

Anthropic gives METR employee-like access to evaluate its AI models

Source copy kept by wheresthe.ai for the story “Anthropic gives METR employee-like access to evaluate its AI models” (<https://wheresthe.ai/d/us-anthropic-gives-metr-employee-like-access-to-evaluate-its-2026-10/>).

Source: <https://www.anthropic.com/news/an-update-on-independent-evaluators>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: ae9187948d60cc2bf440c136503535c70736a9c7bebaa70e01a136bdf2e54e53

Documents the source offers: <https://www-cdn.anthropic.com/files/4zrzovbb/>

[website/0a58d567024a8b448ff15158ebc3625328dfcc1f.pdf](https://www-cdn.anthropic.com/files/0a58d567024a8b448ff15158ebc3625328dfcc1f.pdf); [https://www-cdn.anthropic.com/](https://www-cdn.anthropic.com/files/f61d49fa5596956a5dec75fea0e973bf6a6a8378/Redacted%20Risk%20Report%20August%202026%20.pdf)

[f61d49fa5596956a5dec75fea0e973bf6a6a8378/Redacted%20Risk%20Report%20August%202026%20.pdf](https://www-cdn.anthropic.com/files/f61d49fa5596956a5dec75fea0e973bf6a6a8378/Redacted%20Risk%20Report%20August%202026%20.pdf)

Announcements An update on independent evaluators Oct 15, 2026

Last month, we committed to stronger independent, third-party assessment of our models and safety practices. We also signed the White House Accord on Super Intelligence, in which companies pledged to partner with independent external auditors or evaluators. Today, we’re announcing that we’ve started to give employee-like access to a team from METR, a nonprofit that evaluates frontier AI models.

METR will publish the results of its risk assessments. Over the past year, we’ve worked with METR on several reviews of our Risk Reports. We’re also working with SecureBio and Accenture as evaluators, and are in early discussions with several other organizations, including Hugging Face.

Here, we explain our approach to working with evaluators. Why we work with independent evaluators As models have become more powerful, they’ve required stronger safeguards and more challenging risk assessments. We want to be publicly accountable for the decisions we’re making. But we can’t publicly share all the information relevant to our safety work. Some of it could be useful to malicious actors, or is commercially sensitive, like compute or training information. Instead, we’re giving third-party evaluators deep access to our information, and are having them share evidence and reasoned conclusions about the risks involved in building our models. We aim to close the gap between what evaluators understand about a frontier AI company’s systems and what the company itself understands.

To evaluate frontier models, evaluators need the following: Expertise in assessing the capabilities of frontier models, including understanding the way evaluations can go wrong, such as alignment faking. Independence: they shouldn’t hold equity in the company they’re evaluating, or be financially or operationally dependent on it. Enough access to meaningfully evaluate risks. Beyond access to the latest models, evaluators also need visibility into how those models were trained, into how our safeguards function, and more. Freedom to publish observations and assessments of a model or a process.

Although the analogy isn’t exact, bank supervision operates on similar logic. Banks know more about their own risk positions than outsiders do, so regulators assign resident examiners to work inside the largest banks. If a bank’s risk-taking could threaten the public, third parties try to see that in advance.

For the AI industry, this is a step toward robust safety oversight. We’re moving quickly with multiple evaluators and learning actively about what kinds of partnership work. We’re committed to sharing updates as we go. What we mean by “employee-like access” We’ve onboarded six METR employees with company hardware, office space, and access to internal tools and systems that are roughly on par with what our own risk assessment teams have access to. Outside a small number of clearly defined categories (e.g., data about customers, attorney-client privileged information, or information about our model architectures), Anthropic employees are encouraged to share information with them just as they would with a fellow employee working on alignment risk. 1 METR will publish its analysis of our risk levels, incidents, or practices. Anthropic can’t veto publication, but we can require them to redact information if it falls into predefined categories like the ones above. (Anthropic ultimately decides what information does and does not fall into these categories, although METR reserves the right to make certain statements about the level of access and information it received.) 2 Our engagement with METR is limited to a small team operating under enforceable agreements. Other METR employees are siloed from that team’s work, including researchers who work with us on other projects. Embedded evaluation gives a few qualified people working under Anthropic’s security standards a close view of how we operate. It’s not a way for other

organizations to obtain our information, or that of our customers and partners. Our other evaluator partnerships We're working on several other partnerships. Because these other partnerships are still in early stages, we're providing these evaluators with more limited information, and they have more limited publication rights. Our goal with these early engagements is to make sure that our partners protect confidential information and have the requisite resources and expertise to evaluate our models. As partnerships progress and we build trust, we hope to fully embed more independent evaluators.

We're developing three tiers of access for evaluators. Over time, we hope to have several evaluators at the third tier, with each team publishing their own analysis of our risk posture and key decisions.

In the first tier, evaluators conduct narrowly scoped risk investigations. We decide what access to provide a first-tier evaluator, and evaluators can publish their findings with our approval.

In the second tier, evaluators can publish without our approval, and they have access to more information. We can still require removal of specific categories of information, though evaluators can publish their views on whether our redactions were justified.

The third tier is the arrangement we've described with METR, with employee-like access and publication rights similar to those of the second tier.

SecureBio is in the process of reviewing the chemical and biological weapons chapter of our August Risk Report , after completing and publishing a pilot review of our previous risk report. After this review is complete, we will consider moving to projects with increased access. 3

With Accenture and Faculty, we're working toward an ambitious partnership that includes evaluating and red-teaming models, conducting alignment assessments, testing model safeguards, and building capabilities across our organizations.

We're embedding evaluators voluntarily, and we believe every frontier AI company should do the same.

Footnotes Because these evaluators (like bank examiners) may also work with other companies, we exercise caution around the possibility of inadvertent leaks. Information about architectures presents a high risk of inadvertent leaks, and its relevance to risk assessment is typically low. When it is highly relevant, we will generally seek to share it, though this will require explicit internal approvals. Updated on October 16 to clarify METR's current project terms with Anthropic. Updated on October 15 to clarify SecureBio's current project terms with Anthropic.

Related content

Preparing for the 2026 US midterms Earlier this year, we shared an update on our election safeguards. Here, we outline how we've continued to strengthen these safeguards ahead of this year's US midterms. [Read more](#)
Anthropic opens Singapore office and signs MOU with the Government of Singapore Anthropic has opened an office in Singapore and signed an MOU with the country's government to advance safe, responsible AI adoption. [Read more](#)
2026 Usage Policy update We're publishing a new version of our Usage Policy. In this post, we summarize the changes we've made. [Read more](#)