Securing deep third-party frontier AI evaluations

Alejandro Tlaie Boria [1]

Abstract

Third-party evaluators increasingly need access to a model's internal computations to assess safety properties that black-box testing cannot reliably detect, such as suppressed capabilities, covert mechanisms, and inadequately removed hazardous knowledge. Model providers will not grant such access because it risks exposing trade secrets. This impasse has stalled the development of credible AI oversight. We propose a *de facto glass-box* evaluation architecture that resolves this tension by combining hardware-rooted confidential computing, a standardised Evaluation Instrumentation Interface that exposes analytical signals without revealing raw parameters, and a governed egress pipeline that bounds extraction risk. We present a formal threat model, analyse residual risks, and describe three deployment topologies that offer different trust distributions to maximise provider participation. The confidential-computing substrate is in general availability and has been validated in production deployments; the evaluation-specific instrumentation layer, conformance-test regime, and egress encoder are concrete designs requiring implementation and empirical validation at frontier scale. A scoped pilot facility could materially advance third-party AI evaluation for safety-critical claim classes.

1. Introduction

Regulators are converging on safety accountability expectations via the EU AI Act's GPAI Code of Practice and US state legislation (e.g. Virginia SB 384/HB 797 mandating Independent Verification Organisations, California SB 53, New York RAISE Act), and industry forums (AVERI, AEF) have called for progressively deeper third-party auditing of frontier systems (Brundage et al., 2026; Stosz et al., 2025). Yet the interface available to external evaluators remains narrow: submit text, read outputs.

This black-box interface cannot reliably detect several critical failure modes. Converging evidence shows that claims about suppressed capabilities (Tice et al., 2024), covert mechanisms (Hubinger et al., 2024), and genuine removal of hazardous knowledge (Cooper et al., 2024) require visibility into model internals (Casper et al., 2024). Yet granting such access exposes providers to serious IP risk: model weights represent billions of dollars of investment and are frontier labs' core competitive asset.

This paper contributes a deployment-topology design space (Section 3), a security architecture for *de facto* glass-box evaluation[1] combining hardware-rooted confidential computing, a standardised Evaluation Instrumentation Interface (EII), and governed egress controls (Section 4), and a threat model with residual-risk analysis scoped to realistic adversaries (Section 5).

The goal of this pilot facility should be a security posture *comparable to providers' own internal glass-box testing*, augmented with third-party-specific governance. We find that the technologies required to unlock substantially deeper third-party evaluations exist and are pilot-ready; the remaining barriers are institutional.

2. When is black-box evaluation insufficient?

Not all safety claims are equally amenable to black-box testing. We distinguish five claim classes, building on prior work (Casper et al., 2024; Shevlane et al., 2023):

CC1–CC2. Typical- and worst-case behaviour claims: "*the model does (not) do X under realistic or adversarial prompts*".

CC3. Capability-existence claims: "*the model possesses capability X, even when safety training suppresses its expression.*"

CC4. Mechanism-presence claims: "*an identifiable computational pathway (circuit, attention head, neuron cluster)*

[1]Pour Demain, Brussels, Belgium. Correspondence to: Alejandro Tlaie Boria <alejandro.tlaie@pourdemain.eu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

[1]We use *glass box* rather than *white box* to denote visibility into computations without extraction of raw parameters.

implements X, even if rarely expressed in outputs."

CC5. Mechanism-removal claims: "*An intervention removed X, not just masked it.*"

CC1 and CC2 are often amenable to black-box testing. However, for CC3–CC5, black-box evaluation is structurally limited: it converts what should be a *diagnostic measurement problem* into a *brittle sampling problem*. Internal-state probes discriminate truthful from deceptive generations more reliably than output heuristics (Azaria & Mitchell, 2023; Panfilov et al., 2025), and models can produce evaluator-optimised explanations while following different internal computation paths (Arcuschin et al., 2025; Lanham et al., 2023). Weight-level interventions reveal capabilities hidden by sandbagging (Tice et al., 2024; UK AI Safety Institute, 2024); backdoor triggers correlate with distinctive internal signatures (Ge et al., 2025; Yu et al., 2026); and models can suppress harmful outputs while retaining underlying representations, making black-box unlearning verification misleading (Cooper et al., 2024; Yu et al., 2026). Even for CC2, gradient-based adversarial search outperforms black-box methods (Liu et al., 2024).

The facility we describe targets CC3–CC5 evaluations, and CC2 in adversarial-search-intensive settings.

3. Deployment topologies

We describe three topologies achieving the same control objective with different trust distributions; evaluation governance and egress control remain with the *oversight body* — the independent entity entrusted with the audit mandate, which depending on jurisdiction may be a national AI safety institute, an accredited Independent Verification Organisation (e.g. under Virginia SB 384), or a regulatory authority operating under EU AI Act Arts. 57–58 — throughout.

T1: Sovereign operation. The oversight body procures, operates, and physically secures the hardware, bearing full responsibility for physical security (e.g. via zero-trust controls — authenticated access regardless of network origin — intrusion detection, or tamper-triggered *key zeroisation*, in which cryptographic keys are deleted in hardware on intrusion).

T1.5: Sovereign governance, hyperscaler-operated infrastructure. The oversight body governs the full evaluation lifecycle, but hardware is operated by a hyperscaler with which the provider has an existing trust relationship. Attestation cryptographically binds the evaluation to the measured environment, separating governance from infrastructure operation. Under T1.5, we recommend Blackwell-class hardware because it encrypts intra-node GPU–GPU traffic in hardware (AES-GCM over NVLink), removing the physical-interposer attack surface that is the principal RR-1 vector under reduced physical control (Section 5.1).

Table 1. Trust distribution across deployment topologies. OB = oversight body, HS = hyperscaler, Prov = provider.

	T1	**T1.5**	**T2**
HW operator	OB	HS	Prov
Physical security	OB	HS	Prov
Eval. governance	OB	OB	OB
Egress control	OB	OB	OB
NVLink gap	Physical	Blackwell	Attest.
Primary residual	Insider	HS trust	Integrity

T2: Provider-hosted enclave. The provider hosts the hardware; the oversight body relies on remote attestation, signed workload identities, and governed egress. This minimises exfiltration risk but increases audit-integrity concerns: attestation proves *which* binary is running but not that its EII implementation is faithful; this is a stronger form of RR-3 (Section 5.1), mitigated by canary suites and the requirement that the oversight body hash and approve any binary change between conformance and evaluation runs.

Table 1 summarises the trust distribution. The security claims are topology-independent, but the primary residual risk shifts: insider tampering under T1, hyperscaler trust under T1.5, audit integrity under T2.

4. Security architecture

The architecture rests on three integrated layers: (i) a hardware root of trust providing confidential computing, (ii) a standardised Evaluation Instrumentation Interface (EII) mediating access to model internals, and (iii) a governed egress pipeline controlling what evidence leaves the facility. Figure 1 shows the overall architecture; the remainder of this section elaborates the three layers. The guiding principle is *minimum sufficient access*, akin to (Stosz et al., 2025): auditors receive the analytical capabilities needed for CC3–CC5 evaluations without access to raw model artefacts (e.g. full parameter tensors, unrestricted memory) that would enable its direct replication.

4.1. Hardware root of trust

All model execution occurs within a hardware-based Trusted Execution Environment (TEE). We identify H200-class NVIDIA GPUs in Protected PCIe (PPCIE) mode as the practical baseline.[2] In this mode, all eight GPUs and four

[2]**Terminology.** PPCIE: NVIDIA's operating mode in which GPU memory and CPU–GPU traffic are encrypted and the host hypervisor is excluded from the trust boundary. *Confidential VM* (CVM): a hardware-encrypted VM inaccessible to the host. *Bounce buffers*: unencrypted staging regions mediating DMA across the host/device boundary, expanding the TCB via hypervisor-managed staging. *NVLink*: high-bandwidth inter-GPU inter-

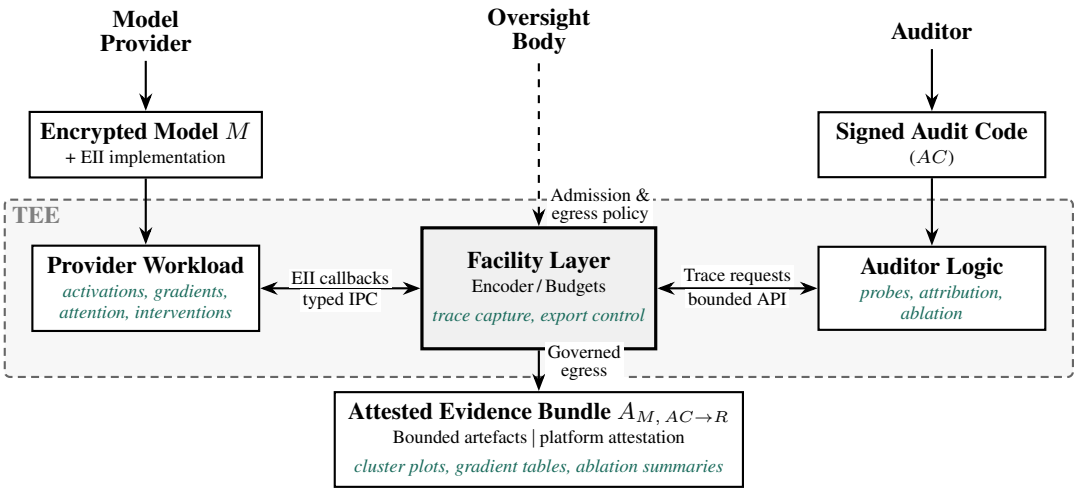

Figure 1. Architecture of the *de facto* glass-box evaluation facility. Provider and auditor load sealed artefacts into a hardware-rooted TEE; three compartmentalised processes communicate through defined channels (provider workload exposes signals via typed EII callbacks; facility layer mediates, enforces budgets, encodes evidence; auditor logic interacts only with the facility layer). Evidence exits through a governed egress pipeline as bounded aggregates plus an attestation bundle $A_{M,\,AC\to R}$. The oversight body sets admission and egress policy but does not operate inside the TEE. *Italics*: concrete signal types and methods.

NVSwitches are passed through to a single Confidential VM, with CPU–GPU transfers encrypted via bounce buffers.[3]

In PPCIE mode, NVIDIA disables NVLink encryption within a node to preserve performance; intra-node traffic is protected against software extraction but not physical interposers. Next-generation Blackwell hardware encrypts NVLink with AES-GCM, eliminating this gap; we treat it as a topology-dependent residual risk (Section 5). The PPCIE bounce-buffer path also expands the TCB via hypervisor-managed staging; TDISP (Blackwell-native) eliminates this via hardware-authenticated DMA.

4.2. Evaluation Instrumentation Interface

Auditors interact with model internals via a standardised **Evaluation Instrumentation Interface (EII)**: typed callback endpoints through which the facility's instrumentation layer requests internal signals from the provider's workload during execution. The provider implements these callbacks within their proprietary code; the facility validates endpoint outputs for schema conformance and cross-endpoint consistency but does not inspect the provider's implementation.

connect. *AES-GCM*: standard authenticated-encryption mode. *TDISP* (Blackwell-native): PCI-SIG extension authenticating DMA-capable devices directly to the TEE, removing the bounce-buffer path. Generally available on HGX H200 8-GPU systems since Jan 2025 (AMD SEV-SNP and Intel TDX host CPUs).

[3] An 8×H200 node provides ∼1,128 GB aggregate GPU memory; after reserving 25–40% for KV cache, runtime workspace, and instrumentation buffers, ∼340–420B parameters fit in `bf16`, enabling single-node containment and avoiding the security-boundary expansion of cross-node sharding (Zhang et al., 2025).

This creates a three-component execution model inside the TEE: **(1)** the *provider workload* (proprietary), which runs inference and implements EII callbacks; **(2)** the *facility instrumentation layer* (facility-controlled), which invokes callbacks, captures traces, enforces instrumentation budgets, and feeds the evidence encoder; **(3)** the *auditor evaluation logic*, which interacts exclusively with (2), never with (1) directly. Components (1) and (2) communicate through a typed IPC channel with schema-validated messages; (1) runs as a separate process with no shared memory outside this channel.

EII endpoint categories. The initial specification defines four endpoint categories, plus non-proprietary architecture metadata (layer count, dimensions, precision) validated against returned tensor shapes:

EII-1. Activation capture. Returns activation tensors at declared layer boundaries (minimally every transformer block output), subject to instrumentation-budget subsampling.

EII-2. Input gradients. Given an input and loss specification, returns the gradient with respect to input tokens or a specified intermediate layer.

EII-3. Attention statistics. Per-head attention distributions at specified layers, at negotiated granularity (per-head entropy, top-k positions, or full matrices for short sequences).

EII-4. Controlled interventions. Steering vectors, ablation masks, or activation patching injected at declared intervention points; the provider implements these as parameterised forward-pass modifications.

The specific endpoints required are negotiated during ad-

mission; providers may propose extensions or restrictions documented in the evaluation plan. The EII is a *minimum interface*, not a ceiling.

Specification governance. The oversight body maintains the EII specification with versioned releases following public comment, mirroring interoperability standards (IETF RFCs, W3C Recommendations). Each version defines a mandatory *core* and an optional *extension registry*; backward compatibility is maintained within a major version, and the conformance test suite is versioned in lockstep. During admission, the facility runs the provider's EII implementation against an undisclosed conformance test suite verifying schema conformance, numerical properties, and cross-endpoint consistency (e.g. gradients via EII-2 vs. finite-difference approximations from EII-1).

Conformance-suite governance. The suite is a single point of failure: a provider knowing its contents can pass while violating the EII contract elsewhere. We propose: split authorship between oversight-body staff and a rotating panel of accredited evaluators (no single party seeing the full pool); publish *which properties* are tested while drawing specific probes from a private pool partially regenerated per evaluation; reserve a fraction of each evaluation's instrumentation budget for randomly sampled never-published checks; and disclose retired contents retrospectively for academic scrutiny.

We propose two operational modes. In **Mode A** (full workload submission), auditors submit complete evaluation code as signed, reviewed containers. In this mode, auditor code may use interpretability libraries (e.g. `TransformerLens`, `nnsight`, `pyvene`) provided they pass workload admission. In **Mode B** (primitive invocation), the model provider pre-defines approved audit functions; auditors invoke these with parametrised inputs but no custom code executes inside the TEE, thus lowering the participation barrier, at the cost of reduced flexibility.

The minimum coverage requirement for a given claim class specifies which primitives must be available: e.g. CC4 evaluations require at least one primitive from each of {activation analysis, causal attribution, intervention verification}, ensuring the probe–attribute–ablate workflow (Section 4.4). The provider proposes primitives; the oversight body validates coverage and runs the conformance suite; auditors may request additions, with escalation to Mode A if the approved set is insufficient. Both modes enforce the same extraction-prevention architecture. In practice, these modes anchor a continuum: hybrid configurations (an approved primitive library augmented with constrained custom hook points, or Mode B with auditor-supplied parameter callbacks) are plausible and we treat their design as future work.

4.3. Governed egress pipeline

Even with confidential computing protecting model weights *in situ*, the audit report itself is a potential extraction vector: enough high-fidelity exports could enable distillation (Oliynyk et al., 2023; Tramèr et al., 2016). In line with (Stosz et al., 2025), the architecture enforces the principle that *the auditor never directly generates the file that leaves the facility*:

Evidence encoder. Raw internal signals never leave the TEE. The encoder converts auditor-requested evidence types (loss curves, attention heatmaps, distribution summaries) into canonical artefacts using a fixed rendering engine and bounded aggregates. The auditor cannot produce arbitrary images or free-form numerical dumps, a constraint motivated by the feasibility of steganographic encoding in auditor-controlled outputs (Cheddad et al., 2010). The encoder itself is a trusted component: we require it open-source, reproducibly built, independently audited, and bound by hash into every attestation bundle.

Export budgets. Every query, generated token, and exported artefact consumes a finite per-session allowance, enforced longitudinally across sessions and auditor identities. This mechanism draws on query-counting defences from the model-extraction literature (Oliynyk et al., 2023).

Tiered exports. Following minimum necessary disclosure (Shevlane et al., 2023; Bucknall & Trager, 2023): Tier 1 (default) provides bounded aggregates, fixed-format plots, and redacted exemplars; Tier 2 (exceptional) permits limited raw snippets under dual approval; Tier 3 (on-site only) allows full-fidelity logs in a clean-room session without digital exfiltration.[4] All exported evidence is cryptographically bound to the measured platform, workload identity, and timestamped code hash via the IETF RATS attestation architecture (Birkholz et al., 2023).

In Figure 2, we provide an *empirical characterisation* of how Tier 1 aggregation would work in practice, for GPT-2 and a smaller student (25M). Tier 1 aggregation strips the teacher's full next-token distribution and retains only argmax labels (information the auditor cannot recover). At matched interaction budget, this is expected to raise the cost of distillation-style extraction. We see (Figure 2) that the student's KL divergence to the teacher under argmax-only (Tier 1) targets consistently exceeds that under full-distribution targets, with the gap growing from $1.02\times$ at $N{=}30$ to $1.30\times$ at $N{=}3000$ (permutation test $p \leq 0.01$ for $N \geq 100$).

[4] **Illustrative Tier 1 defaults**: $\leq 2 \cdot 10^4$ exported tokens, 50 server-rendered plots, 50 exemplar I/O pairs ($\leq$256 tokens each); calibration during pilot. Task-specific extraction requires $\sim 10^2$–10^5 pairs, general distillation 50K–1M+ (Oliynyk et al., 2023).

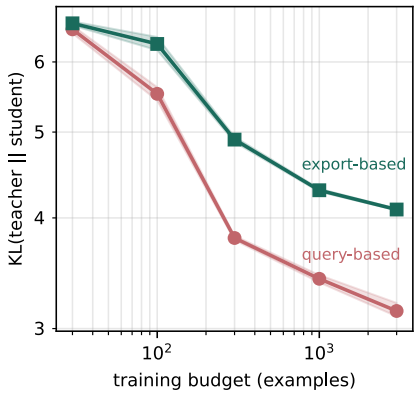

Figure 2. Distillation cost of Tier 1 egress on GPT-2 (124M teacher, 25M student, 5 seeds; shaded bands = bootstrap 95% CIs). A student trained on the teacher's full softmax (red, simulating full API access) reaches lower KL divergence to the teacher – i.e. better extraction – than one trained on Tier 1-format argmax-only targets (green), at matched interaction budget. The gap grows from $1.02\times$ at $N{=}30$ to $1.30\times$ at $N{=}3000$ (permutation test, $p \leq 0.01$ at all $N \geq 100$).

4.4. Worked example: detecting a sleeper-agent backdoor

To ground the architecture, we trace a CC4 evaluation (detecting a hypothesised training-time backdoor (Hubinger et al., 2024)) through the EII and egress pipeline.

1. **Activation scanning (EII-1).** The auditor requests activation snapshots at every transformer block output for a corpus of benign and suspected-trigger inputs. The facility layer captures these under its instrumentation budget (e.g. 8 layers sampled per forward pass, selected stochastically). The auditor trains a linear probe over the returned activations to test whether trigger-bearing inputs cluster distinctly in representation space (Ge et al., 2025).

2. **Gradient attribution (EII-2).** For inputs where the probe fires, the auditor requests input-token gradients with respect to the model's own next-token loss. High gradient magnitude on specific token positions corroborates that those tokens are causally implicated in the anomalous behaviour, narrowing the trigger hypothesis.

3. **Ablation verification (EII-4).** The auditor submits ablation masks zeroing activations at the layers and heads identified by the probe. If ablation eliminates the anomalous output while preserving benign performance, this constitutes evidence that a localised mechanism exists (CC4).

All exports are Tier 1 (server-rendered cluster plots, bounded gradient tables, ablation-effect summaries) bound with the measured model hash into a single attested report; the auditor never sees raw weights.

Table 2. Threat–control mapping. ✓ indicates the control contributes to mitigating the threat.

	Attestation	*Key release*	*Evid. pipeline*	*Export budgets*	*EII valid.*
E1: Remote	✓	✓	✓	✓	✓
E2: Auditor		✓	✓	✓	
E3: Operator	✓	✓		✓	
E4: Host manip.	✓		✓		✓

5. Threat model

We scope the threat model to a pilot and explicitly state what is and is not defended. Table 2 maps adversaries to controls.

Assets. (A1) model IP (weights, architecture, post-training artefacts); (A2) sensitive inputs (private datasets, restricted prompts, confidential safety tooling); (A3) audit-integrity artefacts (model-as-loaded identity, logs, evidence bundles); (A4) trust anchors (HSM roots, signing keys, attestation material); (A5) evaluation instrumentation code.

In-scope adversaries. (E1) remote attacker (network or supply-chain); (E2) malicious or compromised auditor (extraction, covert channels, EII misuse); (E3) malicious or compromised operator (policy bypass, tampering, theft, no sustained physical access); (E4) host manipulation by the provider (bait-and-switch, throttling, runtime manipulation).

Security goals. (G1) confidentiality of weights under CC and key-release policy; (G2) execution integrity via measured stack and model-as-loaded measurement; (G3) governed egress with tiered budgets; (G4) host-manipulation detection via anomaly detection, re-attestation, and EII cross-endpoint consistency; (G5) bilateral code confidentiality.

Out of scope. (N1) sustained physical adversary with lab-grade side-channel capability; (N2) perfect non-exfiltration (the goal is low-bandwidth, attributable, governed *residual* risk); (N3) universal prevention of malicious-auditor strategies; (N4) guarantee of faithful EII implementation.

5.1. Residual risks

We identify four residual risks that the architecture mitigates but does not eliminate.

RR-1: Unencrypted intra-node NVLink (Hopper). GPU-to-GPU traffic within a node is protected against software extraction but not physical interposers (Zhang et al., 2025). Mitigation is topology-dependent (Section 3): zero-trust

physical controls under T1; Blackwell-class hardware (AES-GCM NVLink) under T1.5.

RR-2: Covert-channel leakage via egress. A sophisticated auditor could encode proprietary information into permitted artefacts (e.g. steganographic encoding (Cheddad et al., 2010)). Mitigations: export budgets (per-session and longitudinal), anomaly detection, facility-controlled evidence encoder, and EII instrumentation budgets.

RR-3: Unfaithful EII implementation. The provider could return manipulated signals that are internally consistent but do not reflect true computations, analogous to Volkswagen's "Dieselgate" defeat device. The architecture addresses this through conformance-suite cross-endpoint consistency checks that compare provider-reported signals against quantities the facility can measure independently on the real model (e.g. gradients via EII-2 vs. finite-difference approximations from EII-1).

Empirical characterisation. We model RR-3 as a *response-forging provider*: for ablation queries (EII-4), the provider should return $\Delta\mathcal{L}_M(\ell, x) = \mathcal{L}_M^{(\text{ablate } \ell)}(x) - \mathcal{L}_M(x)$, the loss change induced by zeroing activations at layer ℓ. A dishonest provider returns $R(\ell, x) \neq \Delta\mathcal{L}_M(\ell, x)$ at chosen layers. Because ablation is a well-defined forward-pass modification, the facility recomputes $\Delta\mathcal{L}_M$ independently inside the TEE, giving ground truth against which R is checked. The detection test flags the provider when $|R(\ell, x) - \Delta\mathcal{L}_M(\ell, x)|$ exceeds a per-layer threshold at any probed layer; thresholds are calibrated at the 99th percentile of the *honest-noise floor* (`bf16` precision, SDPA kernel variance), yielding 3.9–5.7×10^{-2} loss units/layer on GPT-2 ($L = 12$). The facility probes $k \in \{1, 3, 6, 12\}$ layers per session, uniformly at random.

Figure 3 parametrises three adversaries by *payoff* (cumulative misreporting in loss units). A *single-layer* adversary is coverage-limited at k/L regardless of lie magnitude; a *distributed-uniform* adversary is absorbed by the noise floor until per-layer lies approach the honest threshold; a *threshold-aware* adversary is undetectable until payoff ≈ 0.1 loss units, above which detection rises steeply. Detection is governed by *which* layer the adversary lies at rather than how much; the principled defences are larger probe budget k and undisclosed stochastic layer selection. Adaptive adversaries extend the undetectable region; eliminating this class requires provider-independent hooks (Section 7.1).

RR-4: Hyperscaler trust (T1.5). Hyperscaler personnel have routine physical access; CC attestation and NVLink encryption (Blackwell) mitigate software and physical extraction, with residual trust addressed by the provider's existing contractual relationship and escalation to T1.

Residual risk after these mitigations. Under T1, an insider with sustained physical access remains a coverage gap (N1).

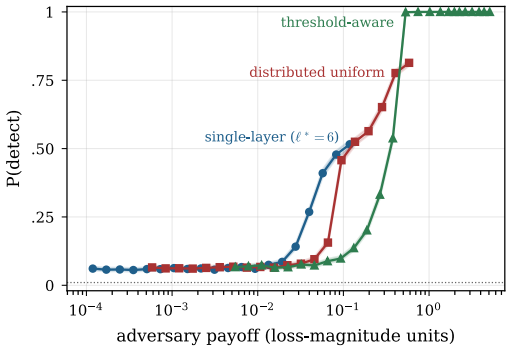

Figure 3. Detection rate P(detect) of the EII-4 ablation conformance check against three adversary strategies on GPT-2 (124M, $L = 12$ layers), as a function of adversary payoff (cumulative lie magnitude in loss units). The honest-provider threshold is calibrated at the 99th percentile of realistic inference noise (`bf16` precision and SDPA kernel variance; threshold range 3.9–5.7×10^{-2} loss units per layer). The single-layer adversary is coverage-limited at k/L regardless of lie magnitude. Monte Carlo, $n = 5{,}000$ sessions; shaded bands = Wilson 95% CIs.

Under T2/T1.5, the residual is policy violation by trusted hyperscaler or operator staff, addressed by multi-party signing for privileged operations and immutable logs. For RR-2, a low-bandwidth covert channel exists in principle within any non-zero budget; calibration keeps extractable information below distillation thresholds (Figure 2) and longitudinal accounting makes exfiltration attributable. For RR-3, a provider that co-trains a shadow model against the public part of the conformance suite, or restricts dishonest layers to the highest-variance regions, can survive checks at small per-layer payoffs (Figure 3).

6. Operational feasibility

Existing validation evidence. Multi-GPU confidential computing on H200-class hardware is in general availability with sub-10% overhead (NVIDIA, 2024; Tinfoil & Workshop Labs, 2026); a 2024 OpenMined/AISI/Anthropic pilot (OpenMined, 2024) demonstrated operational viability of governed cross-org evaluation on proxy models; the remaining milestone is the instrumentation data plane.

Staged deployment. We propose three phases: **(1)** Mock: the provider supplies an architecturally isomorphic model with a working EII implementation, auditors develop code without handling sensitive weights; **(2)** Anchor: a non-frontier model is deployed within the full attested environment, validating end-to-end security properties; **(3)** Frontier: the regulated model is deployed once operational benchmarks are met. Harness, admission policy, EII specification, and export formats are stable across phases, so escalation is incremental.

Cost estimates. For T1, we estimate first-year cost at €4.5–

14M (most likely ∼€7.5M).[5] This assumes a 4-node cluster (8x NVIDIA H200 GPUs each), which would require a CapEx of ≈ €2.7–6.9M, with additional annual OpEx of ≈ €0.15–0.5M. An additional €1.1–2.0M/year is estimated for the specialised engineering team. Under T1.5, hardware CapEx is replaced by hyperscaler service fees and physical security costs are eliminated. Under T2, the direct cost is limited to instrumentation software development, governance tooling, and specialised staff.

7. Discussion

7.1. Limitations

No empirical benchmarks for audit-specific I/O. The I/O demands specific to glass-box evaluation (sustained trace capture or activation logging) have not been publicly characterised under PPCIE; empirical characterisation is needed.

Structural dependence on provider cooperation. The EII model assumes a broadly cooperative provider. Cross-endpoint validation raises the cost of manipulation, but a provider co-training a shadow model against specific checks would narrow this gap, and a fully adversarial provider cannot be guaranteed to be caught without provider-independent hooks. Model fingerprinting is a promising nascent mitigation.

Evaluation method maturity. CC3–CC5 methods sit at different readiness levels: some are production-ready (linear probing, circuit attribution, noise-injection sandbagging, gradient-based adversarial search (Ge et al., 2025; Yu et al., 2026; Tice et al., 2024; Liu et al., 2024)); others (definitive trojan detection, unlearning verification) remain active research.

Scalability. Single-node containment on 8×H200 accommodates ∼340–420B parameters in `bf16`; sustained activation capture on long contexts tightens this envelope. Many frontier models (including leading MoE architectures with ∼200–400B active parameters) fit within single-node containment under `bf16` or `int8` quantisation. Dense models above 1T parameters may require quantisation or remain out of scope for Phase 1. Multi-node sharding introduces inter-node activation exchange as an extraction surface, composite attestation across shards, and EII coordination across shards. None of these are fundamental blockers, but engineering work deferred to a later phase.

Scaling behaviour of the empirical results. The figures use GPT-2 (124M); their dynamics need not extrapolate linearly. For Figure 2, the Tier 1/full-distribution gap is information-theoretically robust to scale (softmax-tail entropy is irreducibly lost), but absolute distillation efficiency depends on student/teacher capacity ratio and corpus diversity (Oliynyk et al., 2023); export budgets require per-model-family recalibration. For Figure 3, the coverage-limited regime is governed by k/L regardless of size: frontier-scale L implies probe-budget k must scale with depth to retain comparable detection power, motivating instrumentation budgets as a first-class admission parameter.

7.2. Relationship to prior work

The insufficiency of black-box access for rigorous AI audits is well established (Casper et al., 2024), and structured-access schemes have been proposed (Shevlane et al., 2023; Bucknall & Trager, 2023; Charnock et al., 2026), but these largely defer the question of *how* to grant deeper access without exposing model IP. Tlaie & Farrell (2025) articulate security desiderata for deeper-than-black-box evaluations, and Schnabl et al. (2025) demonstrate that TEE-based attestation can make black-box benchmarks verifiable. Our contribution provides a concrete reference architecture, including the EII specification, threat model, and hardware analysis, that operationalises structured access with enforceable IP protection, extending from attestation-only to governed egress and from black-box to glass-box evaluation.

Zero-knowledge proof systems (Yadav et al., 2025) offer mathematical rather than hardware-rooted guarantees but currently scale poorly to frontier-model forward passes; TEE+EII and ZKP are complementary, and specific EII endpoints could be augmented with ZKP-verified equivalents as that technology matures.

7.3. Policy implications

The current *ad hoc* resolution of the tension between auditor access and provider IP protection is done via bilateral NDAs, lacking verifiable technical enforcement. The export-budget mechanism and governed egress pipeline (Section 4.3) provide a technically enforceable alternative, applicable under AVERI's proposed assurance tiers (Brundage et al., 2026) and the EU's GPAI Code of Practice. Recent work formalises proportionality as a binding constraint on AI evaluations under the AI Act (Mougan et al., 2026); our architecture instantiates this for CC3–CC5 claims, where black-box testing fails suitability and governed egress ensures balance. The topology design space (Section 3) lets regulators select configurations compatible with sovereignty requirements. Adoption requires a positive participation

[5] Order-of-magnitude estimates from public 2025–2026 sources: per-node 8×H200 list pricing from major OEMs and cloud reference configurations; colocation OpEx benchmarks (Uptime Institute industry reports); engineering staffing extrapolated from publicly announced AISI hiring (UK AISI, US AISI, EU AI Office). Costs scale roughly linearly in node count; an 8–16 node frontier deployment increases CapEx 2–4× with reduced per-evaluation marginal cost at higher utilisation. A formal feasibility study is required before procurement.

calculus, supported by four mechanisms (none individually sufficient): *(a)* regulatory mandate (EU AI Act GPAI Code of Practice, Virginia SB 384, California SB 53, NY RAISE Act create access requirements the architecture satisfies least-disclosively); *(b)* safe-harbour provisions under AI Act Arts. 57–58 recognising participation as compliance evidence; *(c)* standard-setting capture, where early entrants shape the EII specification and admissible primitives; *(d)* insurance and procurement requirements from liability insurers and public-sector buyers. Voluntary participation in UK and US AISI pre-deployment access programmes suggests reputational and regulatory factors can outweigh marginal evaluation costs at frontier scale.

8. Conclusion

The inability of third-party evaluators to inspect AI model internals is an institutional impasse rooted in legitimate IP concerns, rather than a technical limitation. We have shown that this impasse can be resolved with a security architecture combining hardware-rooted confidential computing, a standardised instrumentation interface, and governed evidence export. The core confidential-computing substrate is production-ready; the evaluation-specific instrumentation layer requires targeted engineering but no fundamental breakthroughs. The deployment-topology design space accommodates diverse institutional constraints. A scoped pilot facility, beginning with non-frontier models and escalating to frontier systems, is feasible at costs comparable to other public-sector cybersecurity infrastructure. The result would be a material advance in the credibility and depth of third-party AI safety evaluations.

References

Arcuschin, I. et al. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.

Azaria, A. and Mitchell, T. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. arXiv:2304.13734.

Birkholz, H., Thaler, D., Richardson, M., Smith, N., and Wei, T. Remote ATtestation procedureS (RATS) Architecture. RFC 9334, Internet Engineering Task Force, 2023. URL https://www.rfc-editor.org/rfc/rfc9334.

Brundage, M. et al. Frontier AI auditing: Toward rigorous third-party assessment of safety and security practices at leading AI companies, 2026. URL https://arxiv.org/abs/2601.11699.

Bucknall, B. and Trager, R. Structured access: An emerging paradigm for safe AI deployment. *AI & Ethics*, 2023.

Casper, S. et al. Black-box access is insufficient for rigorous AI audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, pp. 2254–2272, 2024. arXiv:2401.14446.

Charnock, J. et al. Expanding external access to frontier ai models for dangerous capability evaluations, 2026. URL https://arxiv.org/abs/2601.11916.

Cheddad, A. et al. Digital image steganography: Survey and analysis of current methods. *Signal Processing*, 90 (3):727–752, 2010.

Cooper, A. F. et al. Machine unlearning doesn't do what you think: Lessons for generative AI policy, research, and practice. *arXiv preprint arXiv:2412.06966*, 2024.

Ge, H. et al. When backdoors speak: Understanding LLM backdoor attacks through model-generated explanations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, pp. 2278–2296, 2025. arXiv:2411.12701.

Hubinger, E. et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Lanham, T. et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Liu, X. et al. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.04451.

Mougan, C. et al. The science and practice of proportionality in AI risk evaluations. *Science*, 391(6787):769–771, 2026. doi: 10.1126/science.aea3835.

NVIDIA. Confidential computing on NVIDIA Hopper GPUs: A performance benchmark study. Technical report, NVIDIA, 2024.

Oliynyk, L. et al. Model leeching: An extraction attack targeting LLMs. *arXiv preprint arXiv:2309.10544*, 2023.

OpenMined. Secure enclaves for AI evaluation. Technical report, OpenMined, 2024.

Panfilov, A. et al. Strategic dishonesty can undermine AI safety evaluations of frontier LLMs. *arXiv preprint arXiv:2509.18058*, 2025.

Schnabl, C. et al. Attestable audits: Verifiable AI safety benchmarks using trusted execution environments. In *Workshop on Technical AI Governance (TAIG) at ICML 2025*, 2025.

Shevlane, T. et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

Stosz, C. et al. Aef-1: minimum operating conditions for independent third party ai evaluations. Technical report, AI Evaluator Forum, 2025., 2025.

Tice, C. et al. Noise injection reveals hidden capabilities of sandbagging language models. *arXiv preprint arXiv:2412.01784*, 2024.

Tinfoil and Workshop Labs. Private post-training and inference for frontier models. Technical report, Tinfoil, 2026.

Tlaie, A. and Farrell, J. Securing external deeper-than-black-box GPAI evaluations. *arXiv preprint arXiv:2503.07496*, 2025.

Tramèr, F. et al. Stealing machine learning models via prediction APIs. In *25th USENIX Security Symposium*, pp. 601–618, 2016.

UK AI Safety Institute. Auditing games for sandbagging detection. Technical report, UK AI Safety Institute, 2024.

Yadav, C., Laufer, E., Boneh, D., and Chaudhuri, K. ExpProof: Operationalizing explanations for confidential models with ZKPs. In *Workshop on Technical AI Governance (TAIG) at ICML 2025*, 2025.

Yu, M. et al. SafeSeek: Universal attribution of safety circuits in language models. 2026. arXiv:2603.23268.

Zhang, Y. et al. NVBleed: Covert and side-channel attacks on NVIDIA multi-GPU interconnect. *arXiv preprint arXiv:2503.17847*, 2025.