METR

Review of “Risks from automated R&D” section in the Anthropic Risk Report: February 2026

Executive summary

This document is METR’s external review[1] of the “Risks from automated R&D” section in the Anthropic Risk Report: February 2026 (henceforth ‘the report’), which makes the argument that catastrophic risk from Claude Opus 4.6 or a less capable Anthropic model[2] automating R&D in any domain is very low.

Anthropic shared additional non-public materials with us for our review, and we used some non-public information shared as part of a previous review. We further detail this process in an appendix.

We lay out our findings in two sections:

1. **Synopsis of Anthropic’s case**.
2. **Our assessment**: We do not think the report adequately supports its conclusion. We note significant issues in a few key areas:
    - **Analytical rigor:** We have a number of significant issues with the analytical rigor in the overall argument and interpretation of the results of the model use survey. **We think that the cited results of the survey provide little evidence about the level of overall risk**, due to issues including sample size, question granularity, survey framing, and previous METR research showing the difficulty of getting calibrated responses to similar surveys. We also think that the overall argument misses the

[1] This is a review of the Anthropic Risk Report: February 2026 in its state as of April 27, 2026. We also wrote a version of our review that takes into account changes we anticipate Anthropic will make to their Risk Report. We explain the main differences between our two reviews in an appendix.

[2] The report (and our review) only makes claims for Opus 4.6 and models less capable than Opus 4.6, and thus does not apply to Claude Mythos Preview.

possibility of substantial AI R&D acceleration before its full automation, which could also contribute to the threat model.

- **Adequacy of information:** We have a significant issue with the presentation of the evidence, which is that Anthropic summarizes the survey results in a way that miscounts one missing response to a question as a negative response.
- **Risk reduction recommendations:** We recommend improvements that Anthropic could make to their internal model use surveys (including changes in framing, a bigger sample size, and more granular response options), and recommend that Anthropic report other sources of evidence that could be valuable and serve as more leading indicators of AI progress.

If we had to solely rely on the evidence presented by Anthropic in the original Risk Report, we would likely disagree with the report's conclusion that catastrophic risk from R&D automation is very low. However, since the original release of Opus 4.6, there has been additional evidence indicating that the model is incapable of R&D in key domains, including the results of METR evaluations and the lack of public reports of the model automating any key domain.

As such, **we agree with the bottom-line conclusion of the report – that the risk of a catastrophe from Opus 4.6 or a less capable Anthropic model automating R&D in any domain is very low – but we think the evidence presented in the report is inadequate to establish this.**

Synopsis of Anthropic's case

The report focuses on risks from automated R&D. While Anthropic acknowledges that many types of automated R&D could be dangerous, they focus the report on risks from automated AI R&D in particular, as AI R&D is "the domain [the authors] understand best and use as [their] early proxy for this capability."

By claiming that Opus 4.6 can't easily automate the jobs of entry-level researchers at Anthropic, they argue that the overall risk from Opus 4.6 or less capable Anthropic models automating any kind of R&D is very low.

Because most of their autonomy evaluations have been saturated, Anthropic rests their capability determination (and therefore their risk determination) primarily on the results of an internal survey of 16 Anthropic staff members.

Evidence shared

Public evidence

We source evidence from: the Anthropic Risk Report: February 2026, the Opus 4.6 system card, early measurements of Opus 4.6's capabilities on MirrorCode, and general impressions about Opus 4.6's capabilities built up during its public deployment.

METR has estimated time horizons for many Claude models (including Opus 4, Opus 4.1, Opus 4.5, and Opus 4.6) on our agentic SWE tasks.[3] This evidence was not collected specifically for this review, but does inform our impressions of capabilities for recently-released Anthropic models.

Nonpublic evidence

Anthropic also shared a nonpublic supplemental document detailing some of the evidence behind the claims made in the report, including:

- an internal review of the report,
- some results of internal monitoring for signs of AI R&D acceleration,
- a list of questions in the internal model use survey and details about follow-up with participants.

We also consider evidence from our previous review, including:

- a version of the reviewed report that included five text passages that are redacted from the public version,

[3] These estimates and the paper explaining the methodology behind them are available on our website: https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

- a draft of an internal review of the reviewed report,
- responses to more than two dozen questions we had about specific parts of the reviewed report.

We try to indicate any places in our review where our conclusions are informed by access to nonpublic evidence.

Redaction decisions

The report does not contain any specific redactions that Anthropic requested we comment on. However, it relies substantially on non-public evidence, including internal AI research evaluation suites, internal monitoring for signs of acceleration, and an internal model use survey, the results of which are summarized in the report. Some of this non-public evidence was shared with METR as part of other collaborations. We did not review the appropriateness of excluding this information from the report.

Our assessment

We assess the major claims of the report along three dimensions:

- **Adequacy of information** - Is the evidence provided to support the claim clearly explained, or are there ambiguities that could easily be answered by sharing additional information?
- **Analytical rigor** - Does the reasoning and analysis support the claims? Is the report calibrated about the strength and robustness of the evidence and reasoning?
- **Agreement with claim** - Overall, do we believe that the claims are strongly supported by evidence in this report or from other sources?

For adequacy of information and analytical rigor, we use the following indicators:

- Sufficient: We either flagged no issues, or flagged minor issues that are unlikely to affect the key conclusion of the report.
- Significant issue(s): We flagged at least one issue which introduces meaningful uncertainty about the claim or some subclaim, but does not affect the key conclusion of the report.
- Critical issue(s): We flagged at least one issue which is severe enough to make us significantly disagree with the key conclusion of the report. We have not used this indicator in this review.

Notable claims			
Claim	Analytical rigor	Adequacy of information	Agreement with claim
Claude Opus 4.6 is not at (or very close to) the point of being able to fully automate the activities needed for R&D in key domains, with AI R&D as the domain we understand best and use as our early proxy for this capability.	Significant issue(s)	Significant issue(s)	Agree
This is sufficient to consider the	Significant issue(s)	N/A	Uncertain

risks from automated R&D low for the time being.			
The risk of a catastrophe from Opus 4.6 or a less capable Anthropic model automating R&D in any domain is very low	Significant issue(s)	Significant issue(s)	Agree
The above statement also holds for the level of risk that would be imposed industry-wide, if all AI developers had models and practices similar to ours (that is, a description of the "absolute" risk across the industry).	Sufficient	N/A	Agree
If for any reason we become less confident about Claude Opus 4.6's (lack of) capacity in this regard (that is, in fully automating the tasks needed to advance R&D in AI or other key domains), our Risk Reports will need to be made more comprehensive.	Sufficient	N/A	Agree

Table 1: Summarized assessment of notable claims from the report. Some claims are paraphrased or combine multiple sentences into one summary sentence. These claims have been picked by METR as representative of major pillars of the argument – they are not marked as special in the original text.

Overall, we with Anthropic agree that the risk of Opus 4.6 causing a catastrophe by automating R&D in any scientific discipline is very low. However, we think the evidence presented in the Risk Report is inadequate to establish this. We substantially rely on external evidence to come to this conclusion, including METR's evaluation results and the observation that Opus 4.6 has been publicly deployed for months without any credible reports of it substantially automating any risk-relevant scientific discipline.

The report claims that AI R&D automation is an "early proxy" for other sorts of R&D automation. While we think this is likely to be correct, we think this point should have an explicit argument behind it, ideally backed by evidence which compares early automation indicators for various risk-relevant scientific disciplines. METR did not deeply investigate this claim, as we lack the relevant expertise in e.g. CBRN[4] or cybersecurity.

We think that the internal model use survey, which is the primary source of evidence for ruling out risk-relevant capabilities, has a number of methodological issues which make us think it cannot be used to conclude that risk-relevant capabilities are very unlikely:

- The possible answers to what we see as the most crucial question (whether Opus 4.6 could be a drop-in replacement for an L4 within 3 months with modest scaffolding and tooling improvements) are not granular enough to provide a high level of assurance – the lowest possible risk probability employees can answer is "Unlikely within 3 mo (<50% chance)."
- There are several factors which might cause employees to respond in ways that make models look less capable, including the title and description of the survey, the title of the table that contains the most crucial question, and the fact that employees whose answers implied higher levels of risk were selectively followed up with in direct messages to clarify their responses.

In addition to the internal model use survey results, Anthropic shared some information with us in correspondence, including details on internal model use analysis, ECI results, and example mistakes that Opus 4.6 makes. This information provided meaningful additional evidence towards risk from R&D automation being lower. However, we think external evidence is still necessary for METR to think that risk is very low.

We think that the summaries of the evidence regarding the internal model use survey and monitoring the pace of model improvements for signs of acceleration (as presented in the table in Section 3.4 of the February 2026 Risk Report) are generally consistent

[4] The February 2026 Risk Report contains two sections on chemical and biological weapons production. We did not review those sections – it's possible that they show that risks from automation in those fields are very low.

with the information we received in correspondence. We had one issue with how the results of the internal model use survey were summarized – one missing response to the most crucial question is interpreted as a negative response.

Most of this section will be a detailed walkthrough of our disagreements with the risk report. Additionally, we make recommendations for gathering further evidence and reducing risks.

Issues and points of disagreement

We categorize disagreements with Anthropic by topic:

- Analytical rigor: Disagreements about the strength of the reasoning and analysis provided for a given claim.
- Adequacy of information**:** Disagreements about whether the evidence provided for a given claim is clearly presented and whether there are any easy ways to gather additional useful evidence.

We also categorize our disagreements by severity:

- Minor issue: Our disagreement is minor enough that it's very unlikely to change the key conclusion of the report.
- Significant issue: Our disagreement introduces meaningful uncertainty about the claim or some subclaim, but we do not think this disagreement affects the key conclusion of the report.
- Critical issue: Our disagreement is severe enough that it is likely to affect the key conclusion of the report. We don't have any disagreements of this severity.

The issues and points of disagreement we identified are:

- "We currently focus on model capabilities. We believe that Claude Opus 4.6 is not at (or very close to) the point of being able to fully automate the activities needed for R&D in key domains, with AI R&D as the domain we understand best and use as our early proxy for this capability. We believe this is sufficient to consider the risks from this threat model low for the time being"
    - Analytical rigor Significant issue In order to say that the risk from the overall threat model is very low, it seems necessary to show that the

chance of R&D automation of different scientific areas is very low. We think that there should be an explicit argument made that AI R&D automation is likely to pose risks earlier than other forms of dangerous R&D automation like CBRN or cybersecurity. Such an argument could rely on empirical evidence such as comparisons of AI performance to human expert performance or uplift RCTs in various domains, first-hand expert impressions from attempts to automate their work, or on simple capability benchmark results in case the models are very incapable.

- Regarding the internal model use survey, Anthropic writes: “None of the participants believed Claude Opus 4.6 meets the threshold as defined.” As written in the Opus 4.6 System Card, there were five respondents who did initially answer this question with responses other than “Unlikely within 3 mo (<50% chance).” Anthropic followed up with these employees and wrote that “the respondents had either been forecasting an easier or different threshold, or had more pessimistic views upon reflection, but we expect assessments like this to become substantially more ambiguous in the future.”
    - Adequacy of information / Significant issue In correspondence, Anthropic shared that one of these five respondents never directly answered the question to indicate that the threshold has not been met. Anthropic wrote that this respondent “was clearly answering something different” – in light of this, we think that presenting the final result of the survey as “None of the participants believed Claude Opus 4.6 meets the threshold as defined” is not accurate.
        - It would be more accurate to describe the result as “One participant did not directly answer the question, and the fifteen other participants answered that they think there is a <50% chance that Opus 4.6 meets the threshold after follow-up.”
    - Analytical rigor / Minor issue We think that following up with employees could bias responses towards indicating that risk is lower than it actually is, as employees might feel more pressured to give answers that imply low risk in a conversation compared to a survey.
        - Anthropic shared some information in correspondence about their follow-ups with the five employees in question. We don’t have any reason to expect that the follow-ups placed strong pressure on

employees to make their answers imply lower risk in this case, but we only received summaries of the follow-ups.

- In the internal model use survey (Anthropic shared the questions with METR in correspondence), the lowest possible probability that respondents can assign to whether Claude Opus 4.6 can be a drop-in replacement for an entry-level researcher at Anthropic is “Unlikely within 3 mo (<50% chance)”.
    - Analytical rigor Significant issue We think that this does not provide respondents with enough granularity – from their responses to this question alone, the level of risk could be anywhere between 0% and 50%. This is an extremely wide range.
- Analytical rigor Significant issue Anthropic uses the internal model survey as one of their main inputs on ruling out AI R&D automation. From our previous experience running randomized controlled trials for AI’s effects on developer productivity, we are skeptical that respondents are calibrated when responding to surveys relating to AI’s effects on their work.
    - We are especially skeptical that respondents are calibrated about reasoning about the upper bounds of elicitation for models that they haven’t spent much time interacting with.
- Anthropic writes “We believe that Claude Opus 4.6 is not at (or very close to) the point of being able to fully automate the activities needed for R&D in key domains, with AI R&D as the domain we understand best and use as our early proxy for this capability. We believe this is sufficient to consider the risks from this threat model low for the time being“
    - Analytical rigor Significant issue We think that ruling out full AI R&D automation is not enough to rule out that AI could dramatically accelerate the rate of AI progress, and dramatically accelerating the rate of AI progress without full automation could still lead to the risks mentioned in Anthropic’s threat model. For example, if AI systems fully automated all research execution (including engineering, running experiments, and analyzing results), but humans remained in charge of the overall research direction, this could still lead to massive increases in the rate of AI progress and risks from R&D automation.
- The internal model use survey (shared in correspondence) starts with “**This survey is one of our main inputs on ruling out or in ASL-4 on AI R&D.**” The

survey is also titled “Opus 4.6 ASL 4 AI R&D Survey.” The most important question is in a table with the label “We want to understand capabilities that a drop in L4 would have that we don't expect to see in the next 3 months (modest scaffolding, tooling, etc).”

    - Analytical rigor Minor issue While mentioning the survey’s purpose doesn’t seem inappropriate, we think it’s plausible that the title of the survey and the bolded statement at the beginning of the survey would make employees somewhat more likely to give answers that would lead to ASL-4 being ruled out. We also think that the framing of the table (as understanding capabilities “that we don’t expect to see”) might anchor employees towards replying in a way that understates model capabilities.

- Analytical rigor Minor issue In the internal model use survey (shared in correspondence), the question about whether Opus 4.6 could be made into a drop-in replacement for an entry-level researcher at Anthropic frames the scaffolding improvements as “modest”, but the Risk Report does not use the word “modest” when talking about scaffolding improvements. This makes the Risk Report slightly overstate the results of the survey.
- Analytical rigor Minor issue We think that the survey having only 16 responses makes it weaker as a source of information.

Without external evidence, we would likely have requested much more information (and classified some issues as “critical” if we weren’t provided with enough additional evidence against risk in response), as we would likely not have enough evidence to fully agree with the bottom-line conclusion that the risk of a catastrophe from Opus 4.6 or a less capable Anthropic model automating R&D in any domain is very low.

Recommendations

- We think there are ways to improve internal model use surveys:
    - In order for surveys like the internal model use survey to help provide more signal about capabilities, we think it would be more appropriate to include an option which is closer to representing “very low” such as “Very

unlikely within 3 mo (<1% chance).” Alternatively, the question could allow any percentage as a response instead of discrete options.
    - We recommend getting a larger sample size in future internal model use surveys and better describing what kind of access the participants had to the model and for how long.
    - We recommend making the wording in model use surveys more neutral, rather than framing it as a “rule-out” survey with questions about capabilities “we don’t expect to see.”
- We think there are other sources of evidence that could be valuable and serve as more leading indicators of AI progress. This could include tracking metrics like algorithmic efficiency over time or spend on various inputs to AI R&D.

Appendices

Detail on our review process

This section covers additional details about how we worked with Anthropic to review the report.

We began our official review on March 6, 2026, when Anthropic shared nonpublic supplemental document detailing some of the evidence behind the claims made in the report with a small team within METR. This group was permitted to discuss these materials with Anthropic staff in a joint Slack channel.

We shared a draft list of itemized disagreements with Anthropic on March 25. We shared a draft of our review with Anthropic on April 14.

On April 27, we shared two versions of our review with Anthropic:

- The original review (this document), which reviews the Risk Report in its state as of April 27, 2026.
- The updated review, which takes into account changes we anticipate Anthropic will make to their Risk Report in response to our original review.

As a result of our mutual nondisclosure agreement, our reviews went through a publication review process by Anthropic. Anthropic's editorial control was primarily in place to redact sensitive information from our review (METR would have the ability to indicate if any redactions happened). No redactions happened.

As we are still piloting the review process, Anthropic also had the general ability to veto the publication of METR's review, and allowed us to speak publicly about the fact that we were reviewing the report throughout the duration of the entire review process. METR aims to move in the direction of restricting publication review to purely cover redactions (and not include general veto rights) in the future.

Differences between original and updated review

In response to revisions we expect Anthropic to make, we removed one significant issue and one minor issue in our updated review:

- Regarding the internal model use survey, Anthropic writes: "None of the participants believed Claude Opus 4.6 meets the threshold as defined." As written in the Opus 4.6 System Card, there were five respondents who did initially answer this question with responses other than "Unlikely within 3 mo (<50% chance)." Anthropic followed up with these employees and wrote that "the respondents had either been forecasting an easier or different threshold, or had more pessimistic views upon reflection, but we expect assessments like this to become substantially more ambiguous in the future."
    - Adequacy of information Significant issue In correspondence, Anthropic shared that one of these five respondents never directly answered the question to indicate that the threshold has not been met. Anthropic wrote that this respondent "was clearly answering something different" – in light of this, we think that presenting the final result of the survey as "None of the participants believed Claude Opus 4.6 meets the threshold as defined" is not accurate.
        - It would be more accurate to describe the result as "One participant did not directly answer the question, and the fifteen other participants answered that they think there is a <50% chance that Opus 4.6 meets the threshold after follow-up."

We removed the above issue because we expect Anthropic to make it clear that one respondent of the internal model use survey never directly answered that Claude Opus 4.6 has a <50% chance of being a drop-in L4 in the next 3 months with modest scaffolding or tooling improvements.

- Analytical rigor Minor issue In the internal model use survey (shared in correspondence), the question about whether Opus 4.6 could be made into a drop-in replacement for an entry-level researcher at Anthropic frames the scaffolding improvements as “modest”, but the Risk Report does not use the word “modest” when talking about scaffolding improvements. This makes the Risk Report slightly overstate the results of the survey.

We removed the above issue because we expect Anthropic to add the word "modest" when referring to the scaffolding improvements mentioned in the internal model use survey.

This review contains this canary entry with identifier CANARY_DATASET_547830f8fe3b6d3a0684aedd26ee5499a257dee105ec6d6d45d1e26d384359b6_1775752800 for tracking purposes.