that. Each of $B = 2 \times 10^4$ resamples draws indices $i_1, \dots, i_n$ with replacement and recomputes $\bar{s}_k^{(b)} = n^{-1} \sum_j s_{k,i_j}$ for all nine instruments on the *same* resample, which preserves the pairing and reproduces the variation in mechanism mix. Interval estimates are the 2.5th and 97.5th percentiles of $\{\bar{s}_k^{(b)}\}_{b=1}^B$. For a pair we report the observed gap $\Delta_{kl} = \bar{s}_k - \bar{s}_l$, its percentile interval, and $P = B^{-1} |\{b : \Delta_{kl}^{(b)} \leq 0\}|$, the share of resampled corpora in which the ordering does not hold. These describe differences in coded coverage over a corpus with this mechanism mix, not which instrument produces better security outcomes.

The loss model of Equation (1) is evaluated by Monte Carlo with 2×10^5 draws under a fixed seed. Priors are declared rather than estimated: per-tenant loss is lognormal with median US\$50k and $\sigma = 0.8$, a light tail relative to observed cyber losses [56], and the realisation coefficient of Section 4.3 is $\varrho \sim \text{Beta}(1.2, 12)$, whose mean of 0.09 is stated here so that Section 5.5 can put it against the only two observed values. Each draw evaluates Equation (1) at fixed N and fixed $\bar{A}$, so the reported median and 99th percentile are quantiles of the induced distribution of L_{platform} and inherit whatever the priors assert. Global sensitivity uses Spearman rank correlation between each sampled input and the output. Structural sensitivity is handled by running the governance indicators under two competing assumptions about how pillars interact, described in Section 5.7. All code, the coded corpus and the generated tables are released; every number in Sections 5 and 6 is produced by that code.

4. THE ARTEFACT

4.1 Five Risk Domains

A security architect is well served by a decomposition along attack surfaces. A risk committee needs categories where the owner is clear, the impact ties to a business consequence, and the treatment can be mandated by policy. The same vulnerability surface is therefore recut along governance lines into five domains, summarised in TABLE 2.

Table 2: The five governance-grade Risk Domains: accountable owner, source technical vulnerability classes from [12], and the architectural amplification group each source class falls into.

Domain	Definition	Owner archetype	Classes [12]	Groups
RD1 Data Sovereignty & Tenant Isolation	One tenant's data is exposed, processed or referenced for another.	CISO + DPO	B2, B3, B4, C4	G1, G2, G3
RD2 Content Integrity & Adversarial Inputs	Outputs deviate from intent under adversarial prompts or retrieved content.	Chief AI Officer + Trust & Safety	A1–A4, C1–C3	G2, G3
RD3 Operational Agency & Action Risk	An agent takes unauthorized or mis-scoped action through tools.	CTO / Head of Platform	E1, E2	G2, G3
RD4 Supply Chain & Third-Party Trust	Model vendors, MCP servers, datasets or embedding providers introduce risk.	CPO + CISO	B1, D1, D2, D3	G1, G3
RD5 Regulatory & Compliance Exposure	Cross-cutting failure to discharge AI-specific and adjacent obligations on time.	General Counsel / CCO	all (cross-cutting)	—

TABLE 2, gives each domain's definition, owner archetype and source classes; what completes a domain is the lead indicator and the regulatory touchpoint, listed below and in full in Supplementary S4.

RD1, Data Sovereignty and Tenant Isolation: one tenant's data exposed in the service of another, at retrieval, inference, caching or output time. Indicator: detected cross-tenant retrievals per million queries. Touchpoints: GDPR Articles 5, 25 and 32.

RD2, Content Integrity and Adversarial Inputs: outputs deviating from intended behaviour under adversarial content in prompts, retrieved documents or conversation context. Indicator: detected injection attempts, split by direct and indirect delivery. Touchpoints: EU AI Act Articles 13 and 14.

RD3, Operational Agency and Action Risk: action through tools, function calls or downstream integration beyond authorized scope or against the wrong tenant. Indicator: tool invocations executed without tenant scoping. Tallam (2026) [41], shows that authorization propagation across

delegating agents is not reducible to injection control; RD3 covers the invocation boundary and not the transitive delegation chain, which Section 6.4 records as an extension.

RD4, Supply Chain and Third-Party Trust: vulnerabilities, behaviour changes or exfiltration paths introduced by model providers, MCP servers, third-party tools, fine-tuning datasets or vector-database services. Indicator: share of integrated tools holding current attestations. Touchpoints: NIS2 Article 21 and the EU AI Act provider/deployer split.

RD5, Regulatory and Compliance Exposure: obligations under AI-specific and adjacent regulation missed. Every technical class carries a regulatory dimension, and RD5 makes discharging it a separate task under the General Counsel. Indicator: time to readiness per newly applicable regulation.

4.2 Twenty Control Families

FIGURE 1, names all twenty families, so only those needing definition beyond their title are described here. Govern turns on G1, the policy and per-domain risk-appetite statement, without which Assess has no reference point and Mitigate no acceptance criterion; Identify turns on I1, a live inventory of every model, agent, MCP server, vector store and embedding model, without which the rest operates blind; and Assess holds the scoring methodology of Section 4.3 (A1) alongside exercises, pre-deployment sign-off tied to score, and tenant-impact analysis.

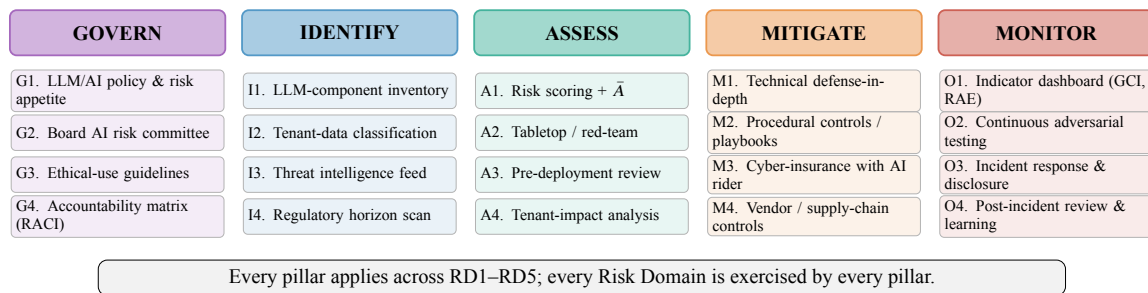


Figure 1: The CSF v2.0 lifecycle operationalized for multi-tenant LLM-SaaS. Twenty control families span five pillars, which are CSF v2.0's six functions regrouped under ERM vocabulary as described in Section 3.2. Section 5.4 reports how often each family was decisive across the 30-incident corpus.

Mitigate holds four families, of which M1 carries the technical work in five layers: input hardening through classifier-based filters such as Llama Guard [57], allow-listed tenant-scoped fields and perplexity scoring of retrieved content; tenant-boundary enforcement through per-tenant vector namespaces, mandatory metadata filtering on every retrieval call, cryptographic tenant tagging and KV-cache isolation; output sanitisation covering PII redaction, code-injection prevention, source attribution and schema validation of tool-call arguments; supply-chain verification through provenance attestations, pinned MCP tool definitions and signed dataset manifests; and continuous monitoring through tenant-scoped logging with cross-tenant anomaly detection. The acceptance criterion is that all five run in production rather than appearing in a plan, and Section 4.4 takes that conjunctive

reading over into the aggregator. The remaining families cover incident playbooks (M2), cyber-insurance with an explicit AI rider (M3), and contractual and technical controls on model providers, MCP publishers and tool authors (M4); Monitor closes the loop through the indicator dashboard (O1), continuous adversarial probing of production endpoints (O2), incident response and disclosure (O3), and post-incident review feeding back into I3, A1 and M1 (O4).

4.3 Amplification: Architectural Bound and Realised Fraction

Standard likelihood-times-impact scoring is extended with a multi-tenancy amplification factor. A single factor read as the expected fraction of tenants affected cannot carry that role, because it merges a property of the architecture, which the provider chooses and an auditor can inspect, with a property of the event, which neither party controls; Section 5.5 shows the two quantities separated by two orders of magnitude on the disclosed cases. The factor is therefore defined in two parts.

Let N be the number of tenants and v a vulnerability class. The *architectural bound* $\bar{A}(v) \in [1/N, 1]$ is the largest tenant fraction the platform's isolation choices permit a single materialization of v to reach: a supremum over conceivable events, fixed once the architecture is fixed. The *realisation coefficient* $\varrho \in [0, 1]$ is the fraction of that bound an actual event attains, absorbing exposure window, traffic overlap and detection latency. The observed amplification is $A_{\text{obs}}(v) = \varrho \bar{A}(v)$, so $A_{\text{obs}} \leq \bar{A}$ by construction and $\mathbb{E}[A_{\text{obs}}(v)] = \mathbb{E}[\varrho] \bar{A}(v)$ when the realisation coefficient is taken independent of the class. The platform-level loss conditional on one materialization is

$$L_{\text{platform}}(v) = L_{\text{per-tenant}}(v) \cdot \varrho \bar{A}(v) \cdot N. \quad (1)$$

The separation matters because $\bar{A}$ is a property of the architecture, which the provider controls and an auditor can inspect, whereas ϱ is a property of the event, which neither party controls. Investment decisions act on $\bar{A}$ alone. The status of the decomposition should be read from TABLE 1, rather than from its algebra: ϱ is nowhere observed apart from A_{obs} , so the two-factor form is an accounting identity adopted for the decisions it separates, and Section 5.5 states the observation that would falsify it.

Rather than assign a point estimate per class, each class falls into one of three architectural groups. Group G1, shared-resource by construction, covers classes living in resources shared by design such as a KV-cache, an MCP tool serving all agents or one model checkpoint serving every tenant, giving $\bar{A} \rightarrow 1$ for B3 and D1–D3. Group G2, tenant-scoped by construction, covers classes confined to one tenant's session, retrieval scope or output path, giving $\bar{A} \rightarrow 1/N$ for A1, A3, A4, B4, C4 and E2. Group G3, configuration-dependent, covers classes whose reach follows isolation choices: shared versus per-tenant vector namespaces (C1–C3), fine-tunes (B1, B2), tool definitions (E1) and sanitised versus raw retrieved content (A2). Here $\bar{A}$ ranges over $[1/N, 1]$ and the assigned value is itself the audit artefact. G3 carries the discriminating power of the instrument: the same per-tenant loss compounds to platform losses one to two orders of magnitude apart depending on one architectural choice. Equation (1) is conditional on occurrence and contains no likelihood term, so L_{platform} is a loss magnitude and not an annualised expectation. Supplying the missing likelihood is open work; AgentRedBench [44], is the most plausible published source for the agentic classes, but we do not use it in any computation here.

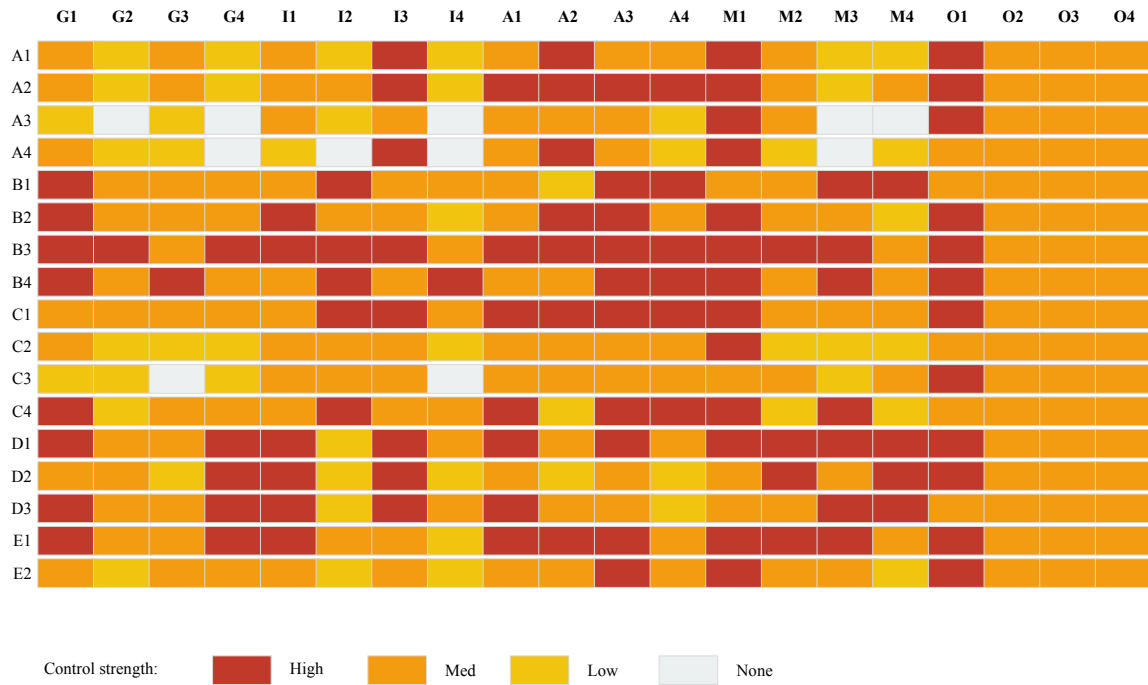


Figure 2: Threat-to-control matrix. Rows are the 17 vulnerability classes of [12], columns are the twenty control families of FIGURE 1. Cell intensity follows the four-level rubric fixed before coding and defined in Section 3.2: *High* the control changes likelihood or impact directly and is auditable, *Medium* it contributes but leaves a residual path, *Low* indirect or partial coverage, *None* no defensible relationship. Coding is author-applied and reviewed internally, so the matrix carries no independent reliability estimate; it is released as machine-readable data for recoding. Section 4.4 computes the Governance Coverage Index and the Residual Amplification Exposure from these weights, and Section 5.7 reports how the resulting priority order compares with the empirical control ranking.

4.4 Measurable Governance Indicators

Three indicators turn the maturity states of Section 4.5 into numbers on a common scale. Let $w_{v,c} \in \{0, 1, 2, 3\}$ be the strength of control family c against class v from FIGURE 2, and $x_c = \ell_c/5 \in [0.2, 1]$ the capability of c at maturity level ℓ_c . The *Control Coverage Score* of a class is the weighted geometric mean of those capabilities,

$$\text{CCS}(v) = \exp\left(\frac{\sum_c w_{v,c} \ln x_c}{\sum_c w_{v,c}}\right) \in [0, 1]. \quad (2)$$

A geometric mean is used because defence in depth is gap-sensitive: a control set with a hole is worth less than its arithmetic average suggests, which is the same assumption the M1 acceptance criterion already makes when it requires all five layers in production. Scaling capability as $\ell_c/5$ rather than $(\ell_c - 1)/4$ keeps $x_c \geq 0.2$, so $\ln x_c$ stays finite and a single immature control depresses the score without collapsing it to zero. The arithmetic alternative $\text{CCS}_+(v) = \sum_c w_{v,c} x_c / \sum_c w_{v,c}$ is retained in the released code, because Section 5.7 shows that one recommendation depends on which of the

two is chosen. The pair is not interchangeable: $CCS \leq CCS_+$ always, with equality only when every weighted capability is equal, and the gap between them is exactly the penalty the gap-sensitive form charges an uneven control set. Averaging over the 17 classes gives the *Governance Coverage Index* $GCI = \frac{1}{|V|} \sum_v CCS(v)$. Weighting the shortfall by architectural amplification gives the *Residual Amplification Exposure*

$$RAE = \frac{\sum_v \bar{A}(v)(1 - CCS(v))}{\sum_v \bar{A}(v)} \in [0, 1], \quad (3)$$

which ranks remediation by how many tenants the uncovered residue can reach rather than by control count. The *Pillar Balance Index* $PBI = 1 - \sigma_{\text{pillar}}/\sigma_{\text{max}}$ records how evenly maturity is distributed, σ_{pillar} being the population standard deviation of the five pillar levels and $\sigma_{\text{max}} = 1.96$ its largest attainable value on a 1–5 scale, reached by a two–three split across the extremes. Multiplying RAE by $N L_{\text{per-tenant}}$ converts it into currency for a risk register, with the caveat on that product carried over from Equation (1): it is a magnitude conditional on occurrence and not an annualised expectation.

One limit of these indices matters more than their definitions. Enumerating all $5^5 = 3125$ attainable pillar profiles, GCI correlates at $r = 0.968$ with plain mean capability, and replacing the 340 coded cells with a uniform matrix moves it by 0.010 on average. The coded weights differentiate *threats*, which is what the residual-priority list uses, but they do not differentiate *organizations* once averaged across threats. An aggregate index built from this matrix is therefore not an independent measurement of governance quality and is not offered as one; what it adds over a bare maturity profile is gap sensitivity, amplification weighting and a route to a currency figure. The choice of aggregator is not cosmetic: the arithmetic form gives a constant marginal gain per level and so recommends the same next step to every organization, while the geometric form points at whichever control is weakest relative to its importance.

4.5 Maturity Model

TABLE 3, gives the per-pillar criteria and Supplementary S4 turns them into a 20-item scored questionnaire. The progression runs from no formal AI risk policy at level 1 to a quantitative register across all five domains at level 3, with level 4 adding measurement: a live dashboard, adversarial testing in CI/CD, risk-differentiated deployment gates and recorded architectural bounds. Level 5 is a research target rather than an attainable state, since formal verification of tenant-isolation invariants and standardised AI-maturity discount schedules have no widely deployed analogue in 2026, so level 4 is the practical assessment ceiling.

5. RESULTS

5.1 Corpus Composition

Across the 30 incidents, indirect prompt injection is the most frequent mechanism (14 of 30), then tool or extension supply-chain compromise (6) and unauthorized agent action (5). Fifteen incidents fall in architectural group G1 and fifteen in G3, and none in G2. That absence is mostly an artefact of the coding rule rather than a finding: an incident takes the maximum group over the classes it

Table 3: Per-pillar capability criteria across the five maturity levels. A completed assessment yields one level per pillar, which the scoring script converts into the three indicators of Section 4.4.

Pillar	L1 Ad-hoc	L2 Defined	L3 Managed	L4 Measured	L5 Optimizing
GOVERN	No AI policy	AI policy documented; risk appetite implicit	Risk-appetite statement per RD; AI risk committee exists	Board reviews indicators quarterly; incidents test appetite	Risk appetite is a public commitment; voluntary disclosure
IDENTIFY	Inventory unknown	LLM stack inventoried; data classification exists	Threat-intel feed in place; horizon scan documented	Inventory live and machine-readable; model SBOM equivalent	Predictive intel feeding pre-emptive controls
ASSESS	Ad-hoc reviews	Feature risk scoring; annual self-assessment	Quantitative risk register; annual red-team	Architectural bound A recorded per class in scoring; continuous tabletops	Formal verification of select invariants; benchmark publication
MITIGATE	No LLM-specific controls	Basic input/output filtering, metadata retrieval filter	Five-layer defence-in-depth deployed; AI insurance rider	Controls measured individually; vendor attestations enforced	Adversarial AI in CI/CD; insurance with maturity discount
MONITOR	Logging only	Manual review of incidents	Quarterly indicator report; documented IR procedure	Real-time dashboard; continuous adversarial testing	Public indicator posture; framework updates published

instantiates, so it lands in G2 only if every class it involves is tenant-scoped, and since direct injection almost always co-occurs with tool execution, the G2 cell is close to unreachable by construction. Eleven incidents do involve at least one G2 class. Disclosure bias toward wide-reaching events plausibly contributes as well, and the corpus cannot separate the two explanations.

TABLE 4, gives the domain distribution. Supply chain and third-party trust dominates at 13 of 30 primary assignments (43%, 95% CI 27–61), ahead of content integrity at 9 (30%) and operational agency at 6 (20%). Tenant isolation is primary in only 2 cases yet implicated in 10. Regulatory exposure is primary in none and implicated in 3, which reflects what security write-ups record rather than what organizations experience, so RD5 cannot be evaluated here.

5.2 Construct Coverage

Twenty-nine of 30 incidents map to at least one of the 17 vulnerability classes (96.7%, 95% CI 83.3–99.4). The exception is LLMjacking [AML.CS0030], where stolen cloud credentials are used to resell access to hosted models. Neither the taxonomy nor the Risk Domains anticipate unmetered consumption of shared model capacity as a category, and the profile reaches it only through third-party trust; a sixth domain covering service abuse is a candidate extension.

Coverage runs the other way as well: eleven of the 17 classes appear and six do not, including embedding inversion, for which laboratory evidence is strong [58], so the taxonomy is broader than the disclosed record can confirm.

Table 4: Risk-Domain distribution across the 30-incident corpus. “Primary” is the mutually exclusive dominant domain and carries a Wilson 95% interval; “Involved” counts every incident touching the domain and therefore sums above 30. RD5 is primary in no incident because vulnerability disclosures do not report regulatory outcomes, not because regulatory exposure is absent.

Domain	Scope	Primary	95% CI	Involved
RD1	Data Sovereignty & Tenant Isolation	2 (7%)	[2, 21]%	10
RD2	Content Integrity & Adversarial Inputs	9 (30%)	[17, 48]%	13
RD3	Operational Agency & Action Risk	6 (20%)	[10, 37]%	12
RD4	Supply Chain & Third-Party Trust	13 (43%)	[27, 61]%	16
RD5	Regulatory & Compliance Exposure	0 (0%)	[0, 11]%	3
Total incidents		30		

Four incidents postdate the taxonomy freeze of 31 August 2025 and all map to pre-existing classes, but that test is weaker than it looks: all four score at the ceiling for this profile, the CSA matrix and the OWASP list alike, so the subset discriminates nothing. The freeze also protects only the 17-class taxonomy, since the Risk Domains, the control families and the comparator rule table were written in 2026 with the whole corpus visible; no part of Section 5.3 is out of sample.

5.3 Comparative Coverage

The headline comparison ranks the eight existing instruments only. Scoring one’s own instrument alongside them cannot settle a ranking, so TABLE 5, places this profile below the rule as a self-scored reference point and the eight-way result stands independently of it.

The CSA AI Controls Matrix leads at 4.00 of 4 (95% bootstrap CI [4.00, 4.00]), then OWASP at 3.87 [3.67, 4.00], NIST AI RMF and MITRE ATLAS at 2.83, ISO/IEC 42001 at 1.90, ISO/IEC 27001 and CSF v2.0 at 1.27, and COBIT 2019 at 0.87 [0.53, 1.20]. The spread is the substantive finding: AI-specific control catalogues cover disclosed multi-tenant failures roughly four times as well as general IT-governance guidance, with the management-system standards between them.

Scored on the same basis, this profile reaches 3.87 [3.70, 4.00]: separated from six comparators by paired bootstrap intervals excluding zero, indistinguishable from the OWASP list ($\Delta = 0.00$, $P = 0.44$), and below the CSA matrix ($\Delta = -0.13$, $[-0.30, 0.00]$). Two sensitivities bound that figure. On the 25 ATLAS cases, the only externally sampled part of the corpus, the profile stands third at 3.84, so parity over all 30 depends on the five cases we added; and setting the four cells where Supplementary S3 records us scoring ourselves below 2,2 back to 2,2 yields 4.00 and a tie at the top. Four self-assigned integers therefore separate third place from first, which is the reason the profile is reported outside the ranking. On these two axes a provider already working from the CSA matrix or the OWASP list gains little here, as expected for a profile built on those instruments rather than against them.

Table 5: Comparative coverage over the 30-incident corpus. RA is risk articulation and CS control specificity, each scored 0–2 per incident by the rule table of Supplementary S3. OW, BI and MT are framework-level scores for accountable-owner assignment, board-visible indicator and explicit cross-tenant amplification; GT is their sum and involves no sampling, being read directly from the published documents. Intervals are 95% percentile bootstrap over 2×10^4 resamples of the incident corpus. **This paper’s profile is separated below the rule and excluded from the ranking**, because we scored all nine instruments ourselves and cannot adjudicate a contest we entered; it is shown for reference only, with the two sensitivities of Section 5.3 attached.

Instrument	Per-incident coverage (n=30)				Framework-level			
	RA	CS	RA+CS	95% CI	OW	BI	MT	GT
CSA AI Controls Matrix v1.0	2.00	2.00	4.00	[4.00, 4.00]	1	1	1	3
OWASP Top 10 for LLM Apps 2025	1.93	1.93	3.87	[3.67, 4.00]	0	0	1	1
NIST AI RMF 1.0 + AI 600-1	1.73	1.10	2.83	[2.50, 3.13]	1	1	0	2
MITRE ATLAS	1.90	0.93	2.83	[2.57, 3.00]	0	0	0	0
ISO/IEC 42001:2023	1.00	0.90	1.90	[1.77, 2.00]	2	1	0	3
ISO/IEC 27001:2022 (Annex A)	0.10	1.17	1.27	[0.93, 1.60]	2	1	1	4
NIST CSF v2.0	0.47	0.80	1.27	[0.97, 1.57]	2	1	0	3
COBIT 2019 (+ ISACA AI guidance)	0.43	0.43	0.87	[0.53, 1.20]	2	2	0	4
<i>Self-scored, excluded from the ranking (Section V.C):</i>								
This paper (deployment profile)	1.97	1.90	3.87	[3.70, 4.00]	2	2	2	6
on the 25 externally sampled cases only			3.84					
without our four self-assigned discounts			4.00					

The separation appears on the other three columns, and it does not depend on a significance test. Ownership, board-visible indicators and cross-tenant amplification together score 6 of 6 here, 4 for COBIT and ISO 27001, 3 for ISO 42001, the CSA matrix and CSF v2.0, 2 for NIST AI RMF, 1 for OWASP and 0 for ATLAS, and no comparator scores above 1 on amplification. These are determinate properties of published documents rather than sampled quantities, so they are read directly. The contribution is therefore the translation into accountable ownership, board reporting and tenant-count-aware exposure, not the enumeration of risks or controls. FIGURE 3, shows both axes together.

5.4 Which Controls Decide Outcomes

Coding each incident for the control families whose operation would most plausibly have prevented or bounded it produces an empirical priority order, reported at atomic granularity. Rolling M1’s four coded layers into one family while leaving the others atomic would put defence-in-depth at 21 of 30, but that is an aggregation artefact, since M1 would fire whenever any sub-control fires while M4 must fire alone. Compared like with like, vendor and supply-chain control leads at 16 of 30 (53%, CI 36–70), then output sanitisation at 15 (50%), input hardening at 7, tenant-boundary enforcement at 6, tenant-impact analysis at 5 and supply-chain verification at 4. One inconsistency is worth recording: Section 4.2 defines M1 as five layers, but only four appear in the coding because continuous monitoring is coded under O2. Most disclosed failures were engineering failures at

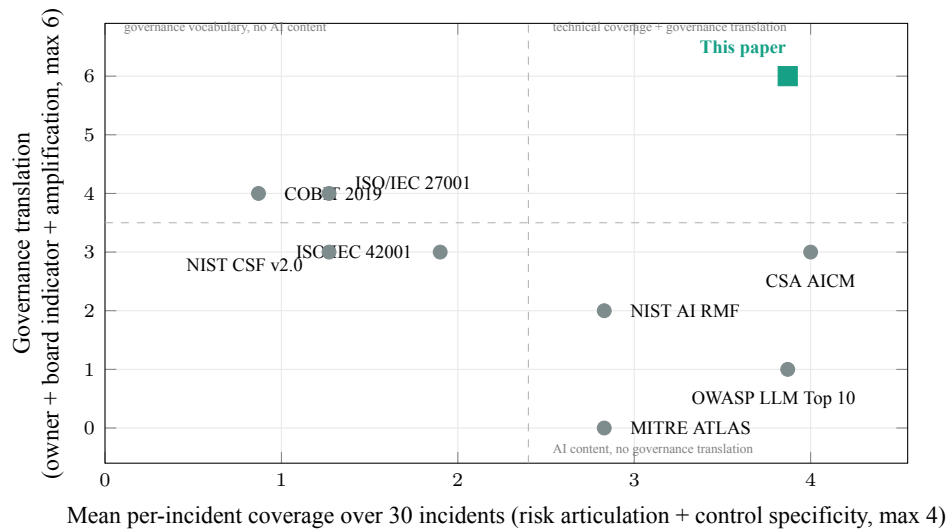


Figure 3: Instrument positioning on the two measured axes. The horizontal axis is mean coverage per incident over the 30-incident corpus, with the caveat that a score is a function of the mechanism set, so horizontal position reflects the mechanism mix of this corpus. The vertical axis is the framework-level governance-translation score, read directly from the published documents. The CSA AI Controls Matrix and the OWASP LLM Top 10 match or exceed this profile on coverage; the separation is vertical, and no comparator scores above 1 on the cross-tenant amplification component. This paper’s point is self-scored and is plotted for orientation, not as a competitive result.

the boundary between the model and the systems around it, which supports sequencing vendor control and output handling first; the low counts for Govern controls do not mean governance is unimportant, since a vulnerability disclosure rarely records whether a policy existed.

5.5 Amplification Against Disclosed Reach

Two incidents in the corpus disclose both a numerator and a denominator. OpenAI reported that the March 2023 Redis defect exposed data belonging to 1.2% of ChatGPT Plus subscribers active during a nine-hour window [51]. Asana reported roughly 1,000 affected customers against a base of over 130,000 paying customers [52], giving about 0.77%, although the correct denominator is the undisclosed opt-in MCP population, so the true conditional rate is higher.

Both are group G1 cases where the architectural bound is $\bar{A} \rightarrow 1$, and neither observed fraction is near 1. That is the finding, and it is what a single-factor amplification term cannot survive: read as an expected affected fraction, such a term over-predicts these two events by two orders of magnitude. Equation (1) accommodates the gap by separating what the architecture permits from what the event realised, and the accommodation must be labelled for what it is. With $\bar{A} = 1$ in both cases ϱ equals the observed fraction, so the decomposition is here a relabelling rather than a tested model: ϱ is free in $[0, 1]$, is never observed apart from A_{obs} , and no available measurement

Table 6: Control families ranked by how often they were coded decisive across the 30 incidents, with Wilson 95% intervals. M1a–M1d are the layers of the defence-in-depth family. Counts are capped at one per incident. Low counts for Govern controls reflect what vulnerability disclosures record rather than what organizations do, as discussed in Section 5.4.

Control	Family	Decisive in	95% CI
M4	Vendor / supply-chain controls	16/30 (53%)	[36, 70]%
M1c	DiD (c) output sanitisation	15/30 (50%)	[33, 67]%
M1a	DiD (a) input hardening	7/30 (23%)	[12, 41]%
M1b	DiD (b) tenant-boundary enforcement	6/30 (20%)	[10, 37]%
A4	Tenant-impact analysis	5/30 (17%)	[7, 34]%
M1d	DiD (d) supply-chain verification	4/30 (13%)	[5, 30]%
I1	LLM-component inventory	3/30 (10%)	[3, 26]%
A3	Pre-deployment review	2/30 (7%)	[2, 21]%
O1	KRI/KPI dashboard	2/30 (7%)	[2, 21]%
O3	Incident response	2/30 (7%)	[2, 21]%
O2	Continuous adversarial testing	1/30 (3%)	[1, 17]%
G4	Accountability matrix (RACI)	1/30 (3%)	[1, 17]%

can contradict it. Falsification requires a case with $\bar{A} < 1$ and a disclosed denominator, which is why TABLE 1, places the decomposition in the untested block. The defensible claim is narrow: reach is not the architectural bound, and separating an inspectable architectural quantity from an uncontrollable event quantity is better bookkeeping than folding both into one number.

Neither data point is even clean: the ChatGPT denominator counts consumer subscribers rather than tenant organizations, and the Asana denominator is the wrong population by the vendor’s own account. Only 2 of 30 disclosures report an affected fraction at all, and that absence, not the choice of model, blocks empirical calibration.

5.6 Uncertainty and Sensitivity

TABLE 7, reports the loss distribution for the worked archetype at $N = 350$ tenants under three architectures. A shared vector store with weak metadata filtering on class C1 ($\bar{A} = 0.5$) gives a median platform loss of \$580k with a 99th percentile of \$7.75M. Verified per-tenant namespaces on the same class ($\bar{A} = 1/N$) give a median of \$3.3k and a 99th percentile of \$44k. A shared inference cache exposing B3 ($\bar{A} = 1$) gives a median of \$1.16M and a 99th percentile of \$15.5M.

These are not estimates but the declared priors pushed through Equation (1), and the prior for ϱ is itself inconsistent with the only two observations, having mean 0.09 against observed 0.012 and 0.008, which inflates every figure in the table.

The shared-to-isolated ratio of 175× might look robust, since $L_{\text{per-tenant}}$ and ϱ cancel between the two rows. It is definitional instead: the ratio is $\bar{A}_{\text{shared}}/\bar{A}_{\text{isolated}} = 0.5/(1/350)$ on every draw, so the simulation recovers a constant we fixed rather than sampled, and $\bar{A}_{\text{shared}} = 0.5$ is a point

taken from the G3 interval without derivation; at 0.1 the same arithmetic gives $35\times$. What survives is the qualitative claim that one isolation decision moves platform exposure by one to two orders of magnitude, which needs no simulation, and the ranking by Spearman correlation in TABLE 8, reproduces the priors' log-dispersions rather than any substantive conclusion, as it must for a product of independent factors.

Table 7: Monte Carlo platform-loss distribution for the worked archetype ($N = 350$), 2×10^5 draws under a fixed seed. Priors are declared, not estimated: $L_{\text{per-tenant}} \sim \ln \mathcal{N}(\ln 50\text{k}, 0.8)$ and $\varrho \sim \text{Beta}(1.2, 12)$. Absolute values inherit those priors and should not be quoted as estimates; the shared-to-isolated ratio of $175\times$ is invariant to both nuisance parameters and is the decision-relevant quantity.

Scenario	Median	Mean	P90	P99
C1 shared vector store, weak filtering	\$578,197	\$1,095,518	\$2,585,450	\$7,653,742
C1 per-tenant namespaces verified	\$3,304	\$6,260	\$14,774	\$43,736
B3 shared inference cache	\$1,156,395	\$2,191,036	\$5,170,899	\$15,307,484

Table 8: Global sensitivity of platform-level expected loss to its three inputs, measured as Spearman rank correlation over the Monte Carlo sample with all inputs varying. The realisation coefficient dominates but is outside a provider's control; the architectural bound ranks second and is the input the instrument targets.

Input	Distribution	Spearman ρ_s with L_{platform}
$A(v)$ architectural	$\mathcal{U}(1/N, 1)$	+0.507
$L_{\text{per-tenant}}$	$\ln \mathcal{N}(\ln 50\text{k}, 0.8)$	+0.493
ϱ realisation	$\text{Beta}(1.2, 12)$	+0.629

5.7 Indicator Behaviour and the Status of the Balance Heuristic

Applying the indicators to the worked archetype at its starting profile (Govern 1, Identify 2, Assess 1, Mitigate 3, Monitor 2) gives GCI 0.330, RAE 0.668 and PBI 0.618; bringing all five pillars to level 3 gives 0.600 and 0.400. Ranking by residual exposure puts cross-tenant leakage first, then compromised model supply chain and tool poisoning, an ordering dominated by the amplification weights, since the coverage term separates classes within a group rather than across them.

Maturity guidance routinely holds that a balanced profile beats an unbalanced one at the same average maturity, and the indices of Section 4.4 make that advice checkable rather than assertable. Enumeration shows the claim is neither true nor false as stated: it is entailed by one composition assumption and empty under the other, and no data here adjudicates between them. Over all 121 profiles with mean maturity exactly 2.0 (TABLE 9), gap-sensitive composition ranks the balanced profile first with a correlation between balance and coverage of +0.90, while additive composition gives -0.000 and ranks it 63rd. The dependence belongs to the aggregator and not to our coded weights: replacing the matrix with uniform weights still leaves the balanced profile first, because at fixed arithmetic mean the weighted geometric mean is maximised exactly when all capabilities are

equal. That is the inequality of arithmetic and geometric means restated in maturity vocabulary, so it holds for every non-negative weight matrix and reports nothing about ours. The balance advice is therefore not wrong so much as silent about its assumption. It holds if a gap costs more than its share of the average, which is what the M1 acceptance criterion already assumes in requiring all five layers in production. Whether real control sets behave that way is empirical, and Section 6.4 states the study that would settle it.

Table 9: The pillar-balance claim is decided by the composition model, not by data. All 121 maturity profiles with mean exactly 2.0 were enumerated and ranked by Governance Coverage under each model. Gap-sensitive composition ranks the balanced profile first and correlates strongly with balance; additive composition is exactly indifferent to it. Replacing the 340 coded matrix cells with uniform weights leaves the balanced profile first under gap-sensitive composition, so the result follows from the aggregator rather than from our coding.

Composition model	rank of (2,2,2,2,2)	corr(PBI, GCI)	best profile
Gap-sensitive (geometric)	1 / 121	+0.904	(2,2,2,2,2)
Additive (arithmetic)	63 / 121	-0.000	(1,1,5,1,2)

5.8 Worked Archetype

A constructed 200-employee B2B vendor with 350 enterprise customers illustrates the instrument end to end in Supplementary S6, carrying no evidential weight and included only to show the assessment, the indicators and a twelve-month roadmap in one place. The one point worth making here is that the indicators reorder the roadmap: at the starting profile the four highest residual exposures are all group G1 classes, so tenant-boundary enforcement and supply-chain verification precede input filtering even though input filtering is the more visible control, and the empirical priority of Section 5.4 points the same way.

6. DISCUSSION

6.1 What the Comparison Establishes

The comparison bounds what this work claims. On naming a disclosed failure and supplying a control for it, the CSA AI Controls Matrix and the OWASP LLM Top 10 are at least as strong as this profile; that axis is a translation of existing catalogues and the measurement says so. What the profile adds sits on the governance axis, and it is not decoration: neither leading catalogue assigns an accountable owner, defines an indicator a board would read, or carries a construct for how many tenants one failure reaches, which are the three things a risk committee acts on.

Two consequences follow for a practitioner. An organization already running the CSA matrix should adopt the ownership mapping, the indicator set and the amplification bound, and should not rebuild its control catalogue; an organization starting from COBIT or ISO 27001 has the governance

machinery and needs the AI content, the opposite deficit. The profile bridges those two populations, and its value scales with how far apart they sit in a given firm.

6.2 Insurance Underwriting

A maturity model of this kind is often said to align with cyber-insurance practice, and the industry evidence will not carry that claim: Munich Re records that AI-specific exposures are often not explicitly covered in standard cyber wordings [16], and Lloyd's expects generative AI to raise the frequency and severity of cyber losses [49], together they show an unsettled market, and neither shows that a maturity score moves a premium.

Recent formal work permits a more precise statement. Zhu models an agentic deployment as a risk state including governance maturity, permission exposure and dependency concentration, and derives an insurability region with governance certification thresholds [46]. In that framework a maturity instrument is useful to the extent it produces discrete auditable evidence tied to those state variables; our levels are built that way, and *A* maps onto dependency concentration directly, since a G1 architecture concentrates the whole tenant base on one resource. The claim is structural compatibility with a published underwriting formalism, not an effect on pricing; whether AI-maturity discount schedules reach standard wordings stays open, as does the wider question of how insurers turn security practice into cover conditions [15, 59].

6.3 Threats to Validity

Circularity. We scored all nine instruments, one of which is ours, and the Risk Domains derive from the taxonomy describing many of the coded mechanisms, so this profile's score is partly definitional. The remedy is structural: the profile is reported outside the ranking rather than defended within it. A weaker remedy would be to observe that the profile does not win, but the whole gap to the leading comparator is four cells we assigned ourselves below the maximum, so a low score is as much ours to choose as a high one. The eight-way comparison retains a milder version of the same problem, because we scored those eight too, and blind scoring by coders with no stake in any instrument remains the correct next step; the out-of-sample subset of Section 5.2 offers no protection.

Coding. One author coded and the second reviewed, both being authors of the source taxonomy, so agreement measures internal consistency and no kappa is reported. Scoring per (instrument, mechanism) pair suppresses ad-hoc discretion but flattens within-mechanism variation, leaving eleven profiles of information rather than thirty. Judgements about instruments we have read but not operated at scale, COBIT in particular, may be conservative, and naming the control that would have prevented an incident is hindsight about systems we did not run.

Corpus. ATLAS applies its own inclusion policy and disclosure favours severe, externally discovered events, so prevalence describes the disclosed record and not the population of platforms. Our own application of the shared-service criterion is inconsistent: consumer assistants are included while single-user desktop agents are excluded, yet developer tools with a similar usage pattern are included on supply-chain grounds, and no written adjudication of all 57 cases exists. ATLAS is

also both sampling frame and scored comparator, which secures its articulation score on 83% of the corpus, and RD5 is unevaluable here.

The quantitative apparatus estimates nothing yet. These constructs organise reasoning about exposure and none of them measures it, which is what the lower block of TABLE 1, declares in advance rather than concedes at the end. Section 5.7 shows the sharper form of the point: for one recommendation the model choice, not the evidence, determines the answer.

The governance dimension is the least-evidenced part of the contribution. It rests on 27 integers scored once each by us, and its amplification component places this profile at the top of a scale built around a construct we introduced. Supplementary S3 publishes the rubric and a per-cell justification so the assignment can be contested, but with $n = 1$ per cell no test is possible. Nothing here shows that adopting the profile reduces incidents or losses.

6.4 Limitations and Future Work

The profile addresses text-only deployments, and multimodal systems add surfaces the source taxonomy does not cover.

Four studies follow from the results above, one per untested construct in TABLE 1. Section 5.3 needs blind scoring: the same rule table applied by coders with no stake in any of the nine instruments, with agreement reported, which would settle whether the ordering is a property of the instruments or of their authors. The composition question of Section 5.7 is testable on a panel of vendors with known control profiles and incident histories, by checking whether an uneven programme suffers more incidents than an even one at matched average maturity, which sub-additive composition predicts and additive composition does not. The realisation coefficient needs a denominator, and the practical route is a disclosure convention reporting affected tenants against exposed tenants. Expert validation remains outstanding: the questionnaire in Supplementary S4 is built for a practitioner panel, and running it is the next study rather than a completed one.

7. CONCLUSION

Multi-tenant SaaS vendors do not need another AI risk framework. They need the existing ones expressed in a vocabulary their risk committees already use, at a level of detail their engineers can act on, with a way of saying how far one failure travels. This paper builds that on top of CSF v2.0 and ISO/IEC 42001: five Risk Domains with named owners and indicators, twenty control families, and an amplification factor split into an architectural bound and a realised fraction.

Testing against 30 disclosed incidents produced results in three directions. Construct coverage held: 29 of 30 incidents instantiate a class of the underlying taxonomy, and the exception, abuse of shared model capacity, marks a gap the decomposition should close. The comparison of eight existing instruments stands independently of our artefact and is the more portable result: AI-specific control catalogues cover disclosed multi-tenant failures roughly four times as well as general IT-governance guidance, and none of the eight carries a construct for how far one failure reaches. On coverage the profile sits behind the CSA matrix and third on the externally sampled subset; what separates

it is ownership, board reporting and tenant-count-aware exposure, where it leads every instrument tested. Two constructs are reported as untested: whether an even maturity profile helps depends on whether controls compose sub-additively, which no available data settles, and an aggregate index from the threat-to-control matrix does not distinguish organizations beyond their mean maturity.

Two findings are worth carrying into practice. Vendor and supply-chain control was decisive in 53% of disclosed cases and output sanitisation in 50%, which argues for sequencing those two first; and only 2 of 30 disclosures reported how many customers were reached, neither against the exposed population. Changing that disclosure practice would do more for empirical AI risk governance than another framework would.

DATA AND CODE AVAILABILITY

The coded incident corpus, the machine-readable threat-to-control matrix, the comparative coverage matrix, the analysis and simulation code, the assessment questionnaire and the scoring script accompany this paper as Supplementary Material. Every figure and table in Sections 5 and 6 is regenerated by `analysis/analysis.py` under a fixed random seed. The MITRE ATLAS data underlying Corpus A was retrieved on 3 August 2026 and the retrieved copy is included so that results remain reproducible as the upstream register changes.

CONFLICT OF INTEREST

The authors declare no financial conflict of interest. We disclose the following methodological conflicts, each also described where it bears on a result. One of the nine instruments compared in Section 5 is our own, and we scored all nine ourselves; no independent coder was involved. The Risk Domains derive from our own prior taxonomy [12]. We coded the threat-to-control matrix from which the indicators of Section 4.4 are computed. We selected the five non-ATLAS incidents in the corpus, and Section 5.3 reports how the comparison changes without them.

References

- [1] <https://www.coso.org/enterprise-risk-management>
- [2] <https://www.iso.org/standard/65694.html>
- [3] Thompson C, Hopkin P. *Fundamentals of Risk Management: Understanding, Evaluating and Implementing Effective Enterprise Risk Management*. 6th ed. London (UK): Kogan Page; 2022.
- [4] <https://csrc.nist.gov/pubs/sp/800/37/r2/final>
- [5] Erl T, Mahmood Z, Puttini R. *Cloud Computing: Concepts, Technology & Architecture*. Upper Saddle River (NJ): Prentice Hall. 2013.

- [6] Kabbedijk J, Bezemer CP, Jansen S, Zaidman A. Defining Multi-Tenancy: A Systematic Mapping Study on the Academic and the Industrial Perspective. *J Syst Softw.* 2015;100:139-148.
- [7] Greshake K, Abdelnabi S, Mishra S, Endres C, Holz T, et al. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications With Indirect Prompt Injection. In: *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*; 2023:79-90.
- [8] Zou W, Geng R, Wang B, Jia J. Poisoned RAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. In: *Proceedings of the 34th USENIX Security Symposium.* 2025:3827-3844.
- [9] Romankiv N. Model Context Protocol (MCP): Key Aspects and Impact on the AI Ecosystem. *Grail Sci.* 2025:605-607.
- [10] Wu G, Zhang Z, Zhang Y, Wang W, Niu J, et al. I Know What You Asked: Prompt Leakage via KV-Cache Sharing in Multi-Tenant LLM Serving. In: *Proceedings of the Network and Distributed System Security Symposium (NDSS).* 2025.
- [11] Zhang Q, Wang EE, Li J, Wang X. Burn-After-Use for Preventing Data Leakage Through a Secure Multi-Tenant Architecture in Enterprise LLM. 2026. Arxiv Preprint: <https://arxiv.org/pdf/2601.06627>.
- [12] Romankiv N, Sytnikov D. Security Challenges of LLM Integration in Multi-Tenant SaaS: Threats, Vulnerabilities, and Mitigations. *Adv Knowl Based Syst Data Sci Cybersec.* 2026;3:443-465.
- [13] Cherednichenko O, Sytnikov D, Romankiv N, Sharonova N, Sytnikova P. AI Agent for Conversational Q&A Over SaaS Codebase Using Large Language Models. In: *Proceedings of the Intelligent Systems Workshop at the 9th International Conference on Computational Linguistics and Intelligent Systems (ISW-CoLInS 2025).* CEUR Workshop Proc. 2025;3983:80-92.
- [14] Romankiv N. Methodology of Context Engineering for Increasing the Relevance of AI Agents in SaaS Systems. *Grail Sci.* 2025:818-820.
- [15] Woods DW, Böhme R. How Cyber Insurance Shapes Incident Response: A Mixed Methods Study. In: *Proceedings of the 20th Workshop on the Economics of Information Security (WEIS).* 2021.
- [16] <https://www.munichre.com/en/insights/cyber/cyber-insurance-risks-and-trends-2025.html>
- [17] <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>
- [18] <https://www.iso.org/standard/42001>
- [19] <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
- [20] <https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf>
- [21] Aven T. Risk Analysis. 2nd ed. Hoboken (NJ): Wiley. 2015.

- [22] Renn O. Risk Governance: Coping With Uncertainty in a Complex World. London Routledge. 2008.
- [23] <https://csrc.nist.gov/pubs/sp/800/30/r1/final>
- [24] <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>
- [25] Falco G, Shneiderman B, Badger J, Carrier R, Dahbura A, et al. Governing AI Safety Through Independent Audits. *Nat Mach Intell.* 2021;3:566-571.
- [26] Mökander J, Schuett J, Kirk HR, Floridi L. Auditing Large Language Models: A Three-Layered Approach. *AI Ethics.* 2024;4:1085-1115.
- [27] <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>
- [28] <https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-guiding-principles-advanced-ai-system>
- [29] Schneeberger D, Stöger K, Holzinger A. The European Legal Framework for Medical AI. In: *Proceedings of the IFIP Cross-Domain Conference on Machine Learning and Knowledge Extraction (CD-MAKE). Lecture Notes in Computer Science.* Springer. 2020:209-226.
- [30] <https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/>
- [31] <https://cloudsecurityalliance.org/artifacts/ai-controls-matrix>
- [32] <https://www.isaca.org/resources/white-papers/2025/leveraging-cobit-for-effective-ai-system-governance>
- [33] Gupta P. AI TIPS 2.0: A Comprehensive Framework for Operationalizing AI Governance. Arxiv preprint: <https://arxiv.org/abs/2512.09114>
- [34] Koch C. From Governance Norms to Enforceable Controls: A Layered Translation Method for Runtime Guardrails in Agentic AI. 2026. Arxiv preprint: <https://arxiv.org/pdf/2604.05229>
- [35] Morgan H. AI Governance Control Stack for Operational Stability: Achieving Hardened Governance in AI Systems. 2026. Arxiv preprint: <https://arxiv.org/pdf/2604.03262>
- [36] Acharya V. Governing the Agentic Enterprise: A Governance Maturity Model for Managing AI Agent Sprawl in Business Operations. 2026. Arxiv preprint: <https://arxiv.org/pdf/2604.16338>
- [37] Surve PP, Shabtai A, Elovici Y. CEDAR-42001: From ISO/IEC 42001 Conformity to Architecture-Aware, Audit-Visible Assurance Posture for AI Cyber-Physical Systems. 2026. Arxiv preprint: <https://arxiv.org/pdf/2606.21276>
- [38] Mumtaz U, Noor R, Mumtaz S. Post-Deployment Accountability in AI Governance: A Cross-Regulatory Empirical Analysis of AI Incidents. 2026. Arxiv preprint: <https://arxiv.org/pdf/2605.16281v3>
- [39] Huwyler H. Standardized Threat Taxonomy for AI Security, Governance, and Regulatory Compliance. 2025. Arxiv preprint: <https://arxiv.org/pdf/2511.21901>
- [40] Errico H, Ngiam J, Sojan S. Securing the Model Context Protocol (MCP): Risks, Controls, and Governance. 2025. Arxiv preprint: <https://arxiv.org/pdf/2511.20920>

- [41] Tallam K. Authorization Propagation in Multi-Agent AI Systems: Identity Governance as Infrastructure. 2026. Arxiv preprint :<https://arxiv.org/pdf/2605.05440>
- [42] Sun H, Liu S, Ma S, Li J, Xiao M, et al. Agent-Assisted Side-Channel Attacks on Non-Prefix KV Cache in RAG. 2026. Arxiv preprint :<https://arxiv.org/pdf/2606.21842>
- [43] Pennas PG, Papaioannou K, Guarnieri M, Doudali TD. PrefixWall: Mitigating Prefix Caching Side Channels in Shared LLM Systems. 2026. Arxiv preprint: <https://arxiv.org/pdf/2603.10726>
- [44] Dingeto H, Leeney W. AgentRedBench: Dynamic Redteaming and Integration-Aware Defense for LLM Agents Over SaaS Integrations. 2026. Arxiv preprint: <https://arxiv.org/pdf/2606.02240>
- [45] Zhu Q. Insurance of Agentic AI. 2026. Arxiv preprint: <https://arxiv.org/pdf/2606.05449>
- [46] Zhu Q. AI-Native Insurance for Agentic AI: Pricing, Underwriting, and End-to-End Automation. 2026. Arxiv preprint: <https://arxiv.org/pdf/2607.13230>
- [47] Hevner AR, March ST, Park J, Ram S. Design Science in Information Systems Research. MIS Q. 2004;28(1):75-105.
- [48] Chrissis MB, Konrad M, Shrum S. CMMI for Development: Guidelines for Process Integration and Product Improvement. 3rd ed. Upper Saddle River (NJ). Addison-Wesley. 2011.
- [49] https://assets.lloyds.com/media/439566f8-e042-4f98-83e5-b430d358f297/Lloyds_/Futureset_/GenAI_/Transforming_/the_/cyber_/landscape.pdf
- [50] <https://atlas.mitre.org/>
- [51] <https://openai.com/index/march-20-chatgpt-outage/>
- [52] <https://www.bleepingcomputer.com/news/security/asana-warns-mcp-ai-feature-exposed-customer-data-to-other-orgs/>
- [53] <https://nvd.nist.gov/vuln/detail/CVE-2025-32711>
- [54] <https://noma.security/blog/forcedleak-agent-risks-exposed-in-salesforce-agentforce/>
- [55] <https://appomni.com/ao-labs/ai-agent-to-agent-discovery-prompt-injection/>
- [56] Maillart T, Sornette D. Heavy-Tailed Distribution of Cyber-Risks. Eur Phys J B. 2010;75:357-364.
- [57] Inan H, Upasani K, Chi J, Rungta R, Iyer K, et al. Llama Guard: LLM-Based Input-Output Safeguard for Human-AI Conversations. 2023. Arxiv preprint. <https://arxiv.org/pdf/2312.06674>
- [58] Morris JX, Kuleshov V, Shmatikov V, Rush AM. Text Embeddings Reveal (Almost) As Much As Text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP); 2023:12448-12460.
- [59] Eling M, Schnell W. What Do We Know About Cyber Risk and Cyber Risk Insurance? J Risk Finance. 2016;17:474-491.