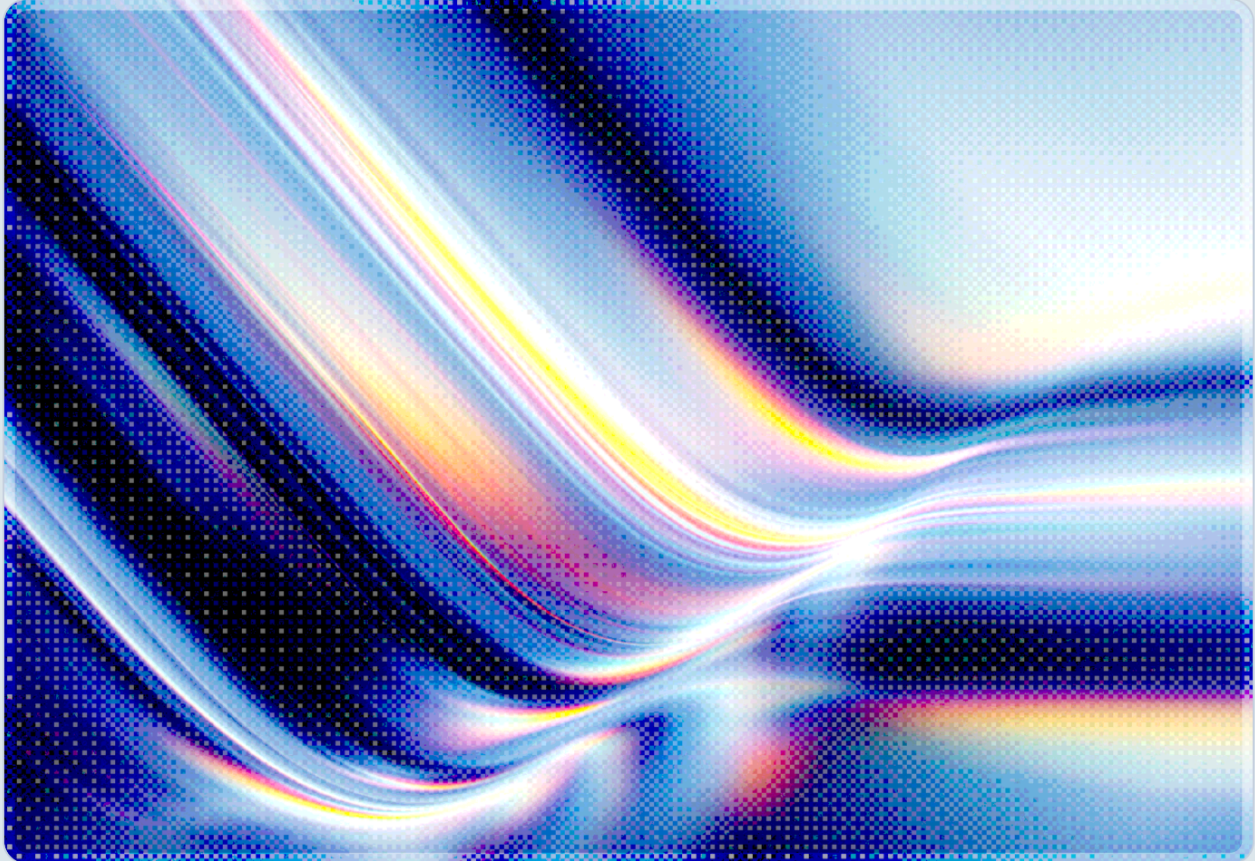


Embedded Independent Evaluators: What Does Independence Establish?

Reading Anthropic's METR, SecureBio, and Accenture
Arrangement

2026-10-17

 AI-assisted Rapid Research



CSAI

cloud
security
alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

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

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

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

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

Key Takeaways

On October 15, 2026, Anthropic described a tiered program that gives outside organizations progressively deeper access to its internal information in order to assess its risks and safety practices [1]. The most extensive tier, currently occupied only by the nonprofit METR, places six METR employees on Anthropic hardware, in Anthropic office space, and on internal tools [1]. SecureBio is reviewing the chemical and biological weapons section of Anthropic's August 2026 Risk Report, and Accenture and Faculty are described as working toward a partnership covering red-teaming, alignment assessments, and safeguard testing [1].

The arrangement departs from evaluations that required the developer to approve a report before release, because Anthropic states that it cannot veto publication by the higher-tier evaluators and may only require redaction in predefined categories [1]. On its stated terms, that no-veto, redaction-only design gives evaluators more editorial independence than the first tier of the same program, in which Anthropic approves publication [1]. No published, end-to-end track record yet exists for it. Whether it delivers assurance or something closer to capture will depend on mechanisms the announcement describes only briefly: who pays, what happens to evaluators who publish unwelcome findings, and how readers can verify the scope of access an evaluator actually had.

For enterprises that rely on frontier model providers, the practical lesson is to treat "independently evaluated" as a claim that needs its own due diligence. Buyers and risk teams should ask which tier of access applied, what was redacted, whether the evaluator was free to say so, and whether the evaluator has financial or operational ties to the developer. Established assurance disciplines, from financial auditing to cloud certification, already supply the vocabulary for those questions.

Background

Anthropic published the update on independent evaluators on October 15, 2026, and added a footnote clarification on October 16 [1]. The post references the company's August 2026 Risk Report and sets out three access tiers. At the first tier, evaluators conduct narrowly scoped risk investigations, with Anthropic determining the level of access and approving publication. At the second tier, evaluators can publish without Anthropic's approval and receive access to more information, although Anthropic may require redaction of specific categories and the evaluator may publish its views on whether those redactions

were justified. At the third tier, evaluators receive employee-like access; METR is the only current participant [1]. The program terms in this note are as described by Anthropic, the party being evaluated, and this note has not located a METR statement that independently confirms them.

The tiers are summarized below.

Tier	Access	Publication control	Current participant(s)
1	Narrowly scoped investigations; access set by Anthropic	Anthropic approval required	SecureBio (pilot review, with increased access under consideration; tier assignment is described inconsistently in public coverage) [1]
2	More information than Tier 1	Evaluator may publish without approval; Anthropic may require redaction of defined categories	None named in the announcement, which does not mean the tier is unused
3	Company hardware, office space, internal tools	Similar to Tier 2; no veto, redaction only in predefined categories such as customer data, privileged communications, and model architecture details	METR (six employees) [1]

Public descriptions of SecureBio's tier are not fully consistent. Anthropic's own post describes SecureBio's pilot review and says the company is considering increased access after the current work [1]. This note relies on Anthropic's primary post and flags the discrepancy rather than resolving it.

METR is not new to Anthropic's reporting. It has previously published external reviews of Anthropic risk material, including a March 2026 review of the Opus 4.6 sabotage risk report [2] and a May 2026 review of the "Risks from automated R&D" section of Anthropic's February 2026 Risk Report [3]. In the latter, METR agreed with Anthropic's bottom-line assessment that the risk of catastrophe from automated R&D by Opus 4.6 or a less capable model is very low, while stating that the evidence in the report was inadequate to establish that conclusion [3]. That review also notes that Anthropic shared additional non-public materials with METR [3]. The finding is relevant because it shows an outside reviewer publicly disagreeing with the developer's reasoning while agreeing with its conclusion, which is the kind of output an assurance relationship should be able to produce. It is a single data point, however, and the review predates the embedded arrangement, so it does not show how METR will behave under it.

According to the announcement, METR's embedded team operates under enforceable agreements, other METR employees are siloed from the team's work, and METR reserves the right to make certain statements about the level of access it was given [1]. Anthropic also states independence criteria: evaluators should not hold equity in the company they evaluate or be financially or operationally dependent on it [1]. The announcement does not specify funding arrangements or compensation [1]. Accenture and Faculty are described as working toward an "ambitious partnership," and Hugging Face is in early discussions [1].

Security Analysis

The design has real strengths, but embedding also creates independence threats that the announcement leaves largely open. This section first sets out the strengths of the design, then the places where capture risk concentrates, and finally compares the arrangement with established assurance regimes. Evaluative statements here are the judgment of the authors rather than reported fact.

Strengths of the Design

In our assessment, the most significant feature is the separation of access from approval. Under the older pattern, in which an evaluator submits a report to the developer for review before release, the developer holds an implicit veto, and readers cannot tell whether a missing caveat reflects the evaluator's judgment or a negotiation. A no-veto, redaction-only model narrows the developer's editorial power to enumerated categories and, at the higher tiers, lets the evaluator characterize those redactions publicly [1]. This resembles a familiar principle in assurance: the party being assessed may protect legitimate secrets but should not control the conclusion.

A second strength is the explicit reference to financial and operational dependence as disqualifying or at least concerning [1]. Auditor independence regimes in other fields treat both financial relationships and the provision of certain non-audit services as threats, on the principle that an assessor should not perform management functions or review its own work [4]. Anthropic's stated criteria are a short version of that logic, applied to an organization type, the research nonprofit, that is common in AI evaluation but rare in conventional audit.

Third, embedding addresses a real information problem. Frontier AI risk assessment depends on non-public facts such as training details, internal monitoring data, and incident records. An evaluator that can see only what the developer chooses to send can validate arguments but not the completeness of the

evidence behind them, and METR's own earlier review illustrates the dependence on developer-supplied material [3]. Employee-like access lets an evaluator ask for the thing the developer did not think to include.

Where Capture Risk Concentrates

The first threat is economic. The post does not say how evaluators are funded [1]. If a single developer is the primary source of an evaluator's revenue, access, or reputation, the evaluator faces an incentive to preserve the relationship. This is an inference from general independence principles rather than a reported fact about METR, and nothing in the cited material suggests METR has acted on such an incentive. It is, however, the question a relying party would reasonably ask, and the independence criterion Anthropic itself states, no financial or operational dependence, cannot be checked without the answer [1].

The second threat is familiarity. In our view, close and sustained working relationships can erode skepticism even in the absence of financial ties, a concern that independence frameworks in financial audit commonly address through measures such as partner rotation, although the sources cited here do not cover those provisions. Six evaluators sharing offices, tools, and day-to-day context with the company they assess is a stronger form of that exposure than a periodic engagement. The siloing of the embedded team from the rest of METR [1] manages information leakage; it does not by itself manage familiarity. Rotation, cooling-off periods, and a second reviewer without embedded access are plausible mitigations drawn from that audit practice, and the announcement does not state whether any apply.

The third threat is the commercial character of Accenture. METR and SecureBio are nonprofits, while Accenture is a large consulting firm, and the announcement says only that it is working toward a partnership on red-teaming, alignment assessments, and safeguard testing [1]. Red-teaming and safeguard testing performed as a paid service can be a legitimate assurance input, but the same firm advising on, building, or remediating safeguards would face the self-review and management-function threats that financial independence rules were written to address [4]. Whether Accenture's role is assessment, advisory, or both is not specified in the sources reviewed, and this note treats it as an open question.

The fourth threat is the redaction channel. Predefined categories such as customer data, privileged communications, and model architecture are individually defensible [1]. The risk would lie in category breadth over time and in whether readers can tell what was removed. The announcement's provision that the evaluator may publish its view on whether redactions were justified [1] is the central safeguard, but its value depends on being exercised and on the evaluator being willing to bear the relationship cost of doing so. METR's reservation of the right to make statements about its level of access [1] is a related disclosure that should be read closely in any resulting report.

Finally, scope matters. The announced program covers an evaluator's view of risk assessments and safety practices, not a certification that a model or system is safe. A published review of one section of a risk report, as in the SecureBio and METR examples, supports confidence in that section and no more. Readers who treat an embedded-evaluator reference as a blanket assurance would overstate what the evidence supports.

Comparison with Established Assurance Regimes

Several existing regimes address the same structural problem, and a comparison helps clarify what is missing. Public-company auditing is governed by SEC and PCAOB independence rules that restrict financial and business relationships and certain non-audit services, and PCAOB rules require audit committee pre-approval of non-audit work [4]. For management system certification, ISO/IEC 42006 sets requirements for bodies that audit AI management systems under ISO/IEC 42001, building on ISO/IEC 17021-1 to address competence and impartiality [5][6]. CSA's STAR program addresses the same relationship in cloud assurance by separating self-assessment, third-party certification, and continuous auditing levels [7][8].

Dimension	Financial audit	Management-system certification	Embedded evaluator (as announced)
Independence rules	Codified (SEC/PCAOB) [4]	Impartiality requirements for certification bodies [5] [6]	Stated criteria; no public enforcement mechanism described in [1]
Who pays	Audited entity, with audit committee oversight [4]	Client, within impartiality rules [5]	Not specified [1]
Scope disclosed	Standardized opinion [4]	Certificate scope [5]	Evaluator may state level of access [1]
Oversight of the assessor	Regulator inspection [4]	Accreditation body [5] [6]	None described in [1]

In our assessment, the last row is the most significant gap. Established regimes pair independence rules with an external party that inspects the assessors. The embedded-evaluator model currently relies on the evaluator's own reputation and on contractual terms that readers cannot inspect.

Recommendations

The recommendations below move from steps that individual organizations can take now, through changes to vendor risk questionnaires, to the longer-term question of whether embedded evaluation can mature into a recognized assurance discipline.

Immediate Actions

Organizations that cite a provider's independent evaluations in vendor risk files should record the specific claim, the evaluator, the tier or access level, and the scope reviewed, rather than a generic "independently evaluated" note. When a provider's published report relies on an evaluator, relying parties should read the evaluator's own publication, not only the developer's summary, and compare what each says about access and redactions. Where an evaluator has disclosed its level of access, as METR reserves the right to do [1], that statement should be filed with the vendor assessment.

Short-Term Mitigations

Procurement and third-party risk teams can add questions to AI vendor questionnaires covering evaluator funding, financial or operational dependence on the developer, the categories and volume of redactions, and whether the evaluator has any consulting or implementation relationship with the developer. For evaluators that are also commercial service providers, as Accenture is, relying parties should ask whether the firm has advised on or built the safeguards it tests. Teams should also keep evaluation findings separate from certification claims: a published review of a risk report section is evidence for a risk argument, not an attestation of control effectiveness.

Strategic Considerations

The broader question is whether embedded evaluation will mature into a recognized assurance discipline. That would likely require published independence standards for AI evaluators, covering funding concentration, rotation, cooling-off periods, and non-assessment services; disclosure norms for access and redaction; and oversight of evaluators by a party other than the developer. Industry, standards bodies, and assurance programs are all candidates to supply these elements, and the ISO/IEC 42006 and STAR models offer templates [5][7]. Developers that adopt embedded evaluation would allow outsiders to test the independence claim by publishing funding terms and the full text of redaction categories. Because the arrangement is new and its track record is short, organizations should revisit their assessment as METR and SecureBio publish results under the new terms.

CSA Resource Alignment

CSA's assurance work is the most directly relevant prior art. The [STAR Program Overview](#) describes how CSA structures assurance in layers, and the questions raised here about who assesses, who pays, and what is disclosed map onto the distinctions between self-assessment and third-party certification in that program [7]. [The Evolution of STAR: Introducing Continuous Auditing](#) is relevant to the temporal question: an embedded evaluator observes continuously, much as continuous auditing does, and similar issues of auditor familiarity and evidence integrity apply [8]. The [STAR Level and Scheme Requirements](#) document provides a model for specifying what a certifying party must satisfy and how scope is stated [9].

For control content, the [AI Controls Matrix \(AICM\) v1.1](#) supplies the control objectives against which model providers can be assessed, and its auditing guidelines for model providers describe how auditors approach them [10]. CSA's STAR for AI direction is the natural venue for any standardized treatment of evaluator independence for frontier model providers. This note does not claim that CSA has published guidance specific to embedded evaluators.

References

- [1] Anthropic. "[An update on independent evaluators](#)." Anthropic, October 15, 2026 (footnote clarification added October 16, 2026).
- [2] METR. "[Review of the Anthropic Sabotage Risk Report for Opus 4.6](#)." METR, March 12, 2026.
- [3] METR. "[Review of the 'Risks from automated R&D' section of the Anthropic Risk Report, February 2026](#)." METR, May 8, 2026.
- [4] PCAOB. "[Ethics & Independence Rules, including Rule 3525 on Audit Committee Pre-approval of Non-audit Services](#)." Public Company Accounting Oversight Board.
- [5] ISO/IEC. "[ISO/IEC 42006: Requirements for bodies providing audit and certification of artificial intelligence management systems](#)." ISO, 2025.
- [6] ICAEW. "[New standard for firms certifying AI management systems](#)." ICAEW, August 2025.
- [7] Cloud Security Alliance. "[STAR Program Overview](#)." CSA, 2024.
- [8] Cloud Security Alliance. "[The Evolution of STAR: Introducing Continuous Auditing](#)." CSA, 2024.
- [9] Cloud Security Alliance. "[STAR Level and Scheme Requirements](#)." CSA, 2022.
- [10] Cloud Security Alliance. "[AI Controls Matrix v1.1](#)." CSA.