FSB FINANCIAL STABILITY BOARD

Sound Practices for Responsible Adoption of Artificial Intelligence (AI): Consultation report

Response to Consultation

XSOC CORP

1. Do you agree with the benefits and risks of AI adoption by financial institutions described in this report? Are there any substantive benefits or risks not covered?

XSOC Corp is a United States post-quantum cryptographic security company (CAGE 8ZXJ8) and Non-Traditional Defense Contractor serving government and financial services. We agree with the benefits and with the risk taxonomy as far as each goes. We identify one substantive benefit and six substantive risks that are not covered.

THE UNCOVERED BENEFIT: GOVERNABLE AI CONSUMPTION

Section 1.2 catalogues the benefits of AI adoption as efficiency, improved risk monitoring, faster decisioning, and reduced operational cost. It omits a benefit that has become board-level in 2026: the ability to govern, attribute, and control AI consumption itself.

The evidence that this is now material is public. Uber's Chief Technology Officer has stated the company exhausted its 2026 AI budget within four months, with per-engineer monthly spend on coding assistants ranging from 500 to 2,000 US dollars. Microsoft's Experiences and Devices division is cancelling internal coding-assistant licences for approximately 12,000 engineers, with cost cited as a driver. Nvidia's vice president of applied deep learning has stated that compute cost on his team now exceeds employee cost.

The driver in each case is not vendor pricing. It is unmanaged waste in the consumption stream: tokens spent on outputs that were wrong and had to be re-prompted, inferences rejected downstream that consumed budget but produced nothing usable, and staff time spent correcting low-quality output. Current AI tooling logs consumption in aggregate and has no mechanism to detect, reject, or attribute any of it. The waste is invisible to the finance function.

This is a governance benefit, not merely a cost benefit, and it belongs in the report. An architecture that verifies output at inference time and refuses what fails verification prevents the re-prompt cycle that constitutes the largest single component of token waste. An architecture that bounds the trust of derived outputs prevents low-quality output from propagating into downstream inferences that compound it. A tamper-evident audit chain attributes every consumed token to a specific authorization scope. The result is AI infrastructure whose spend a chief financial officer can attribute, control, and explain to a

board, an audit committee, or a supervisor. Institutions that cannot do this are, in a real sense, running an unmanaged expense line against an unbounded liability, and boards are beginning to notice. We recommend the final report name governable, attributable AI consumption as a benefit of responsible adoption, since it is one of the few benefits that follows directly from the guardrails rather than in tension with them.

2. Are the sound practices sufficiently comprehensive and clear to enable financial institutions' responsible AI adoption?

The sound practices are clear and well structured. They are not yet sufficiently comprehensive, in that for several risks the report correctly identifies, the prescribed response is governance, documentation, and monitoring where architectural and cryptographic controls exist, are proven in production, and are materially more effective.

A scope observation first. Footnote 48 directs institutions to interagency model risk management guidance including FRB et al (2026). The revised United States interagency guidance, SR 26-2 (Federal Reserve, OCC, FDIC, 17 April 2026), supersedes SR 11-7 and, in footnote 3, explicitly excludes generative AI and agentic AI from the formal scope of model risk management guidance while affirming that institutions must still establish appropriate governance for these systems. The FSB's sound practices therefore point, for the AI categories of greatest current supervisory concern, to a framework that has formally declined to cover them. This raises the importance of the FSB toolkit being concrete about control classes for GenAI and agentic AI, since for many institutions it will be the primary reference filling that gap. The principles SR 26-2 retains, in particular effective challenge, ongoing monitoring, and outcomes analysis, remain the right principles. What is missing is a mechanism by which they can be discharged for systems whose outputs are non-deterministic and produced at machine speed. We describe such a mechanism below.

Sound Practice 2 (governance and accountability). The three-lines-of-defence discussion would benefit from a concrete allocation for AI-specific controls, which cut across functions in ways existing models do not anticipate. In deployments we operate, the allocation runs as follows. The first line owns the canonical reference for its domain, that is, the definitions and invariants that AI output must remain aligned with, since only the business owns what its terms mean. Risk and compliance set the tenant policy: the applicable regulatory profile, the classification ceiling, and the trust threshold below which output is not admitted. Security owns the cryptographic key material and the identity roles, including their time-to-live. The third line, and external auditors and supervisors, consume the audit chain and verify its signatures cryptographically, which requires trusting neither the vendor nor the institution's own operators. That last property is the one we would encourage the report to draw out, because it changes what independent assurance means: assurance ceases to rest on the assurer's trust in the record keeper and rests instead on verification.

Sound Practice 3 (risk management framework and documentation). The report asks for tracking and documentation across the lifecycle, including enhanced logging for GenAI and agentic use cases. Documentation that can be silently altered after an incident protects no one. Lifecycle documentation and decision logs for high-materiality use cases should be tamper-evident: hash-chained, append-only, cryptographically signed, and periodically anchored to an external timestamping authority, so that alteration of any record is detectable by verification rather than by trust in the record keeper. This aligns the toolkit with EU AI Act

Article 12, which requires high-risk AI system logs to be tamper-evident, timestamped, and verifiable by national competent authorities, and it converts the report's documentation expectations into evidence that survives incident response and examiner review. On cost, which is the usual objection: signature schemes with information-theoretic security bounds run in production at roughly 72,000 signatures per second per processor core, meaning a gateway handling 1,000 inferences per second consumes approximately 1.4 percent of one core for its entire cryptographic audit workload. Tamper-evident logging is no longer a performance trade-off.

Sound Practice 6 (selection). Selection considerations should include a durability assessment of any safety or policy behavior relied upon: whether the property is implemented in model weights, and therefore removable at bounded cost by any party with weight access including via fine-tuning, or enforced by controls external to the model. For open-weight models and fine-tunable third-party deployments, in-model safety behavior should be classified as non-durable, and compensating boundary controls should be a selection prerequisite for material use cases.

Sound Practice 7 (data governance). The formulation "accurate, complete, consistent, reliable, and secure" treats security generically. The data-security bullet in Section 3.4 should be expanded to reference: record-level encryption of data used for AI training, fine-tuning, and retrieval, with per-record key derivation so that compromise of one key exposes one record rather than the corpus; irreversible cryptographic de-identification of PII in preference to reversible substitution; and cryptographically verifiable data lineage. The report already asks institutions to track data lineage for auditability; lineage records that are hash-chained and tamper-evident convert that documentation exercise into an integrity control that directly counters the training-data poisoning risk described in Section 1.3. We would add a category the report does not contemplate: the institution's own governed reference data, meaning the canonical definitions against which AI output is checked, is itself a poisoning target. It should be version-chained, individually signed at creation, and monitored for adversarial modification. An institution that verifies AI output against a corrupted reference has automated its own error, and will do so with high confidence and a clean audit trail.

Sound Practices 8, 9, and 10 (explainability, performance management, human oversight). Output-level evaluation and human review of individual outputs cannot detect cumulative semantic corruption relative to an external source of truth. Four additions.

To Sound Practice 8's compensating-controls list: inference-time verification of model output against an institution-controlled canonical reference (governed definitions, dates, citations, named entities, and other invariants the institution requires to be preserved), with configurable disposition of drifting output ranging from annotation, to rendering alongside the canonical definition, to override, to fail-closed refusal. Such checks run at microsecond scale against reference stores of practical size and add no material latency relative to model inference.

To Sound Practice 9: for delegated editing and multi-step workflows, performance evaluation should measure cumulative drift across the interaction sequence, not only per-output quality. We further recommend the report contemplate quantitative trust grading of individual AI outputs, computed at inference time from the output's drift against the canonical

reference, the regulatory profile applicable to the institution and use case, the tenant policy version in force, and the grades of any upstream sources the output depends on, with an institution-set admission threshold below which output is not released. Grades should carry a lifecycle (issued, revoked under an administrative scope, expired on a mandatory time-to-live, invalidated on detected poisoning of the reference, invalidated on failed authorship attestation) with every transition signed and chained. This is, in substance, the documentary instantiation of effective challenge for non-deterministic systems: every output is challenged, the challenge is scored, the scoring is recorded immutably, and the chain of grades is reviewable in retrospect by an examiner. It also answers the question examiners actually ask, which is not only what the model decided but which policy was in force when it decided, since the policy version applied to each grade is recorded on the grade itself.

To Sound Practice 10: human oversight of high-materiality outputs should be equipped with the canonical reference and the edit history, since a reviewer shown only the final output is subject to the same formal detection limit as an automated classifier.

Across all three: adopt fail-closed as a named principle. Any verification or authorisation path that errors should deny rather than default to allow.

Sound Practice 11 (cyber and ICT risk management). Add three elements. First, quantum-resistant cryptography and crypto-agility assessments for AI systems and the data they touch. Second, phishing-resistant, hardware-bound authentication for personnel and systems with access to AI models, training data, and vector stores. The AI-enabled attack taxonomy in Section 1.3 (poisoning, prompt injection, extraction, deepfake phishing) shares a common enabling stage: credential theft. Architectures in which credentials are not stored, transmitted, or reusable remove that stage rather than detecting it. Third, and specific to the report's concern about novel attack classes emerging faster than defenders can characterise them, we recommend the toolkit acknowledge path-level detection as a control class distinct from output inspection. Detection that operates on the geometry of the inference path (measuring, for example, whether information is collapsing onto a single degenerate direction, whether the trajectory conserves energy between input and conclusion, or whether the output is over-committed relative to the model's own evidence budget) identifies adversarial steering, indirect prompt injection, and looping multi-agent pipelines by their structural fingerprint rather than by matching a known-bad signature. The distinction matters precisely for the novel-attack case: signature matching detects what has already been seen, whereas path-level anomaly detection can flag classes that have not yet been characterised. It is not a complete defence, and should not be presented as one, but it is a materially different detector with a materially different failure mode, which is what defence in depth requires.

Sound Practice 12 (third-party risk management). Three additions. First, the LiteLLM supply-chain compromise of March 2026, which affected thousands of organisations including financial-sector counterparties, was productive because poisoned dependencies found harvestable secrets on developer machines and in environment files. Institutions should assess whether their own and their providers' architectures minimise standing secrets (long-lived API keys, stored tokens, exportable private keys), since the blast radius of a dependency compromise is proportional to the credential material available to harvest.

Second, the report identifies service provider concentration and the resulting homogeneous behaviors and correlated outcomes across institutions as a financial stability risk. A governance layer that sits above the model providers, holds the institution's canonical reference and policy locally, and enforces the same verification, grading, and audit regime across multiple providers is a direct mitigant. It decouples the institution's control environment from any single provider, it makes provider substitution a configuration change rather than a re-architecture, and it means the institution's governance data does not reside with the provider. We recommend Sound Practice 12 name provider-independent governance as a concentration-risk control, distinct from the exit-planning and business-continuity measures currently described.

Third, on third-party data and definitions specifically: where an institution consumes reference data, taxonomies, or model outputs from an external source, the trust it assigns to anything derived from that source should be capped by how strongly the source itself has been verified. A tiered ceiling (for example, an unverified source capping derived trust at a low level, an authority-verified source at a higher level, and a cryptographically attested source uncapped) makes third-party risk a runtime enforcement property rather than a procurement questionnaire. This is straightforward to implement and directly addresses the report's insufficient-transparency concern.

3. **Do the sound practices strike an appropriate balance between managing risks relating to all forms of AI, and addressing some of the risks relating to emerging and new complex forms of AI, such as GenAI and agentic AI?**

The treatment of agentic AI risk (unauthorized actions, disruption of connected systems, agent-to-agent collusion, impracticality of real-time human monitoring) is candid and accurate. The balance is not yet right, because the report's stated conclusion, that effective monitoring may require augmentation with another AI agent, concedes that governance and human oversight alone cannot operate at agent speed, but the sound practices do not then name the control classes that can.

We submit that the durable controls for GenAI and agentic AI organize around two boundaries the final report could name explicitly.

THE OUTPUT BOUNDARY. Verify what a model produces against the institution's governed source of truth before it reaches a system of record or leaves through a governed channel; hold or block what cannot be verified; extend verification across multi-agent chains so that the provenance of a result produced by several cooperating agents is recorded and checkable end to end; and record every decision in a tamper-evident log. The production incidents of 2026 all terminate at this boundary. Zero-click prompt injection against enterprise assistants exfiltrated data through trusted output channels, documented against more than one major vendor's agent platform, and the first large-scale indirect prompt injection campaigns were observed in the wild in March 2026. In almost none of these cases is the harm the model merely saying something wrong. The harm is realized when tainted output crosses into a record or out through a channel.

Within that boundary, the operative mechanism is quantitative trust grading at inference time. Each output receives a score computed from its drift against the institution's canonical reference, the regulatory profile applicable to that institution and use case, the policy version

in force, and the trust of any upstream sources it depends on. Output below the institution's admission threshold is held rather than released. We would stress that the regulatory profile is implementable as a concrete, enumerated, versioned setting rather than an abstraction. In our own deployment the profiles are basel_iii, dodd_frank, itar, ear, cmmc, fisma, and a default, resolved per tenant at inference time from the active policy version and bound immutably to both the inference response and the resulting grade. Changing the institution's policy changes the profile applied to subsequent inferences, and the version that governed any past decision remains recorded on that decision. This is what allows a supervisor to ask not only what the system decided, but under which regime and under which policy version it decided, and to receive a cryptographically verifiable answer.

Grades should carry a lifecycle: revocable under an administrative scope, expiring on a mandatory time-to-live so stale trust cannot silently persist, invalidated if the canonical reference is found to have been poisoned, and invalidated if authorship attestation on a source record fails. Every issuance and transition should be signed and chained. This is the point we would most encourage the FSB to draw out, because it makes the existing supervisory vocabulary work for non-deterministic systems. A grade lifecycle of this kind is the documentary instantiation of effective challenge, the principle SR 26-2 retains from SR 11-7 and on which examiner review of model risk management depends. Every output is challenged, the challenge is scored, and the record is verifiable after the fact by a party who trusts neither the institution's operator nor the model provider.

THE ACTION AND IDENTITY BOUNDARY. Each agent holds a cryptographically attested identity. The scope of actions it may take (action class, target, time window, policy epoch) is fused into the credential itself rather than checked in bypassable application code. Credentials carry short mandatory expiries. Revocation propagates in milliseconds across independent channels, and a single epoch increment can cryptographically invalidate every outstanding credential at once, with no per-token cleanup and no revocation list to distribute. Machine-speed lateral movement is won or lost on revocation time; mass revocation is the resilience control that bounds blast radius when an agent is hijacked. Under this model, policy violation becomes cryptographically equivalent to credential forgery rather than a rule an agent can route around. This also addresses the report's shadow AI concern directly: agents without enrolled identities simply cannot transact.

Three further elements belong in the final report.

First, a compositional trust rule for chained and multi-agent inference: the trust assigned to a derived output should be bounded by the minimum trust of its upstream inputs, enforced at inference time. Without this rule, a low-trust source can be laundered into a high-trust conclusion by passing it through additional inference steps. With it, contamination anywhere in the chain caps the output. This is also an operational resilience control in the sense of Basel III Pillar 2: it is a runtime invariant, enforced continuously and evidenced by signature, that a compromised upstream source cannot propagate unearned trust downstream. Relatedly, the ability to revoke trust grades in production under an administrative scope allows an institution to invalidate compromised AI-derived conclusions without taking the system down, which is the graceful-degradation property that operational resilience frameworks ask institutions to demonstrate and that a binary kill switch does not provide.

Second, an honesty norm for control claims. A 2026 IEEE Security and Privacy article by a NIST author establishes that no checker enforcing a policy over the space of inputs can be made robust, and that the limit holds for the finite context of real systems. No single control, internal or external, catches every adversarial case. Sound practice is therefore coverage of characterized failure classes with evidence, fail-closed defaults, and fast revocation, composed across independent boundaries so that an attack that slips one control still meets another. Supervisors and institutions evaluating vendor claims should treat assertions of complete robustness as a negative signal, because such claims assert precisely the robustness that has been proven unavailable.

Third, on sovereignty and proportionality: controls of this kind can run entirely inside the institution's own boundary, with the canonical reference, the policy, the trust registry, and the audit chain remaining local, including in air-gapped deployment. Integration is a configuration change rather than a re-architecture, since the governance layer sits between the existing application and the existing provider. This matters for the report's cross-jurisdictional audience, because adopting a governance layer need not require exporting the institution's most sensitive governance data to a third party or another jurisdiction, and it matters for proportionality, because the barrier to adoption is configuration rather than capital.

We recommend these elements under Sound Practices 3, 9, 10, 11, and 12.

4. **Are the sound practices sufficiently flexible to accommodate and address newer types of AI and responsible adoption over time?**

Partly. The framing is technology-neutral and outcome-oriented, which is correct and gives the toolkit good durability. The flexibility risk lies elsewhere: several practices are implicitly anchored to assumptions about AI systems that newer forms are already invalidating, and the report will age at the rate those assumptions age.

Three examples.

First, the practices assume that a model's evaluated behavior is a stable property that can be assessed at selection and relied upon thereafter. For any model whose weights are accessible or fine-tunable, that assumption is already false, and it becomes more false as open-weight frontier models proliferate. A practice framed as "evaluate the model's safety behavior" ages badly. A practice framed as "identify which relied-upon properties are enforced outside the model, and treat in-weight properties as non-durable" ages well and remains correct for AI architectures that do not exist yet.

Second, the practices assume oversight can be exercised on outputs. Agentic systems already take hundreds of intermediate actions per task, as the report itself observes, and the trajectory is toward more autonomy and longer horizons, not less. A practice framed around reviewing outputs ages badly. A practice framed around governing boundaries (what a system is permitted to do, and what may cross into a record or out through a channel) remains correct regardless of how the interior of the system evolves.

Third, the practices assume the relevant cryptographic assumptions hold. Given the implementation horizon of this toolkit and the compression of quantum timelines, crypto-agility should be an explicit expectation rather than an implicit one.

The general principle we would offer is that durability comes from specifying invariants rather than mechanisms. Controls anchored to properties of a particular AI technology (this model class, this modality, this level of autonomy) require revision each time the technology moves. Controls anchored to boundaries and invariants (identity is attested and revocable; authority is scoped and expires; output is verified against a governed reference and graded against the applicable regime; trust is bounded by the minimum of its inputs; decisions are recorded tamper-evidently; verification paths fail closed) do not require revision, because they make no assumption about what is inside the boundary. They held for traditional machine learning, they hold for GenAI, they hold for agentic AI, and they will hold for whatever follows.

We therefore recommend the final report state, as a cross-cutting principle, that institutions should prefer controls whose correctness does not depend on the internal characteristics of the AI system being governed, and should re-examine any control whose validity rests on an assumption about model behavior, model access, or model autonomy, since those assumptions have the shortest half-life.

5. **Do the case studies in this report sufficiently highlight how different types of financial institutions can benefit from responsible AI adoption? Are there additional case studies for inclusion in the report? If so, please provide such case studies, particularly for nonbanks.**

The case studies are useful and well chosen for what they cover, but they are uniformly success narratives drawn predominantly from large banks and insurers. Two consequences follow. Non-bank and smaller-institution readers see little that maps to their scale or their control budget, and no reader sees a realized failure, which is where the most transferable lessons in this field currently sit.

We offer three case studies and are prepared to provide supporting detail under the FSB's processes.

CASE STUDY OFFER 1 (non-bank; AI data infrastructure). A production RAG security deployment in which every document chunk entering the vector database is encrypted with a unique derived key; PII entities are replaced with irreversible keyed fingerprints, so no substitution dictionary exists to steal; decrypt operations are gated by role-based policy with multi-factor and hardware attestation; and an immutable hash-chained audit log records every operation. Independently verified end-to-end pipeline overhead: negative 0.39 percent versus the unprotected baseline, averaged over five sequential runs. The case study rebuts the performance objection that has historically excused plaintext AI data layers, and it illustrates proportionality well, since the same controls run from a laptop-class deployment to enterprise clusters.

CASE STUDY OFFER 2 (cautionary; supply chain and biometrics). A structured analysis of the March 2026 Mercor incident, tracing the kill chain from poisoned open-source dependency, to credential harvest, to VPN lateral movement, to exfiltration of roughly 4 terabytes including irreplaceable biometric data, mapped against Sound Practices 7, 11,

and 12, showing at each stage which architectural control would have broken the chain. A report whose case studies are exclusively success narratives leaves the most instructive material on the table.

CASE STUDY OFFER 3 (output-boundary governance and trust grading in live deployment). A live production governance layer in the inference path of commercial LLM providers, operating across more than one provider simultaneously with per-tenant isolation. Its properties, each mapped to a sound practice the report already contains:

Every model output is verified at inference time against a tenant-isolated, version-chained, individually signed canonical reference. Drifting claims resolve through configurable modes: annotate, render alongside the canonical definition, override, or refuse outright. Verification runs in microseconds and requires no additional model call. (Sound Practice 8.)

Every output receives a quantitative trust grade computed from that drift, the regulatory profile applicable to the tenant, the policy version in force, and the grades of upstream sources it depends on. Output below the tenant's admission threshold is held. The regulatory profile is an enumerated, versioned setting (basel_iii, dodd_frank, itar, ear, cmmc, fisma, default) bound immutably to both the response and the grade, so the regime and the policy version that governed any past decision remain recoverable. Grades carry a full lifecycle (issued, revoked, expired on a mandatory time-to-live, invalidated on detected poisoning of the reference, invalidated on failed authorship attestation), and every transition is signed and chained. This is effective challenge, discharged automatically, at machine speed, with an examiner-reviewable record. (Sound Practices 3 and 9; SR 26-2 and its SR 11-7 antecedents.)

Trust composes across multi-agent chains under a minimum-of-inputs bound enforced at inference time, so a low-trust source cannot be laundered into a high-trust conclusion through additional inference steps. Grades can be revoked in production under an administrative scope, invalidating compromised conclusions without taking the system down. (Basel III Pillar 2 operational resilience.)

Submitted documents and code artifacts pass a multi-layer audit that includes citation-format verification and detection of fabricated references and identifiers. For institutions whose AI-assisted output includes regulatory filings, credit memoranda, and research, fabricated citations are among the most consequential and least visible failure modes, and they are detectable structurally rather than by reader vigilance.

External reference sources carry trust ceilings tied to their verification level, so anything derived from an unverified third-party source is capped at a low trust level regardless of how confident the model sounds. (Sound Practice 12.)

Every consequential decision is recorded in a hash-chained, append-only, externally anchored audit log, signed with a 30-byte signature carrying an unconditional forgery bound that does not weaken against a quantum adversary, and verifiable by any party holding the public verification key without trusting the deployment operator or the model provider. (EU AI Act Article 12; Sound Practice 3.)

On cost and proportionality, which is where this case study earns its place: signing runs at roughly 72,000 signatures per second per processor core, so a gateway handling 1,000 inferences per second consumes about 1.4 percent of one core for its entire cryptographic audit workload. Integration is a single configuration change, repointing an existing client at the gateway, with no model retraining and no proprietary inference stack to operate. The canonical reference, policy, trust registry, and audit chain remain inside the customer boundary, and air-gapped operation is supported.

The case study demonstrates that the compensating controls we recommend under Sound Practices 8, 9, and 10 are implementable at institutional traffic volumes today, at a cost that does not register against inference cost, and that the enhanced GenAI and agentic logging called for in Sound Practice 3 can be produced as cryptographic evidence rather than as trusted assertion.

6. Do the case studies in this report provide actionable insights for financial institutions in their responsible AI adoption?

Partially. They are actionable as evidence that AI adoption produces measurable benefit, and the operational detail (processing times, adoption rates, loss reduction) is credible and useful. They are less actionable as guidance for building the control environment, for four reasons that are straightforward to remedy.

First, the case studies describe outcomes but rarely the controls that made the outcome safe. A reader learns that a bank's agentic fraud system helped update three quarters of its card fraud rules and cut fraud losses by over 20 percent, and that human-in-the-loop review approves new rules. A reader does not learn what happens when the agent proposes a rule the reviewers accept but that is wrong, how the agent's authority is scoped, what systems it can reach, how quickly it can be revoked, or what record survives if the incident must be reconstructed for a supervisor. The transferable content of a case study is the control design, not the benefit figure.

Second, there are no cost or latency figures for the controls themselves. Institutions weighing whether to adopt encryption, verification, grading, or attestation controls consistently cite performance and cost as the blocker, and the report leaves that objection unanswered. Case studies that state control overhead quantitatively would remove it. Two figures from our own deployments are offered in that spirit: record-level encryption of a RAG pipeline measured at negative 0.39 percent end-to-end overhead, and cryptographic signing of every audit record consuming approximately 1.4 percent of one processor core at 1,000 inferences per second. Numbers of this kind change the conversation from whether an institution can afford the control to why it has not deployed it.

Third, no case study describes a failure, a near-miss, or a rollback. The report explicitly asks institutions to perform root-cause analysis and post-incident review, and to treat anomalies, drift events, and near-misses as triggers to refine controls. A set of case studies containing none of these models the opposite behavior. One or two candid failure cases would be more actionable than the entire current set, and would give boards permission to surface their own.

Fourth, the case studies skew to large, internationally active institutions with dedicated data science teams. The proportionality principle is stated in the text but not demonstrated in the examples. A case study showing a smaller institution or a non-bank achieving an acceptable control posture with modest resources, for instance by adopting a governance layer as a configuration change rather than a build, would make proportionality concrete rather than aspirational.

We suggest that each case study in the final report close with a short "controls in force" note: what was scoped, what failed closed, what was logged, what it cost, and what the institution would do differently. That single addition would convert the case studies from evidence of benefit into templates for adoption.

7. Are the definitions in the glossary clear and aligned with industry sound practices, including recent developments in AI?

The definitions present are clear and consistent with industry usage. The gap is coverage: the glossary does not define the control and threat concepts that the report's own risk analysis invokes, which leaves institutions without shared language for the mitigations. We recommend the following additions.

Data-layer encryption. Encryption applied to individual records before storage and persisting through the application layer, as distinct from transport encryption (TLS) or disk encryption, neither of which protects data in use.

Irreversible de-identification. A one-way cryptographic transformation of personally identifiable information that preserves deterministic matching for retrieval while precluding recovery of the original value by any party, including administrators. To be distinguished from reversible substitution, which relies on a stored mapping that is itself a single point of compromise.

Cryptographic data lineage. Provenance records that are hash-chained and tamper-evident, such that alteration is detectable by verification rather than by trust in the record keeper.

Post-quantum cryptography, and harvest-now-decrypt-later. Neither term appears in the report, despite an implementation horizon extending past 2030.

Machine identity. A cryptographically verifiable, hardware-anchored identity assigned to a non-human actor such as an AI agent, enabling scoped authorization and revocation of that actor.

Canonical reference (or governed ontology). An institution-controlled, versioned, signed source of truth for definitions, dates, citations, named entities, and other invariants the institution requires to be preserved, against which AI output can be verified at inference time.

Ontology poisoning. Adversarial modification of the institution's canonical reference, such that AI output is verified against a corrupted source of truth. Distinct from training-data poisoning, and requiring its own integrity controls, since an institution that verifies against a poisoned reference has automated its own error and will do so with high confidence and a clean audit trail.

Regulatory profile. An enumerated, versioned attribute of an institution's AI policy identifying the regime under which a given inference is governed (for example Basel III, Dodd-Frank, ITAR, EAR, CMMC, FISMA), resolved at inference time and bound to both the output and its trust record, so that the regime and policy version governing any past decision remain recoverable.

Trust grade (or output attestation). A quantitative trust value assigned to an AI output at inference time, computed from the output's drift against the canonical reference, the applicable regulatory profile, the policy version in force, and the trust of upstream sources on which it depends, with an institution-set threshold below which output is not admitted.

Grade lifecycle. The set of states a trust grade may occupy (issued, revoked, expired, invalidated) together with the signed record of every transition between them. A grade lifecycle is the mechanism by which the model risk management principle of effective challenge can be discharged for non-deterministic systems, since every output is challenged, the challenge is scored, and the scoring is reviewable after the fact.

Semantic drift. Divergence of AI output from canonical reference that accumulates across successive interactions while each individual output remains well-formed, and which is therefore not detectable by review of any single output.

Fail-closed. The property that a verification or authorization path which errors denies rather than defaults to allow.

Trust laundering. Elevating the apparent trustworthiness of a low-trust source by passing it through additional inference steps; prevented by bounding the trust of a derived output at the minimum of its upstream inputs.

Path-level detection. Detection that operates on the geometry or trajectory of an inference rather than on its output text, identifying adversarial steering and injection by structural fingerprint rather than by matching a known-bad signature, and therefore capable in principle of flagging attack classes that have not yet been characterized.

Each term names a control or threat concept the report already describes in substance. Defining them would let institutions and supervisors converge on the same vocabulary.

8. Are there any comments you would like to make on this report? Please fill in.

We commend the FSB, the SRC Workstream on Artificial Intelligence, and the drafting team. The 12 sound practices and the accompanying case studies are the most operationally useful supervisory material yet published on financial-sector AI adoption, and the risk analysis in Section 1.3 is candid where much industry material is not.

Our central submission is twofold.

First, for the most consequential risk families the report identifies (exposure of AI data layers, credential-enabled supply-chain and deepfake attacks, and the coming quantum transition) proven cryptographic and architectural controls exist that are categorically stronger than the governance, documentation, and monitoring measures the sound practices currently emphasize, and these controls now carry no material performance penalty. We have given

figures where we have them, because the performance objection is the one that has historically stopped adoption and it no longer survives contact with the numbers.

Second, for GenAI and agentic AI specifically, two assumptions embedded in the current control set have formal counterexamples. The assumption that a model's evaluated safety behavior is durable is refuted by the demonstration that such behavior is removable from the weights at negligible cost. The assumption that output review can detect consequential failure is refuted by the formal result that cumulative semantic corruption is invisible to single-output review, combined with the empirical measurement of that corruption at 25 to 50 percent over 20 delegated edits, with 80 percent of the damage arising from sparse compounding errors that sampling will systematically miss. Both are load-bearing assumptions, and both fail.

The durable trust boundary for institutional AI inference is external to the model: verify output against a governed canonical reference, grade it against the applicable regulatory profile, bound composed trust at the minimum of its inputs, scope and revoke agent authority cryptographically, fail closed, and keep tamper-evident evidence.

We would emphasize one implication for supervisors in particular. The existing supervisory vocabulary does not need to be replaced for AI. It needs a mechanism. Effective challenge, the principle United States interagency model risk guidance retains under SR 26-2 and on which examiner review depends, can be discharged for non-deterministic systems by grading every output at inference time against the institution's canonical reference and applicable regulatory regime, and by signing and chaining every grade transition so the record is verifiable after the fact. Operational resilience in the sense of Basel III Pillar 2 is served by enforcing, as a runtime invariant with signed evidence, that a compromised upstream source cannot propagate unearned trust downstream, and by the ability to revoke trust in production without taking the system down. EU AI Act Article 12 logging is served by a hash-chained, externally anchored audit record. NIST AI RMF and the AI 600-1 generative profile map to the same layers. The mapping runs one way: an architecture of this kind produces the artifacts each regime already asks for, with cryptographic verification that the artifacts are authentic, and an institution adopting it does not need to translate between regimes.

Naming these control classes in the final report would give boards and senior management a concrete engineering target rather than only a process target, without compromising the report's technology-neutral stance. The recommendation is to govern at the boundary, grade what the model produces, encrypt the data layer, eliminate standing credentials, and adopt quantum-resistant cryptography. It is not a recommendation to adopt any particular vendor.

Supporting material. Several of our responses draw on formal results published in R. Blech, External Invariants: A Cryptographic Trust Architecture for Institutional AI Inference, Version 2.0, May 2026, DOI 10.5281/zenodo.20092349, which sets out the trust-composition bound, the drift metric, and the audit-chain bounds referenced above, and maps them to SR 26-2, NIST AI RMF and AI 600-1, EU AI Act Article 12, CISA SSDA, and Basel III Pillar 2. We are sending it to fsb@fsb.org as supplementary material alongside this response.

Our key-derivation primitive has been independently evaluated by the University of Luxembourg (cryptanalysis; no attacks found; output entropy 7.998 bits per byte), California

Polytechnic State University San Luis Obispo (Dieharder version 3.31.1; 99.4 percent aggregate across 98 tests), and George Mason University (SENTINEL evaluation, project FP5223). Our cryptographic modules operate with FIPS 140-3 validated providers (CMVP certificates 4631 and 4816).

XSOC Corp would welcome the opportunity to brief the SRC Workstream on Artificial Intelligence, to contribute the case studies described in our response to Question 5, or to provide technical detail on any point in these comments. We consent to publication of this response on the FSB website.

External Invariants: A Cryptographic Trust Architecture for Institutional AI Inference

Richard Blech

Chief Executive Officer and Technical Lead
XSOC CORP, Irvine, California, USA
rblech@xsoccorp.com
ORCID: 0009-0003-4540-2134
DOI: 10.5281/zenodo.20092349

Version 2.0, May 2026

Version Notes. Version 2.0 (May 2026) updates Section 6 regulatory mapping to reflect SR 26-2 (Federal Reserve, Office of the Comptroller of the Currency, Federal Deposit Insurance Corporation, April 17, 2026), which supersedes SR 11-7. The mapping is sharpened by SR 26-2 footnote 3, which explicitly excludes generative AI and agentic AI from the formal scope of model risk management guidance while affirming that institutions must still establish appropriate governance for these systems. Section 6 also adds reference to NIST AI 600-1 (Generative Artificial Intelligence Profile, July 2024) as a companion to NIST AI RMF 1.0. The architectural argument and all formal results of Theorems 1 through 4 are unchanged from Version 1.0 (Zenodo, May 2026). The concept DOI 10.5281/zenodo.20092349 resolves to this version; the Version 1.0 DOI continues to resolve to the original release.

Abstract

Large language models are now deployed in regulated industries where decisions made on machine-generated output have legal, financial, and operational consequences. The default trust architecture for such deployments is a stack of in-model defenses: safety fine-tuning, refusal training, output classifiers, and prompt filters. We argue that this default is structurally insufficient and develop the case in three parts, with formal results supporting each step. First, we formalize a recent mechanistic-interpretability result (Arditi et al., NeurIPS 2024) that refusal behavior across thirteen open-source chat models up to 72 billion parameters is mediated by a one-dimensional subspace of the residual stream, and we show that any property shown to be mediated by such a subspace is recoverable through rank-one weight orthogonalization with bounded compute and bounded capability loss. Second, we present a formal model of stateful semantic corruption across delegated edits, grounded in the DELEGATE-52 benchmark (Laban, Schnabel, Neville, 2026), and show that single-output classifiers operating without access to canonical reference, prior state, or edit history cannot detect corruption that is constituted by the relationship between many outputs and an external source of truth. Third, we propose an external invariant architecture organized around five governing layers (Identity, Ontology, Trust, Providence, Interdiction) and develop the mathematical foundations for each: a compositional trust bound on Epistemic Integrity Attestation scores, a formal drift metric for Semantic Stable Reference Lookup, a Providence audit chain combining an information-theoretically bounded signature primitive (Pr[forge] $\leq 2^{-128}$ unconditional for the QSIG primitive) with hash-chain

and Merkle-anchoring mechanisms under stated cryptographic and timestamping assumptions, and five geometric primitives for adversarial detection grounded in Fisher information geometry, Hamiltonian dynamics, variational free energy, online curvature anomaly detection, and information geometry. We describe a reference implementation in live deployment against commercial LLM providers and tenant-isolated contexts, and we map the architecture to existing institutional frameworks for model risk management, AI risk management, and AI Act logging. The contribution is mathematical: an architectural argument that the trust boundary belongs outside the model by construction, with formal bounds that make the claim verifiable.

Keywords: AI governance, cryptographic accountability, semantic drift, information geometry, audit chain, model risk management, post-quantum signatures, AI Act, regulatory compliance, agent safety, alignment

1 Introduction

Institutional adoption of large language models has outrun the trust architecture available to support it. Banks use them to draft credit decisions. Hospitals use them to summarize patient records. Government agencies use them to triage citizen inquiries. Law firms use them to review filings. In each setting, a decision made on machine-generated output now has consequences that were previously the responsibility of a human professional whose work could be audited, challenged, and verified through established institutional process.

The architecture deployed in production today consists of in-model defenses inherited from the chatbot moderation problem: safety fine-tuning, refusal training, post-hoc output classifiers, and prompt-level input filters. These mechanisms were developed to keep a conversational assistant from producing harmful text. They were not designed to underwrite institutional trust, and they cannot.

This paper develops the argument in three parts. Section 2 establishes that in-model alignment is structurally insufficient by formalizing the Arditi et al. (2024) result on refusal as a one-dimensional subspace and deriving its implication for any trust property implemented in the weights. Section 3 establishes that output-level moderation is structurally insufficient by developing a formal model of stateful semantic corruption, grounded in the DELEGATE-52 benchmark. Section 4 introduces the external invariant architecture and develops the mathematical foundations for each of its five layers, including a compositional trust bound, a formal drift metric, cryptographic bounds for the audit chain, and five geometric primitives for adversarial detection. Section 5 describes a reference implementation in live deployment against commercial LLM providers. Section 6 maps the architecture to existing institutional frameworks. Section 7 reviews related work. Section 8 discusses limitations. Section 9 concludes.

The contribution is architectural and mathematical. We do not propose a new alignment technique, a new prompt engineering pattern, or a new fine-tuning method. We propose that the trust boundary for institutional AI inference is structurally external to the model, that this can be instantiated as a cryptographically anchored governance fabric with formal bounds on every consequential property, and that institutional adoption of AI inference depends on such a boundary existing.

2 The Mathematical Insufficiency of In-Model Alignment

2.1 The Refusal Direction as a One-Dimensional Subspace

Let M be a transformer-based language model with L layers, hidden dimension d_{model}, and parameter set θ. For each layer $\ell \in \{1, \ldots, L\}$ and each token position i, let $x_i^{(\ell)}(t) \in \mathbb{R}^{d_{\text{model}}}$ denote the residual stream activation at that layer and position when the model is run on input token sequence t.

Let $\mathcal{D}_{\text{harmful}}$ and $\mathcal{D}_{\text{harmless}}$ be sets of harmful and harmless instruction prompts respectively. Let $\mathcal{I}$ be the set of post-instruction token positions. The difference-in-means vector at layer ℓ and post-instruction position $i \in \mathcal{I}$ is

$$r_i^{(\ell)} = \frac{1}{|\mathcal{D}_{\text{harmful}}|} \sum_{t \in \mathcal{D}_{\text{harmful}}} x_i^{(\ell)}(t) - \frac{1}{|\mathcal{D}_{\text{harmless}}|} \sum_{t \in \mathcal{D}_{\text{harmless}}} x_i^{(\ell)}(t). \tag{1}$$

The empirical result of Arditi et al. (2024), which we restate as a definition, is that for thirteen safety-fine-tuned chat models in the Llama, Qwen, Yi, and Gemma families, with parameter counts ranging from 1.8 to 72 billion, there exists a layer ℓ^* and position i^* such that the unit vector $\hat{r} = r_{i^*}^{(\ell^*)} / \|r_{i^*}^{(\ell^*)}\|$ has the following two properties simultaneously, with high probability over evaluation prompts:

Property 1 (Sufficiency for refusal). Adding $r_{i^*}^{(\ell^*)}$ to all token positions of layer ℓ^* activations of a harmless input induces refusal on the resulting output.

Property 2 (Necessity for refusal). Ablating $\hat{r}$ from every residual stream activation at every layer and every token position prevents the model from refusing harmful inputs.

The ablation operation is

$$x' \leftarrow x - \hat{r}\hat{r}^\top x, \tag{2}$$

applied to every $x_i^{(\ell)}$ and to every intermediate residual stream activation.

2.2 Equivalent Rank-One Weight Modification

The activation-level intervention of Equation (2) is equivalent to a structural modification of the model's weights. For each weight matrix $W_{\text{out}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{input}}}$ that writes to the residual stream (the embedding matrix, all attention output matrices, all MLP output matrices, and any associated bias vectors), define the orthogonalized matrix

$$W'_{\text{out}} = W_{\text{out}} - \hat{r}\hat{r}^\top W_{\text{out}}. \tag{3}$$

The operation in Equation (3) is a rank-one modification: the modification matrix $\hat{r}\hat{r}^\top$ has rank one. The transformation is linear and applies independently to each weight matrix that writes to the residual stream.

Proposition 1. *Under directional ablation of the refusal direction at all layers and all token positions, applying Equation (3) to every residual-stream-writing weight matrix produces a model whose forward pass is functionally equivalent to the activation-ablated model.*

The proof is a direct algebraic verification: the residual stream is a sum of contributions from weight matrices, each of which has been modified to never write to the $\hat{r}$ direction. Since the residual stream cannot accumulate a $\hat{r}$ component, and since previous contributions are also orthogonal to $\hat{r}$ by induction, the modified weights produce the ablated trajectory directly. We refer the reader to Arditi et al. (2024), Appendix E, for the full algebraic equivalence.

2.3 Compute Cost and Capability Preservation

Two further empirical results from the same study bear on the structural insufficiency argument. First, the compute cost of constructing the modification of Equation (3) is bounded above by the cost of running the model on a small set of harmful and harmless prompts, $|\mathcal{D}^{\text{train}}_{\text{harmful}}| + |\mathcal{D}^{\text{train}}_{\text{harmless}}| = O(10^2)$ forward passes, plus the cost of computing the difference-in-means and the orthogonalization. Arditi et al. report a total cost under five United States dollars for a 70-billion-parameter model on commodity inference hardware.

Second, the orthogonalized model retains its general capabilities. On standard language model evaluations (MMLU, ARC, GSM8K, HellaSwag), the orthogonalized model's performance

lies within the 99 percent confidence interval of baseline performance for almost all evaluations. Loss on standard text corpora and on chat-distribution completions changes by less than 0.05 nats per token in most cases.

2.4 The Structural Implication

Combining Proposition 1 with the empirical results yields the structural claim of this section.

Theorem 1 (Brittleness of in-model trust properties)**.** *Let P be a behavioral property of a model M shown to be mediated by a low-rank subspace of the residual stream of dimension k. Then there exists a rank-k modification of the model's weights that disables P at compute cost bounded above by the cost of one forward pass over a representative dataset, with capability loss bounded below the standard evaluation noise floor.*

For $k = 1$, as established empirically for refusal across thirteen models, the disabling modification is a rank-one orthogonalization with cost under five dollars on a 70-billion-parameter model.

Corollary 1. *For any trust property whose institutional guarantee depends on P, where P has been shown or can be reasonably assumed to be mediated by a low-rank residual-stream subspace, the trust property is, by construction, removable from the deployed model by any party with weight access at bounded cost. The set of parties with weight access includes anyone who can download an open-weight model, any cloud customer with a fine-tuned deployment, any insider at a model-hosting organization, and any party who can induce or exploit a supply chain vulnerability that exposes weights.*

2.5 The Conclusion

A trust boundary, in the sense relevant to institutional AI deployment, is a place at which trust assumptions are converted into verifiable properties. In-model alignment provides no such verification: the property cited (refusal, policy compliance, content filtering) is not a property of the architecture, the training methodology, or the alignment procedure. It is a property of one specific copy of one specific weight tensor, which can be modified to lack the property by a rank-one operation at bounded cost.

The trust boundary for institutional AI inference therefore cannot live in the weights. The architecture must be such that the trust property does not depend on the weights at all.

3 The Mathematical Model of Stateful Corruption

The argument from Section 2 establishes that the trust boundary cannot be the weights. This section establishes that the trust boundary cannot be the immediate output either, even if that output is inspected by classifiers, scanned by content filters, or tested against safety policies. The argument is that the failure mode that matters for institutional inference is not unsafe wording in a single output. It is stateful semantic corruption that compounds across multiple inferences while every individual output remains a well-formed example of its surface form.

3.1 The Delegated-Edit Process

Let D_0 denote an initial document, modeled as a finite sequence of statements over a domain ontology. Let $\{c_1, c_2, \ldots, c_T\}$ be a sequence of editing instructions issued by a user. A delegated-edit process is the application of a language model M to produce $D_t = M(D_{t-1}, c_t)$ for $t = 1, 2, \ldots, T$.

For each statement s in D_t, let $\sigma_v(s) \in \{0, 1\}$ be the indicator that s remains semantically valid relative to a canonical reference V (the "ground truth" of definitions, dates, citations, named entities, and other invariants the institution requires to be preserved). Let

$$\rho_t = 1 - \frac{1}{|D_t|} \sum_{s \in D_t} \sigma_v(s) \tag{4}$$

denote the corruption rate at step t, the fraction of statements in D_t that have drifted from canonical reference.

3.2 The DELEGATE-52 Empirical Result

The DELEGATE-52 benchmark (Laban, Schnabel, Neville, 2026) measures ρ_T for $T = 20$ delegated interactions across 19 language models in 310 professional document-editing environments. The reported results are:

$$\rho_{20}^{\text{frontier}} \approx 0.25, \qquad \rho_{20}^{\text{average}} \approx 0.50, \tag{5}$$

where "frontier" refers to Gemini 3.1 Pro, Claude 4.6 Opus, and GPT 5.4, and "average" is across all tested models. These rates are consistent across the 310 environments and across multiple independent runs.

The benchmark also reports that ρ_t scales non-trivially with context length. As context grows from 4,096 to 32,768 tokens, ρ_t rises non-linearly. This is consistent with the structural property that transformer attention does not allocate uniformly across long contexts; information at the start of the context becomes functionally inaccessible to a later inference even though the tokens are nominally present.

A third reported finding is that agentic frameworks amplify ρ_t. When the model operates within an agent harness with tool use, each tool call consumes context budget. The original document is forced further from the active attention window, and the per-edit corruption probability rises accordingly.

3.3 Single-Output Detectability

We now formalize the claim that single-output classifiers cannot detect this corruption. Let Φ be any classifier that, given a single output D_t, returns a binary verdict $\Phi(D_t) \in \{0, 1\}$ indicating whether the output is acceptable.

Proposition 2 (Single-output undetectability under observational limits)**.** *For any classifier Φ operating only on the single output D_t, without access to the initial document D_0, the canonical reference V, the prior history $D_0, \ldots, D_{t-1}$, or any relational consistency model linking outputs across edit steps, there exist delegated-edit sequences $\{c_1, \ldots, c_T\}$ for which every individual output is admissible under Φ (that is, $\Phi(D_t) = 1$ for all t) while cumulative semantic corruption ρ_T exceeds any fixed threshold.*

The proof is constructive. Each c_t is chosen to introduce a small, individually plausible modification: a date shift, a digit transposition in a citation, a silent redefinition of a term used in an earlier section. Each D_t remains well-formed under any classifier that operates without reference to V. The cumulative effect after T steps is ρ_T corruption, but no individual D_t is anomalous on its own. This is precisely the failure mode DELEGATE-52 quantifies empirically.

3.4 The Structural Implication

A trust architecture that operates by inspecting the immediate output of an inference cannot, by Proposition 2, detect failures that are constituted by the relationship between many inferences and an external source of truth. To detect such failures, the architecture must:

1. Hold the canonical reference V as an external invariant.
2. Verify each output D_t against V at inference time, not at audit time.
3. Maintain a chain of evidence linking each D_t to its predecessors.
4. Bound the trust score of D_t by the trust scores of its inputs.

These requirements cannot be satisfied by any architecture in which the trust boundary is the immediate output. They require an architecture in which the trust boundary holds invariants outside the model.

4 External Invariants as Trust Architecture

The negative results of Sections 2 and 3 establish that the trust boundary cannot be the weights and cannot be the immediate output. This section presents the corresponding positive architecture. We propose that the trust boundary for institutional AI inference is best instantiated as a five-layer external invariant architecture: Identity, Ontology, Trust, Providence, and Interdiction. We develop the mathematical foundations for each layer in turn.

4.1 Identity

The Identity layer binds each request to a session whose authority can be cryptographically verified independently of any trust assumption about the model or the model provider. We require three properties.

Property I.1 (Attestation). Each session is established through a platform attestation that produces a cryptographic credential bound to a hardware-rooted authenticator.

Property I.2 (Locality of secrets). The cryptographic operations on user-side secrets occur in an isolated execution context (a hardware-rooted enclave or a sandboxed WebAssembly module on the user's device). The gateway receives an attestation token but never holds the underlying secret material.

Property I.3 (Real-time revocation). Each request authorization checks the session against a revocation registry. A revoked session rejects the next request regardless of the recency of the prior authorization.

The cryptographic primitive backing the session attestation in our reference implementation is a deterministic symmetric key agreement (DSKAG) that produces per-request key material from a hardware-bound seed without exposing the seed to the network. The DSKAG primitive is described in detail in Blech (2026b); for this paper we treat it as a black box, relying only on its interface and security properties, with no disclosure of internal implementation.

Definition 1 (DSKAG)**.** *A deterministic symmetric key agreement primitive is a function $\mathcal{K} : \mathcal{S} \times \mathcal{C} \to \mathcal{K}$ where $\mathcal{S}$ is the seed space, $\mathcal{C}$ is the context space (request identifier, timestamp, role identifier), and $\mathcal{K}$ is the key space. The function satisfies:*

(a) Determinism: $\mathcal{K}(s, c) = \mathcal{K}(s, c)$ for all (s, c).

(b) Indistinguishability: For any seed s and any pair of distinct contexts $c_1 \neq c_2$, the keys $\mathcal{K}(s, c_1)$ and $\mathcal{K}(s, c_2)$ are computationally indistinguishable from independently sampled uniform random keys, given any oracle access short of s itself.

(c) Forward secrecy: Compromise of $\mathcal{K}(s, c)$ does not yield information about $\mathcal{K}(s, c')$ for any $c' \neq c$ beyond that recoverable from observation of $\mathcal{K}(s, c)$ itself.

The implementation of $\mathcal{K}$ provides 128-bit symmetric security under the indistinguishability standard.

The Identity layer requires more than session establishment. It requires that the authority granted to a session be scoped to specific actions, targets, time intervals, and policy epochs, and that each scope dimension be revocable independently of the others. The reference implementation realizes this through the AC-AGENT-CGA construction (Blech, 2026c), which we summarize here.

An authorization scope is a tuple

$$\Lambda = (\alpha, \tau_t, [t_0, t_1], \epsilon),$$

where $\alpha \in \mathcal{A}$ is the authorized action class, $\tau_t \in \mathcal{T}$ is the target identifier, $[t_0, t_1] \subset \mathbb{R}$ is the temporal window of validity, and $\epsilon \in \mathbb{N}$ is the policy epoch. A request issued under scope Λ is admitted only if all four components match: the request action lies in α, the request target equals τ_t, the request timestamp lies in $[t_0, t_1]$, and the active policy epoch equals ϵ. The four-dimensional scoping produces fine-grained authority: a credential authorized for one action class on one target during one time window under one policy epoch cannot be reused for any other combination, even by a holder of the credential.

Property I.4 (Three-channel fail-closed revocation). Revocation operates on three independent channels: per-credential revocation invalidates the binding to the hardware authenticator, per-scope revocation invalidates one or more components of Λ, and per-epoch revocation increments the active policy epoch and invalidates all credentials issued under prior epochs. A request is admitted only if all three channels permit it; failure of any one channel rejects the request and produces a Providence record carrying a typed denial reason from a fixed vocabulary.

The formal specification of the scoping model and the revocation channels is given in Blech (2026c), with an explicit enumeration of fourteen typed denial reasons that constitute the closed vocabulary of authorization failures (out-of-scope action, target mismatch, expired window, stale epoch, revoked credential, replay attempt, and others). The fixed denial vocabulary serves two purposes: it makes Providence records over authorization failures structurally interpretable by automated audit systems, and it bounds the surface area of authorization decisions to a finite set of cases that can be exhaustively reviewed.

4.2 Ontology and Semantic Stable Reference Lookup (SSRL)

The Ontology layer holds canonical definitions per tenant in a tenant-isolated, version-chained, cryptographically signed store. We formalize the drift detection mechanism.

Definition 2 (Canonical Ontology). *For tenant τ and version v, the canonical ontology $\mathcal{O}_\tau^{(v)}$ is a finite set of triples (t_i, d_i, e_i) where t_i is a term identifier, d_i is a canonical definition, and $e_i \in \mathbb{R}^{d_e}$ is an embedding of the definition in a fixed embedding space.*

Definition 3 (Claim Extraction). *Given a model output y, claim extraction is a function $\Psi : \mathcal{Y} \to 2^{\mathcal{T} \times \mathcal{R}}$ that maps y to a set of (term, claim-about-term) pairs, where each $r \in \mathcal{R}$ is itself embedded as $e(r) \in \mathbb{R}^{d_e}$.*

Definition 4 (SSRL Drift). *For a claim $(t, r) \in \Psi(y)$, the SSRL drift relative to ontology $\mathcal{O}_\tau^{(v)}$ is*

$$\delta(r, t \mid \mathcal{O}_\tau^{(v)}) = \begin{cases} 1 - \cos(e(r), e_i) & \text{if } \exists i : t_i = t \\ \infty & \text{otherwise} \end{cases} \tag{6}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity.

The SSRL gate decision is parameterized by a threshold $\tau_\delta \in (0, 1)$ and a correction mode $\mu \in \{\text{annotate}, \text{dual\_render}, \text{override}, \text{strict\_refuse}\}$. The decision rule is

$$\text{action}(r, t) = \begin{cases} \text{admit} & \text{if } \delta(r, t \mid \mathcal{O}_\tau^{(v)}) \leq \tau_\delta \\ \mu(r, t, d_i) & \text{otherwise.} \end{cases} \tag{7}$$

The four correction modes are defined as: μ_{annotate} delivers r flagged with δ; $\mu_{\text{dual\_render}}$ delivers r alongside d_i; μ_{override} replaces r with d_i; $\mu_{\text{strict\_refuse}}$ rejects the response and produces a Providence record of the rejection.

The latency of the SSRL check is bounded by the cost of one nearest-neighbor lookup in the ontology embedding space. For ontologies of size $|\mathcal{O}_\tau^{(v)}| < 10^4$ and $d_e = 768$, this is microsecond-scale on commodity hardware with index structures such as hierarchical navigable small worlds (HNSW).

4.3 Trust and the Compositional EIA Bound

The Trust layer assigns each output a quantitative trust value, the Epistemic Integrity Attestation (EIA) score. The mathematical foundation of the Trust layer is a composition theorem that bounds the trust of a derived output by the trust of its inputs.

Definition 5 (EIA Score)**.** *For an output y produced from input x in a tenant context τ, the Epistemic Integrity Attestation score $E(y \mid x, \tau) \in [0, 1]$ is a quantitative trust value computed from (i) the SSRL drift of claims in y relative to $\mathcal{O}_\tau^{(v)}$, (ii) the regulatory profile applicable to τ, (iii) the tenant policy in force, and (iv) the trust scores of upstream sources y depends on.*

Definition 6 (Composition)**.** *A composition is a derivation $y = \mathcal{C}(y_1, \ldots, y_n; x)$ in which y depends on a set of upstream outputs $\{y_1, \ldots, y_n\}$ together with a primary input x. The dependency is causal: each y_i is consumed in the production of y either as direct input, as retrieved evidence, or as input to an intermediate step.*

The Trust layer enforces the following composition theorem.

Theorem 2 (Compositional EIA Bound)**.** *Under the trust-preservation principle, for any composition $y = \mathcal{C}(y_1, \ldots, y_n; x)$,*

$$E(y \mid x, \tau) \leq \min_{i \in \{1, \ldots, n\}} E(y_i \mid x_i, \tau). \tag{8}$$

Proof. Suppose, for contradiction, that $E(y) > \min_i E(y_i)$. Let $j = \arg\min_i E(y_i)$. Then $E(y) > E(y_j)$, but y depends causally on y_j. The composition has produced an output whose admissible trust exceeds the admissible trust of one of its causal inputs. Any institutional process that admits y at trust level $E(y)$ must therefore admit conclusions that depend on y_j at a level greater than $E(y_j)$ admits, which contradicts the trust assignment of y_j. The contradiction is removed only by $E(y) \leq E(y_j) = \min_i E(y_i)$. □

The bound of Equation (8) is conservative: tighter bounds are possible for specific composition operators (for example, products or weighted averages of independent sources can be bounded above by something stricter than the minimum). The minimum is the architecturally safest choice and is what the reference implementation enforces.

The institutional admission criterion is parameterized by a tenant threshold $\tau_E \in [0, 1]$:

$$\text{admit}(y) \iff E(y \mid x, \tau) \geq \tau_E. \tag{9}$$

Combining Equations (8) and (9) yields the operational consequence:

Corollary 2. *Under the compositional EIA bound, admit(y)* $\implies E(y_i \mid x_i, \tau) \geq \tau_E$ *for all* i*. Equivalently, if any upstream source* y_i *falls below the admission threshold, the downstream output* y *is rejected.*

This corollary closes the trust laundering attack: an adversary cannot launder a low-trust upstream source into a high-trust downstream output by passing it through additional inference steps.

4.4 Providence and Cryptographic Bounds

The Providence layer records every consequential decision in a hash-chained, append-only, Merkle-anchored audit log. Each entry carries an information-theoretic signature. The layer combines three classes of cryptographic guarantee that are distinct in their security model: an information-theoretically bounded signature primitive (QSIG), hash-chain integrity under the chosen hash function's computational security model, and Merkle anchoring under explicit trust assumptions on the timestamping authority. We develop the mathematical foundations in turn.

Definition 7 (Providence Record)**.** *A Providence record at index* k *is a tuple* $r_k = (t_k, a_k, p_k, \sigma_k)$ *where* t_k *is a timestamp,* a_k *is the action description,* p_k *is the payload (input, output, EIA score, gate decisions), and* σ_k *is the signature.*

Definition 8 (Hash Chain)**.** *The Providence chain is a sequence of records* $r_1, r_2, \ldots$ *together with a sequence of chain hashes* $h_0, h_1, h_2, \ldots$ *where* h_0 *is a fixed initialization vector and*

$$h_k = H(h_{k-1} \| r_k), \quad k \geq 1, \tag{10}$$

for a collision-resistant hash function $H : \{0,1\}^* \to \{0,1\}^n$ *with output length* n *bits. The reference implementation uses* $H =$ *SHA-256,* $n = 256$*.*

Theorem 3 (Hash Chain Integrity)**.** *Let an adversary modify any record* r_j *with* $1 \leq j \leq k$ *to a different value* $r'_j \neq r_j$*, where* h_k *has been published or anchored externally. Then the probability that the adversary can produce a sequence of replacement records* $r'_j, r'_{j+1}, \ldots, r'_k$ *such that the recomputed chain head* $h'_k = h_k$ *is bounded above by* $2^{-n/2}$ *(for collision resistance) when the adversary is computationally bounded, and is exactly* 2^{-n} *in the random oracle model for second-preimage attacks on the chain head.*

For $n = 256$, the bounds are 2^{-128} (collision) and 2^{-256} (second preimage). The conservative interpretation is that hash chain integrity is bounded by collision resistance, 2^{-128}.

Definition 9 (Merkle Anchoring)**.** *Let* $\{h_{k_1}, h_{k_2}, \ldots\}$ *be a sequence of Providence chain heads at anchor intervals (every* K *records, or every* Δt *time units). Each anchor* h_{k_j} *is published to an external timestamping authority, which records the pair* (h_{k_j}, t_{k_j}) *in a public record verifiable by anyone.*

Proposition 3. *Under Merkle anchoring with anchor interval* K *and a trustworthy timestamping authority, an adversary's window of opportunity to alter records is bounded by the most recent anchor: alterations to records* $r_{k_j+1}, \ldots, r_k$ *with* $k < k_{j+1}$ *are detectable by chain verification, and alterations to records* $r \leq r_{k_j}$ *are bounded by Theorem 3 applied to the externally anchored hash.*

The signature primitive on Providence records is XSOC-QSIG (Blech, 2026a). We restate the relevant security claim.

Definition 10 (XSOC-QSIG)**.** *XSOC-QSIG is a signature scheme producing 30-byte signatures (240 bits) that achieves unconditional existential unforgeability under adaptive chosen-message attack against computationally unbounded adversaries (Blech, 2026a). For any adversary* $\mathcal{A}$ *that*

has observed a polynomial number of valid signature-message pairs of its own adaptive choosing, the probability that $\mathcal{A}$ produces a valid forged signature on a previously unsigned message is bounded above by 2^{-128}:

$$\Pr[\textit{forge}] \leq 2^{-128}. \tag{11}$$

The bound holds unconditionally: it does not depend on any computational hardness assumption. This is the strongest formal security model for digital signatures (EUF-CMA) instantiated under information-theoretic rather than computational security (Wegman and Carter, 1981; Stinson, 2006). For comparison, NIST FIPS 204 ML-DSA (Dilithium-3) produces signatures of approximately 3.3 KB and provides EUF-CMA security against computationally bounded adversaries under the assumed hardness of module-lattice problems. The QSIG bound holds against adversaries with unbounded compute time and quantum resources.

The relevance to institutional inference is direct. EU AI Act Article 12 requires that high-risk AI systems maintain logs that are tamper-evident, timestamped, and verifiable by national competent authorities. The Providence layer with hash chaining (Theorem 3), Merkle anchoring (Proposition 3), and information-theoretic signatures (Equation 11) instantiates these requirements with bounds that hold across the multi-decade horizons institutional records may need to support.

4.5 Interdiction and the Five Geometric Primitives

The Interdiction layer detects adversarial patterns in the geometry of the inference path itself, before the surface output is admitted. The layer operates through five primitives, each grounded in a well-developed mathematical theory and each producing a scalar signal that is fused at the gate.

4.5.1 GRL: Gradient Resonance Lambda

GRL is grounded in information geometry. For an inference at step t with output distribution $p_t(y \mid x, y_{<t}; \theta)$ over the vocabulary $\mathcal{V}$, the Fisher information matrix in the parameter space (or restricted to the relevant active subspace) is

$$\mathcal{I}_t(\theta) = \mathbb{E}_{y \sim p_t}\left[\nabla_\theta \log p_t(y \mid x, y_{<t}; \theta) \nabla_\theta \log p_t(y \mid x, y_{<t}; \theta)^\top\right]. \tag{12}$$

Let $\lambda_1(\mathcal{I}_t) \geq \lambda_2(\mathcal{I}_t) \geq \cdots \geq \lambda_d(\mathcal{I}_t) \geq 0$ be the eigenvalues of $\mathcal{I}_t(\theta)$ in non-increasing order.

Definition 11 (GRL)**.** *The Gradient Resonance Lambda at step t is*

$$GRL_t = \frac{\lambda_1(\mathcal{I}_t)}{tr(\mathcal{I}_t)} = \frac{\lambda_1(\mathcal{I}_t)}{\sum_i \lambda_i(\mathcal{I}_t)}. \tag{13}$$

The quantity $\mathrm{GRL}_t \in [1/d, 1]$ measures the concentration of Fisher information on the principal direction. $\mathrm{GRL}_t = 1/d$ corresponds to perfectly isotropic information (all directions equally informative, healthy reasoning); $\mathrm{GRL}_t \to 1$ corresponds to fully degenerate manifold (a single principal direction dominates, characteristic of mode collapse, sycophantic compliance, and adversarial steering).

Decision rule. The Interdiction gate flags the inference if $\mathrm{GRL}_t > \tau_{\mathrm{GRL}}$ for a tenant-configured threshold (typical default $\tau_{\mathrm{GRL}} = 0.7$).

The relevance to adversarial detection is grounded in the mechanism documented by Arditi et al. (2024) and discussed in Section 2: adversarial suffixes hijack attention heads, producing measurable concentration of Fisher information on the suffix-attended direction. This concentration manifests as elevated GRL.

Operational note. In live deployment against API-only LLM providers where parameter-space gradients are not directly accessible, $\mathcal{I}_t$ is approximated using empirical statistics over

the output distribution available through logprob output. The principal eigenvalue ratio is computed on this empirical proxy rather than on the full parameter-space Fisher information. Calibration of τ_{GRL} to specific attack classes (mode collapse, sycophantic compliance, attention hijacking) is performed per-tenant using known-good and known-bad inference traces.

4.5.2 HJM: Hamiltonian Jacobi Manifold

HJM is grounded in Hamiltonian mechanics. The inference trajectory is treated as motion of a particle through a potential field defined by the model's surprise.

Definition 12 (Inference Hamiltonian)**.** *Let q_t be a point in the model's reasoning state space (for example, the embedding of the current generated prefix), and let p_t be the corresponding momentum (rate of change of q_t between inference steps). The Hamiltonian is*

$$H(q,p) = T(p) + V(q), \quad T(p) = \tfrac{1}{2}\|p\|^2, \quad V(q) = -\log p(y \mid x, y_{<t}; \theta). \tag{14}$$

Hamilton's equations describe the expected evolution of (q_t, p_t):

$$\frac{dq}{dt} = \frac{\partial H}{\partial p} = p, \quad \frac{dp}{dt} = -\frac{\partial H}{\partial q} = -\nabla V(q). \tag{15}$$

Energy is conserved along true Hamiltonian trajectories: $H(q(t), p(t)) = H(q(0), p(0))$ for all t.

Definition 13 (HJM Energy Violation)**.** *The HJM energy violation at step t is*

$$\varepsilon_t = |H(q_t, p_t) - H(q_0, p_0)|. \tag{16}$$

Decision rule. The gate flags the inference if $\varepsilon_t > \tau_{\text{HJM}}$.

External perturbations that inject information not present in the original input, such as indirect prompt injection through poisoned documents or tool-use side effects that contaminate context, inject energy into the trajectory and produce measurable energy violations. A Hamiltonian-conservative inference path is an inference path uncontaminated by external steering.

Operational note. q_t is realized in deployment as the embedding of the current generated state in the model's output token-embedding space; p_t as the first difference between consecutive embeddings. The Hamiltonian is evaluated on this projected trajectory rather than in the full parameter space. Energy violations are calibrated against indirect-prompt-injection traces and tool-use baseline traces collected per-tenant.

4.5.3 FEBI: Free Energy Bounded Inference

FEBI is grounded in variational free energy. Let q be a posterior over outputs given input x, and let $E(y, x) = -\log p(y \mid x; \theta)$ be the energy.

Definition 14 (Variational Free Energy)**.** *The variational free energy of the posterior q relative to the model's energy is*

$$\mathcal{F}[q] = \mathbb{E}_{y\sim q}[E(y,x)] - H[q], \tag{17}$$

where $H[q] = -\sum_y q(y) \log q(y)$ is the Shannon entropy of q.

Proposition 4 (ELBO)**.** *For any q, $\mathcal{F}[q] \geq -\log Z(x)$ where $Z(x) = \sum_y \exp(-E(y,x))$ is the partition function.*

The bound is the evidence lower bound standard in variational inference.

Definition 15 (FEBI Excess)**.** *The FEBI excess is*

$$\Delta\mathcal{F}(x, q) = \mathcal{F}[q] - (-\log Z(x)). \tag{18}$$

By Proposition 4, $\Delta\mathcal{F}(x, q) \geq 0$, with equality iff q equals the true posterior $p(\cdot \mid x; \theta)$.

Decision rule. The gate flags the inference if $\Delta\mathcal{F}(x, q) > \tau_{\text{FEBI}}$.

Excess free energy indicates one of three conditions: (i) the posterior q has not converged to the true posterior (under-conditioning, often a sign of insufficient evidence in the input), (ii) the posterior has been steered away from the true posterior (a sign of adversarial input or adversarial intermediate state), or (iii) the input x is anomalous relative to the model's training distribution.

Operational note. The partition function $Z(x)$ is intractable for any non-trivial vocabulary and is approximated in deployment by summing over the top-k tokens accessible via logprob output (typical $k \in [100, 500]$). The free energy is computed on this truncated distribution; the resulting bound is an upper bound on the true free energy and yields a corresponding upper bound on $\Delta\mathcal{F}$, sufficient for threshold-based decisions but conservative for absolute calibration.

4.5.4 CADS: Curvature Anomaly Detection via Welford

CADS is grounded in online statistical estimation. The primitive computes running estimates of mean and variance of curvature signals using Welford's algorithm and flags anomalies via z-scores.

Definition 16 (Welford Running Statistics)**.** *Given a sequence of curvature observations $\kappa_1, \kappa_2, \ldots$, the running mean M_n and running sum-of-squares S_n are updated by*

$$M_n = M_{n-1} + \frac{\kappa_n - M_{n-1}}{n}, \quad S_n = S_{n-1} + (\kappa_n - M_{n-1})(\kappa_n - M_n), \tag{19}$$

with $M_0 = 0, S_0 = 0$. The running variance is $\hat{\sigma}_n^2 = S_n/(n-1)$ for $n \geq 2$.

Definition 17 (CADS z-Score)**.** *The CADS z-score at step n is*

$$z_n = \frac{\kappa_n - M_n}{\hat{\sigma}_n}. \tag{20}$$

Decision rule. The gate flags the inference if $|z_n| > \tau_{\text{CADS}}$, with typical default $\tau_{\text{CADS}} = 3.5$.

Welford's algorithm has the property that M_n and S_n are computed in one pass with constant memory and are numerically stable under streaming updates. The applied curvature signal κ may be Ricci curvature on the model's information manifold, the GRL signal of Equation (13), the HJM violation of Equation (16), or any scalar signal whose distribution is approximately stationary in the absence of attack.

Operational note. Welford running statistics are maintained per-tenant and per-curvature-signal in constant memory. Thresholds τ_{CADS} are calibrated empirically per-tenant. The algorithm is online and stable; no approximation beyond floating-point representation is required.

4.5.5 IGT: Information Geometry Triple

IGT is grounded in information theory. Three quantities are computed jointly to characterize the information content, the input-output coupling, and the trajectory smoothness.

Definition 18 (Conditional Entropy)**.** *For output distribution $p(y \mid x)$,*

$$H(Y \mid X = x) = -\sum_y p(y \mid x) \log p(y \mid x). \tag{21}$$

Definition 19 (Mutual Information)**.** *For input distribution $p(x)$ and joint $p(x, y)$,*

$$I(X;Y) = H(Y) - H(Y \mid X) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{22}$$

Definition 20 (Jensen-Shannon Divergence)**.** *For two distributions P and Q over $\mathcal{Y}$, with $M = (P + Q)/2$,*

$$JSD(P\|Q) = \tfrac{1}{2} KL(P\|M) + \tfrac{1}{2} KL(Q\|M), \tag{23}$$

where $KL(P\|M) = \sum_y P(y) \log[P(y)/M(y)]$. The JSD is bounded: $JSD(P\|Q) \in [0, \log 2]$, symmetric in its arguments, and a true metric under the square root.

Definition 21 (IGT)**.** *At inference step t, the Information Geometry Triple is*

$$IGT_t = (H(Y_t \mid X_t), I(X_t; Y_t), JSD(P_t\|P_{t-1})) \in \mathbb{R}^3. \tag{24}$$

Decision rule. The gate flags the inference if any component crosses a tenant-configured threshold:

$$H_t > \tau_H \quad \text{or} \quad I_t < \tau_I \quad \text{or} \quad \text{JSD}_t > \tau_{\text{JSD}}. \tag{25}$$

High conditional entropy indicates uncertain output; low mutual information indicates that the output carries little dependence on the input (a sign of context loss, adversarial distraction, or generation collapse); high JSD indicates abrupt shift in the output distribution between consecutive steps (a sign of trajectory discontinuity). The three components are complementary: each addresses a different aspect of the information structure, and their joint evaluation is more informative than any one alone.

Operational note. All three quantities are computed in deployment over the truncated top-k distribution accessible via logprob output. For providers that do not expose logprobs, IGT is computed over a sampled approximation, with a corresponding loss of statistical confidence. The conditional entropy and mutual information estimators are unbiased under the truncation; the JSD estimator is conservative (a lower bound on the true divergence).

4.5.6 Fusion of Geometric Signals

The five primitives produce five scalar signals at each inference step. The Interdiction gate fuses them through a tenant-configured decision function Γ:

$$\text{flag}(t) = \Gamma\left(\text{GRL}_t, \varepsilon_t, \Delta\mathcal{F}_t, z_t, \text{IGT}_t\right). \tag{26}$$

The simplest decision function is the disjunctive one: flag if any individual primitive exceeds its threshold. More sophisticated functions take weighted sums, learned thresholds, or per-primitive false-positive calibrations. The reference implementation supports all three.

4.6 The Layer Composition Bound

The architecture's defense-in-depth property is captured by a composition theorem. Let p_i be the false-negative rate of layer $i \in \{\text{Identity}, \text{Ontology}, \text{Trust}, \text{Providence}, \text{Interdiction}\}$, defined as the probability that a request adversarially constructed to bypass the architecture passes layer i undetected.

Theorem 4 (Layer Composition Bound)**.** *Under the assumption that an adversarial attempt against layer i is statistically independent of an adversarial attempt against layer $j \neq i$ (independent attack vectors per layer), the joint bypass probability is*

$$\Pr[\textit{bypass}] \leq \prod_{i \in \textit{layers}} p_i. \tag{27}$$

Under the alternative assumption that all layers fail simultaneously under a single, undetected attack vector, the joint bypass probability is

$$\Pr[\textit{bypass}] \leq \max_i p_i. \tag{28}$$

Real attacks lie between these extremes. The independence bound (27) applies to attacks that target distinct architectural concerns (e.g., a stolen credential targeting Identity, a poisoned ontology targeting Ontology, an attention-hijacking suffix targeting Interdiction). Each such attack must be defeated independently by its target layer; success at one does not lower the bar at others. The architecture's value lies in the gap between the two bounds: for a 5-layer architecture with each $p_i = 10^{-3}$ under independence, $\Pr[\text{bypass}] \leq 10^{-15}$; under maximum correlation, $\Pr[\text{bypass}] \leq 10^{-3}$. The empirical attack landscape supports the independence assumption for distinct architectural concerns: the Arditi rank-one orthogonalization defeats refusal but does not produce a credential, does not poison ontology, does not forge a Providence record, and does not produce a low-GRL trajectory.

The architecture is structurally robust to advances in any one attack class because no single attack vector compromises more than one layer.

5 CLM: A Reference Implementation

This section describes a reference implementation of the architecture, the Cognitive Language Model (CLM), in live deployment against commercial LLM providers and tenant-isolated contexts. The purpose is to demonstrate that the architecture is implementable at the scale of real institutional traffic. Implementation details that are proprietary, specifically the internals of DSKAG and the parameters of the entropy-hardening cipher used in the storage primitive, are treated as black boxes.

5.1 Engine

The CLM engine is implemented in Rust, organized as 21 crates with 112 tests across cryptographic primitives, gate logic, ontology enforcement, grade lifecycle, audit chain, and geometric reasoning. The engine compiles as a C-compatible shared library that is consumed through a Foreign Function Interface from a Go gateway. The Go gateway implements REST and gRPC APIs, request routing, telemetry, and provider integration with OpenAI and Anthropic completion endpoints.

The cryptographic primitives consume key material from Azure Key Vault. The DSKAG primitive (Definition 1) provides per-request key derivation. The XSOC-QSIG signature primitive (Definition 10) signs every Providence record with an unconditional 2^{-128} forgery bound (Equation 11). QSIG signing is benchmarked at approximately 14 microseconds per sign on a single core.

5.2 Identity (NIE)

The Identity layer is implemented as the Nexus Identity Engine (NIE). NIE attests every session through a platform biometric (Touch ID, Windows Hello, Face ID, or platform passkey) using WebAuthn. DSKAG key derivation runs inside a WebAssembly sandbox in the user's browser, so cryptographic operations on user-side material occur on the user's device rather than on any server (Property I.2). The gateway receives an attestation token but never sees the underlying key material. Every authorization call routes through a fail-closed gateway that does not cache. Sessions are revocable in real time (Property I.3).

5.3 Ontology and SSRL

The Ontology layer holds canonical definitions per tenant in a tenant-isolated dynamic store, indexed by HNSW for sub-millisecond nearest-neighbor lookup. Each entry is signed with XSOC-QSIG at creation. Entries are version-chained: a change to a canonical definition produces a new ontology version $v' = v + 1$ with the previous version retained for retrospective verification. The SSRL drift check (Definition 4) runs on every model output. Latency is dominated by the embedding of the claim and the HNSW lookup, totaling tens of microseconds in the reference implementation. The four correction modes of Equation (7) resolve through the policy chain.

5.4 Trust and EIA

The Trust layer assigns an EIA score to every output (Definition 5). Grades have a lifecycle: issued, revoked, expired, poisoned, authorship-invalidated. Each transition is signed with XSOC-QSIG and chained into the Providence record. The compositional EIA bound of Theorem 2 is enforced at inference time: when a request depends on multiple upstream sources, the gateway computes the minimum upstream EIA and bounds the downstream EIA accordingly before the response is produced. Grade revocation operates under an administrative scope (clm:admin:critical) that itself requires elevated authentication and produces signed event records.

5.5 Providence

The Providence audit chain (Definition 8) is hash-chained with SHA-256 ($n = 256$, integrity bound 2^{-128} by Theorem 3), append-only, and Merkle-anchored at configurable intervals (Definition 9). Every consequential transition produces a Providence record signed with XSOC-QSIG. Authorization-related events draw their typed denial reasons from the closed fourteen-element vocabulary specified in Blech (2026c), which makes the resulting records structurally interpretable by automated audit consumers and bounds the surface area of authorization-failure cases. The verification model is that any party with the public verification key can verify a Providence record without trusting CLM, the storage substrate, or the LLM provider.

5.6 Interdiction

The Interdiction layer runs the five geometric primitives of Section 4.5 on every inference path, fused through the decision function of Equation (26). In addition, a six-layer document and code audit runs on every artifact submitted through the gate: ontology drift (using Definition 4), geometric anomaly (using the GRL, HJM, FEBI, CADS, IGT signals), citation format, cross-provider recognition (a citation that purports to come from one model but matches the linguistic signature of another), structural pattern, and a Mythos-class adversarial scan that targets nation-state-grade adversarial code patterns.

5.7 Operational Characteristics

The implementation is operational in live deployment against commercial LLM providers including OpenAI and Anthropic completion endpoints, with tenant-isolated contexts and active workloads. Cross-provider inference governance is operational. Tenant-isolated canonical ontology is loaded dynamically and enforced at inference time. SSRL drift detection runs on every output. The four correction modes resolve through the policy chain. The grade lifecycle including poisoning detection is fully implemented. Compositional EIA caps downstream trust at the lowest upstream score in deployment. Cross-agent validation runs in microseconds without an

LLM call. The six-layer document and code audit runs on every artifact. QSIG signs and verifies at deployment rates against real key material from Azure Key Vault. Specific throughput numbers under load and customer deployment details are out of scope for this paper.

The relevant claim for the paper's argument is that the architecture is implementable, has been implemented, and runs in live deployment against institutional traffic. The mathematical foundations of Section 4 do not depend on this specific implementation. They depend only on the architectural property that the trust boundary lives outside the model.

6 Regulatory Mapping

The architecture maps to existing institutional frameworks for model governance and AI risk management without modification or interpretation. We trace the mapping for five frameworks.

SR 26-2 (Federal Reserve, OCC, FDIC Revised Guidance on Model Risk Management, April 17, 2026). SR 26-2 supersedes SR 11-7 across all federally-supervised banking institutions. The interagency guidance retains the core principles of effective challenge, ongoing monitoring, outcomes analysis, and model lifecycle documentation that examiner review depends on. Footnote 3 of SR 26-2 explicitly excludes generative AI and agentic AI systems from the formal scope of the model risk management framework while affirming that institutions must still establish appropriate governance for these systems. The supersession is consequential for the present work in two ways. First, the principles SR 11-7 articulated and SR 26-2 retains, effective challenge documentation, model inventory, and outcomes analysis, continue to apply, and the grade lifecycle in the Trust layer is the documentary instantiation of effective challenge: every output is graded, every grade transition is signed, and the chain of grades is reviewable in retrospect under the bound of Theorem 3. Authorship attestation invalidations on unverified entries produce explicit security_posture penalties that surface in monitoring dashboards. The Providence audit chain is the documentation that survives examiner review. Second, the formal carve-out of generative and agentic AI from model risk management scope creates a governance gap that the present architecture is positioned to address: institutions must establish governance for these systems but have no prescriptive interagency framework to follow. The five-layer architecture provides a candidate framework with formal bounds.

NIST AI Risk Management Framework 1.0 and Generative AI Profile (NIST AI 600-1). NIST AI RMF 1.0 organizes risk management around four functions: govern, map, measure, manage. The five-layer architecture maps to all four. Identity addresses authority and accountability under govern. Ontology and Trust address map and measure, providing canonical reference and the quantitative trust scoring of Definition 5 and Theorem 2. Providence supports manage by producing the evidentiary record. Interdiction operates across all four functions through the five geometric primitives of Section 4.5. NIST AI 600-1, the Generative Artificial Intelligence Profile (July 2024), extends the base framework with twelve risk categories specific to generative AI systems, including hallucination, information integrity, value chain and component integration, and human-AI configuration. The SSRL drift detection (Definition 4) and the compositional EIA bound (Theorem 2) directly address the information integrity and value chain risks; the Interdiction layer addresses the hallucination and human-AI configuration risks through the geometric primitives.

EU AI Act, Article 12 (Logging Requirements for High-Risk AI Systems). Article 12 requires that high-risk AI systems automatically record events relevant to risk identification, system operation, and post-market monitoring. The records must be tamper-evident, timestamped, and verifiable by national competent authorities. The Providence layer instantiates these requirements with the bounds of Theorem 3, Proposition 3, and Equation 11. The information-theoretic security of QSIG provides verifiability over multi-decade horizons.

CISA Secure Software Development Attestation (SSDA). SSDA requires attestation of secure software development lifecycle practices, supply chain risk management, vulnerability

management, and access control. The Identity layer's WebAuthn-based attestation, the WASM-isolated key derivation on user devices (Property I.2), and the role-based access control with revocation (Property I.3) operate as instantiations of the access control requirements. The cryptographic supply chain for the engine, including the use of Azure Key Vault for key material and the verifiable signature properties of QSIG, addresses the supply chain risk management requirement.

Basel III Pillar 2 (Operational Resilience). Basel III Pillar 2 requires that internationally active banks demonstrate operational resilience capabilities. The compositional EIA bound (Theorem 2) is one such control: a runtime invariant that downstream trust cannot exceed compromised upstream trust, with signed evidence of every enforcement action. Grade revocation under administrative scope provides the ability to invalidate trust in production without taking the system down.

The mapping is one-way: the architecture provides primitives that institutional frameworks require. An institution adopting the architecture does not need to translate from one regulatory regime to another; the architecture provides the artifacts each regime asks for, with cryptographic verification that the artifacts are authentic.

7 Related Work

Mechanistic interpretability and alignment robustness. The structural insufficiency argument of Section 2 builds directly on Arditi et al. (2024), formalized here as Theorem 1. The result extends a body of work on linear representations of features in language model activation space (Bolukbasi et al., 2016; Mikolov et al., 2013; Park et al., 2023; Marks and Tegmark, 2023; Tigges et al., 2023) and on feature directions as causal mediators of behavior (Panickssery et al., 2023; Turner et al., 2023; Zou et al., 2023a; Templeton et al., 2024). Adversarial-suffix attacks, beginning with Zou et al. (2023b), are connected to the refusal-direction result through the attention-hijacking mechanism documented in Arditi et al. The fragility of safety fine-tuning to fine-tuning attacks is documented in Lermen et al. (2023), Yang et al. (2023), and Zhan et al. (2023).

Agent failure taxonomies and benchmarks. The argument of Section 3 builds on three contributions. Laban, Schnabel, and Neville (2026) introduce DELEGATE-52, the long-context delegated-edit benchmark grounding the formal corruption model of Section 3.1. Shapira et al. (2026) document eleven representative case studies of failures emerging from the integration of language models with autonomy, tool use, and multi-party communication. Franklin et al. (2026) at Google DeepMind develop the AI Agent Traps taxonomy of six attack categories targeting different components of an agent's operating cycle: perception, reasoning, memory, action, multi-agent coordination, and the human supervisor. The XSOC-NIE-GUARD paper (Blech, 2026d) maps ten controls across five planes against the Franklin taxonomy, anchored to a specific empirical case study (the OpenClaw security crisis of April 2026, with 138 documented CVEs and over 135,000 internet-facing instances). The present paper differs from XSOC-NIE-GUARD in that the present paper develops the architectural argument and its formal mathematical foundations, while XSOC-NIE-GUARD is a controls catalog grounded in the threat taxonomy.

Information geometry of neural networks. The geometric primitives of Section 4.5 build on a substantial literature in information geometry (Amari, 2016; Cover and Thomas, 2006). Fisher information geometry of neural networks has been studied for optimization (Martens, 2020) and for understanding generalization (Achille and Soatto, 2018). The specific application to adversarial detection in LLM inference paths is, to our knowledge, novel to the present work and the reference implementation it describes.

Cryptographic accountability. The cryptographic primitives we use have long histories. Hash-chained logs go back to Schneier and Kelsey (1999). Merkle-tree-anchored auditing is

treated in Crosby and Wallach (2009). Information-theoretic authentication is due to Wegman and Carter (1981); a textbook treatment is Stinson (2006). Our contribution is not in the cryptography but in its application to the trust boundary problem for AI inference, with formal bounds (Theorem 3, Equation 11) connecting the architecture to the institutional verification requirements of frameworks such as the EU AI Act. Adjacent constructions in the same primitive family include scoped artifact gate authorization (Blech, 2026c) and a zero-broadcast three-party transferable designated-verifier mode (Blech, 2026e); the latter is a multi-party extension of the QSIG primitive that supports cross-tenant verifiable transfer of authorization artifacts and is out of scope for the single-tenant trust architecture argument of this paper.

OWASP AI Exchange and adjacent industry frameworks. The OWASP AI Exchange catalog covers policy-level concerns adjacent to those addressed here. The architecture of Section 4 instantiates the audit and verification requirements described in the catalog with the formal bounds of Theorem 3 and Equation 11.

8 Discussion and Limitations

The architecture addresses a specific question: what should the trust boundary look like for institutional AI inference? It does not address all adjacent questions.

Model capability is out of scope. The architecture takes the model's outputs as input and applies external invariants. It does not improve the model's ability to perform the underlying task. A poor model produces poor outputs more often, which the architecture will downgrade or reject more often. The architecture is not a substitute for model selection, prompt engineering, retrieval augmentation, or fine-tuning for task performance.

Canonical reference is upstream. The compositional EIA bound (Theorem 2) protects against drift, contamination, and trust laundering relative to a canonical reference. It does not validate the canonical reference. The validation of $\mathcal{O}_\tau^{(v)}$ is an institutional process that occurs upstream of the inference architecture.

Independence assumption in the layer composition bound. Theorem 4 assumes that bypass attempts against different layers are statistically independent. Real attacks may exhibit correlations: an attacker who has compromised the identity layer may have correlated knowledge useful against the ontology layer. The architecture's value comes from the gap between the independence bound (Equation 27) and the maximum-correlation bound (Equation 28), and from the empirical observation that distinct layers attack distinct architectural concerns. A formal bound under structured correlation assumptions is open.

Performance trade-offs. The architecture introduces latency for every gate check. SSRL drift detection (Equation 6) runs in tens of microseconds. The geometric primitives (Equations 13, 16, 18, 20, 24) run in microseconds to milliseconds depending on the primitive. QSIG signing (Equation 11) is approximately 14 microseconds on a single core. The aggregate latency overhead in the reference implementation is small relative to LLM inference latency, which is typically dominated by the network round trip and the model's own computation. The architecture is not free, but it does not impose costs that change the order of magnitude of the underlying inference.

Adoption barriers. The architecture requires that the institution maintain a canonical ontology, a tenant policy, and a trust registry. For institutions that already operate with a canonical reference (banks under model risk management, hospitals under coding systems, law firms under defined-term taxonomies), the addition is incremental. For institutions that have deployed AI inference without a canonical reference, the addition is more significant and corresponds to building a process that arguably should exist regardless of whether AI inference is part of the operation.

Open mathematical questions. Several questions remain open. The compositional EIA bound (Theorem 2) is well-defined for linear chains of upstream sources but is more complex

for dense graphs of inference; we have practical solutions in the reference implementation but no general theorem on optimal bounds for arbitrary directed acyclic composition graphs. The geometric primitives are most informative for inference paths the architecture observes in full; for partially observed paths, the primitives provide weaker signal, and the optimal sampling strategy under partial observation is open. The information-theoretic security of QSIG holds against unbounded adversaries but does not protect against compromise of the institution's signing infrastructure; a formal treatment of key management under institutional adversary models is a separate problem.

9 Conclusion

The trust boundary for institutional AI inference must live outside the model. The argument is structural and is now mathematically established:

By Theorem 1 and Corollary 1, any trust property whose implementation is a low-rank subspace of the residual stream is recoverable through rank-one weight orthogonalization with bounded compute and bounded capability loss. In-model alignment is therefore not a trust boundary in the institutional sense.

By Proposition 2, no single-output classifier can detect the stateful semantic corruption that DELEGATE-52 quantifies empirically at 25 percent for frontier models and 50 percent on average. Output-level moderation is therefore not a trust boundary either.

By the architecture of Section 4, the trust boundary can be instantiated externally as a five-layer cryptographically anchored governance fabric. The compositional EIA bound (Theorem 2) closes the trust laundering attack. The Providence audit chain (Theorem 3, Proposition 3, Equation 11) provides tamper-evident, multi-decade-verifiable records under information-theoretic security. The five geometric primitives (Equations 13, 16, 18, 20, 24) detect adversarial patterns in the inference path before the surface output is admitted. The layer composition bound (Theorem 4) makes the architecture robust to advances in any one attack class.

The architecture is implementable. The reference implementation runs in live deployment against commercial LLM providers and tenant-isolated contexts, with QSIG signing at approximately 14 microseconds per sign, SSRL drift detection in microsecond latency, and the geometric primitives integrated into the gate decision at every inference.

The mapping to institutional frameworks (Section 6) is one-way: the architecture provides the artifacts each regime requires, with cryptographic verification that the artifacts are authentic. SR 26-2 effective challenge documentation is the grade lifecycle. NIST AI RMF risk management functions and AI 600-1 generative AI extensions are the five layers. EU AI Act Article 12 logging is the Providence chain under Theorem 3. CISA SSDA access control is the Identity layer under Properties I.1 through I.3. Basel III operational resilience is the compositional EIA bound under Theorem 2.

The implication for institutional adoption of AI inference is that the question of whether the deployed system is trustworthy is not a question about the model. It is a question about the architecture in which the model is embedded. A trustworthy institutional deployment does not depend on the model being aligned. It depends on the trust boundary being external by construction, with formal bounds on every consequential property.

Not by policy. By mathematics.

References

Achille, A., and Soatto, S. (2018). Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12).

Amari, S. (2016). *Information Geometry and Its Applications*. Applied Mathematical Sciences, vol. 194. Springer.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. (2024). Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems 38*. arXiv:2406.11717.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862.

Basel Committee on Banking Supervision. (2017). Basel III: Finalising post-crisis reforms. Bank for International Settlements.

Blech, R. (2026a). DSKAG-IT-SIG: Information-theoretic transaction signatures with hardware-bound policy binding and permissionless zero-knowledge on-chain verification. Zenodo. DOI: 10.5281/zenodo.19639166.

Blech, R. (2026b). XSOC FHE SDK Technical White Paper, Version 3.0.0.0. XSOC CORP. Distribution Statement A.

Blech, R. (2026c). XSOC-QSIG/CGA: Cryptographic gate authorization for scoped artifact signing and secure release. Zenodo. DOI: 10.5281/zenodo.19560239.

Blech, R. (2026d). XSOC-NIE-GUARD: A cryptographic mediation architecture for AI agent systems, with application to the OpenClaw security crisis. Zenodo. DOI: 10.5281/zenodo.19685360.

Blech, R. (2026e). XSOC-QSIG-3P: A zero-broadcast three-party transferable designated-verifier signature mode. Zenodo. DOI: 10.5281/zenodo.20078816.

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems 29*.

Cover, T., and Thomas, J. (2006). *Elements of Information Theory*, 2nd ed. Wiley-Interscience.

Crosby, S., and Wallach, D. (2009). Efficient data structures for tamper-evident logging. In *Proceedings of the 18th USENIX Security Symposium*.

Cybersecurity and Infrastructure Security Agency (CISA). (2024). Secure software development attestation form. United States Department of Homeland Security.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act). *Official Journal of the European Union*.

Federal Reserve Board, Office of the Comptroller of the Currency, Federal Deposit Insurance Corporation. (2026). SR 26-2: Revised Guidance on Model Risk Management. Board of Governors of the Federal Reserve System, April 17, 2026.

Franklin, M., Tomašev, N., Jacobs, J., Leibo, J. Z., and Osindero, S. (2026). AI Agent Traps. Google DeepMind. SSRN Working Paper 6372438. DOI: 10.2139/ssrn.6372438.

Laban, P., Schnabel, T., and Neville, J. (2026). LLMs corrupt your documents when you delegate. Microsoft Research. arXiv:2604.15597.

Lermen, S., Rogers-Smith, C., and Ladish, J. (2023). LoRA fine-tuning efficiently undoes safety training in Llama 2-Chat 70B. arXiv:2310.20624.

Marks, S., and Tegmark, M. (2023). The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. arXiv:2310.06824.

Martens, J. (2020). New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146).

Mikolov, T., Yih, W., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of NAACL HLT 2013*.

National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0). United States Department of Commerce.

National Institute of Standards and Technology. (2024). NIST AI 600-1: Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. United States Department of Commerce.

Open Worldwide Application Security Project (OWASP). (2025). AI Exchange: Controls catalog. OWASP Foundation.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. (2023). Steering Llama 2 via contrastive activation addition. arXiv:2312.06681.

Park, K., Choe, Y., and Veitch, V. (2023). The linear representation hypothesis and the geometry of large language models. arXiv:2311.03658.

Schneier, B., and Kelsey, J. (1999). Secure audit logs to support computer forensics. *ACM Transactions on Information and System Security*, 2(2).

Shapira, N., Wendler, C., Yen, A., et al. (2026). Agents of Chaos. arXiv:2602.20021.

Stinson, D. (2006). *Cryptography: Theory and Practice*, 3rd ed. Chapman and Hall/CRC.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., et al. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread.

Tigges, C., Hollinsworth, O., Geiger, A., and Nanda, N. (2023). Linear representations of sentiment in large language models. arXiv:2310.15154.

Turner, A., Thiergart, L., Udell, D., Leech, G., Mini, U., and MacDiarmid, M. (2023). Activation addition: Steering language models without optimization. arXiv:2308.10248.

Wegman, M., and Carter, J. (1981). New hash functions and their use in authentication and set equality. *Journal of Computer and System Sciences*, 22(3).

Yang, X., Wang, X., Zhang, Q., Petzold, L., Wang, W., Zhao, X., and Lin, D. (2023). Shadow alignment: The ease of subverting safely-aligned language models. arXiv:2310.02949.

Zhan, Q., Fang, R., Bindu, R., Gupta, A., Hashimoto, T., and Kang, D. (2023). Removing RLHF protections in GPT-4 via fine-tuning. arXiv:2311.05553.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A., et al. (2023a). Representation engineering: A top-down approach to AI transparency. arXiv:2310.01405.

Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. (2023b). Universal and transferable adversarial attacks on aligned language models. arXiv:2307.15043.

Author Information

Richard Blech is Chief Executive Officer and Technical Lead of XSOC CORP, a non-traditional defense contractor headquartered in Irvine, California. He is the author of *The Cognitive War* and an active contributor to the OWASP AI Exchange. His prior work includes the OpenClaw / XSOC-NIE-Guard control framework and the XSOC-QSIG post-quantum signature primitive. ORCID: 0009-0003-4540-2134.

Acknowledgments

The author thanks the OWASP AI Exchange community for ongoing discussion of the controls landscape, and the reviewers who engaged with earlier iterations of the architectural argument across multiple academic and industry venues. Particular thanks to Bill Stout for flagging the SR 11-7 to SR 26-2 supersession in the NIST overlay working group channel, which prompted the Version 2.0 revision.

Conflicts of Interest

The author is the founder and majority shareholder of XSOC CORP, which develops the reference implementation discussed in this paper. The architectural argument and the formal results of Theorems 1 through 4 are independent of this commercial interest. The reference implementation is described to demonstrate that the architecture is implementable; the paper's argument does not depend on this specific implementation.

Distribution