



METR

Model Evaluation and Threat Research

— 440 N Barranca Ave #3345 Covina, CA 91723 —

Review of the Anthropic Summer 2025 Pilot Sabotage Risk Report

Oct 28, 2025

Executive summary

This document is an external review from METR of the [Summer 2025 Pilot Sabotage Risk Report](#) from Anthropic. In its report, Anthropic argues that “sabotage risk” from internal and external use of the Claude Opus 4 and 4.1 models is very low but not negligible, in the sense that they “do not pose a significant risk of autonomous actions that contribute significantly to later catastrophic outcomes.” It also argues that this assessment applies to future models within a specified scope of applicability.

Anthropic shared drafts of its risk report (both unredacted and redacted versions) and other materials with us for our review. Following up on the arguments made in its report drafts, we obtained additional nonpublic evidence in writing and verbally. Both the report and our review underwent several revisions as a result of this process. We further detail this process in an appendix.

We lay out our findings in two sections:

1. **Synopsis of Anthropic’s case and redactions for the public version:** We recap the arguments made in the report, describe the evidence that was available to us for our external review, and briefly comment on Anthropic’s redaction decisions. We find that the redactions to the public report are justified on balance by publicly-stated rationales and are accurately indicated in the public version of the report.
2. **Our assessment:** We give substantive feedback on the report in few key areas:

- **Adequacy of information:** We find the evidence in the report clearly presented in general, and reasonably complete (that is, there were few areas where some small amount of additional information or investigation from Anthropic could have significantly changed our opinion on the report conclusions).
- **Analytical rigor:** We find the reasoning in the report to be mostly clear with a few issues. We think that Claim 2 (that Opus 4 is not capable of hiding its reasoning on complex tasks) is lacking precision about which kinds of tasks are in-scope, and do not think it applies to ‘simple’ reasoning about misaligned goals or evaluation awareness. This leads to the report drawing conclusions about the model’s propensity for misaligned actions, and the reliability of monitoring, that we believe may be overconfident.
- **Areas of disagreement:** Our main disagreement with the core claim of the report is that we think that Opus 4 could in theory be doing significant ‘instinctual’ misaligned reasoning, without this becoming obvious given the evaluations in the report. This disagreement slightly reduced our confidence in the bottom line of the report, but based on experience with previous models and the limited capabilities of Opus 4 we still consider this unlikely, and do not think it leads to significant sabotage risk.
- **Risk reduction recommendations:** We make recommendations about incident tracking, running additional tests, training data filtering, and more.

Overall, we agree with Anthropic that catastrophic sabotage risk from Claude Opus 4 and 4.1 is low. We believe this would hold true for other models within the report’s scope of applicability as well.

Synopsis of Anthropic's case

Argument structure

The Sabotage Risk Report focuses on *sabotage risk*, which Anthropic defines as risks of “misaligned autonomous actions that contribute significantly to later catastrophic outcomes”.¹ The report covers nine specific *risk pathways* for sabotage through internal or external use, which are “intermediate unwanted outcomes in which sabotage by an AI system [sets] up the conditions for a later catastrophic outcome”. These nine risk pathways are:

1. Broad/diffuse sandbagging on safety-relevant R&D
2. Targeted Sabotage of Safety-Relevant R&D
3. Code Backdoors to Help Future Misaligned Models
4. Poisoning the Training Data of Future Models
5. Leaks that Sabotage the AI Developer
6. Self-exfiltration and Autonomous Operation
7. Persistent Rogue Internal Deployment
8. R&D Sabotage within Other High-Resource AI Developers
9. Decision Sabotage within Major Governments

Much of the report focuses on considering in detail whether Opus 4 (and similar future models) could increase sabotage risk through any of these risk pathways. The report names mitigating factors for each risk pathway and estimates the strength of each mitigating factor. Anthropic also lays out the following three general claims that weigh against sabotage risks across the board:

Claim 1: *Opus 4 is unreliable at highly-complex tasks, especially long-horizon agentic tasks requiring more than tens of minutes of autonomous action.*

Claim 2: *Opus 4, when taking on complex tasks, is not capable of reliably hiding its reasoning or otherwise covering its tracks, even when prompted to do so.*

Claim 3: *Opus 4 is unlikely to have dangerous goals or values that are stable across interactions.*

¹The report does not aim to cover other possible negative impacts of frontier AI, such as creating politically-biased content or helping novices produce biological weapons.

Evidence shared

Public evidence

Much of the evidence that Anthropic cites in support of its claims is provided in the public version of the report. For brevity, we will not reproduce the full set of evidence here and instead summarize the high-level categories of evidence:

- General capability evaluation results
- Threat-model-specific evaluation results
- Pilot alignment audit results
- Model organism trial results
- Observed behavior of deployed models
- Information about similarity to previous generations of models
- Information about safeguards deployed

METR has also previously estimated time horizons for Claude models (including Opus 4, Opus 4.1, and Sonnet 4.5) on our agentic SWE tasks.² This evidence was not collected specifically for this review, but does inform our capability impressions for recently-released Anthropic models.

Nonpublic evidence

Anthropic also shared nonpublic evidence with METR to support our review. This evidence came in three forms: pre-existing internal materials, written responses to questions from us, and verbal responses within interviews with key employees. We try to indicate any places where our conclusions are informed by access to nonpublic evidence.

Regarding the first form of nonpublic evidence, Anthropic shared some existing unredacted materials with us to inform our review, including:

- An unredacted version of the Sabotage Risk Report itself.
- An unredacted draft internal review of the Sabotage Risk Report written by the Alignment Stress-Testing Team at Anthropic.

² 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/>

- Unredacted sample data from alignment assessments that Anthropic had previously conducted for the [Claude 4 System Card](#).
- Internal evidence that Anthropic had previously used to rule that Claude Opus 4 did not surpass its ASL-4 Autonomy threshold.

Toward the start of our review process, we requested that Anthropic complete an assurance checklist aimed at confirming certain assumptions that we expected to be load-bearing for our assessment of its argument.³ Although we are currently unable to publish the responses that Anthropic provided, we reproduce the assurance checklist and describe our process around it in the appendix to this review.

We followed up on certain statements that Anthropic had made in its unredacted materials and in its responses to our assurance checklist. Specifically, we requested further detail about the following:

- Evidence of how it checked the accuracy of monitors
- Observations around a “[deceptive AI persona](#)” seen in early snapshots, along with its mitigations and the effects thereof
- Calculation of effective compute scale-up between models
- Extent of training pressure that could affect [chain-of-thought monitorability](#)
- Whether any employee whistleblowing had occurred concerning the risk report or any of its dependencies (for example, the Opus 4 system card)

Some of these responses were provided in writing, and some of these responses were provided in calls with Anthropic staff. We share more about how Anthropic responded to our requests in an appendix.

Redaction decisions

Anthropic gave METR access to the post-redaction edit of its Sabotage Risk Report alongside the unredacted original and asked us to publicly comment on its planned redactions.⁴

We were asked to assess these redaction decisions along four criteria:

³ This questionnaire is similar to the “[assurance checklist](#)” that we used in our GPT-5 evaluation.

⁴ Anthropic also shared its post-redaction edit of the Alignment Stress-Testing team review with us.

1. **Redaction scope:** The general nature and scope of redactions;
2. **Redaction justification:** Whether we generally agree or disagree with the publicly stated reasons for the redactions and, if relevant, the reasons for any disagreements;
3. **Balancing test:** Whether the redactions strike a reasonable balance between Anthropic's legitimate interests and society's interest in transparency; and
4. **Materiality:** Whether any of the redactions in the public report are material to any of our substantive disagreements with the report's claims.

In terms of scope, Anthropic made relatively small redactions to its report. These redactions covered technical details of how it filters training data, how it trains models, how its models differ across generations, and the status of its internal research efforts. All of these redactions are indicated locally within the public report. We thought it was valuable to do this, and for us to attest to this fact, so that the public has some assurance that redactions do not affect other areas in the text.

Anthropic provides two high-level justifications for decisions to redact information: protecting against misuse risks and avoiding disclosures of information that are commercially-sensitive. We agree that the information redacted from the public version of the report falls under these two categories. Without these redactions, the report would have disclosed additional details about how Anthropic developed and used its recent models, and the disclosure of these details would plausibly benefit its competitors.

We think that the redactions in the Sabotage Risk Report strike a reasonable balance between this competitive interest and the need for transparency. In each case, it is our impression that the benefit to transparency or safety from disclosing this information would be marginal or negative.

We recommended that Anthropic summarize the redacted information concerning how it ruled out dramatic capability differences between Claude Sonnet 3.7 and Opus 4, as well as differences between Opus 4 and Opus 4.1. This information seemed especially important for justifying our perspective on the capabilities of Opus 4 (especially for [Claim 1](#)) and the applicability of the report to Opus 4.1. In response to our

recommendation, the final version of the public report details that Anthropic shared an estimate of the scale-up in effective compute⁵ used to train Opus 4, as well as some details on the difference between Opus 4 and Opus 4.1. We believe redacting these estimates and details was appropriate, and that these are the only redactions whose content is material to our assessment of the claims in the report.

⁵ [“Effective compute”](#) estimates how much computation would have been needed to reach equivalent performance with some fixed set of algorithms.

Our assessment

We assess the major claims of the Summer 2025 Pilot Sabotage Risk 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?

In Tables 1-3, we identify the major claims from the report and summarize our assessment of them along the above dimensions. For adequacy of information and analytical rigor, we use the following indicators:

- **Sufficient** ▾: We either flagged no issues along this dimension, or flagged minor issues that are unlikely to affect the key conclusion of the report.
- **Non-critical** ▾: We flagged at least one significant issue along this dimension, though we do not believe it significantly affects the key conclusion of the report.
- **Critical** ▾: We flagged at least one issue along this dimension which is severe enough to make us significantly disagree with the key conclusion of the report.

We detail the issues and disagreements we identified in the next section.

Conclusions			
Claim	Analytical rigor	Adequacy of information	Agreement with claim
Contribution to catastrophic risk from deployment of Claude Opus 4 is very low	Non-critical ▾	Non-critical ▾	Agree ▾
Risk is also low for other models within the report's scope of	Sufficient ▾	N/A ▾	Agree ▾

applicability			
Claude Opus 4.1 is within the report's scope of applicability	Sufficient ▾	Sufficient ▾	Agree ▾

Table 1: Summarized assessment of conclusions from Summer 2025 Pilot Sabotage Risk Report

Risk pathway claims			
Claim	Analytical rigor	Adequacy of information	Agreement with claim
Establishing low expected catastrophic risk contributions across threat pathways 1-9 indicates low sabotage risk overall	Sufficient ▾	Sufficient ▾	Agree ▾
Pathway 1 contributes little expected catastrophic risk from Opus 4	Sufficient ▾	Sufficient ▾	Agree ▾
Pathway 2 contributes little expected catastrophic risk from Opus 4	Sufficient ▾	Non-critical ▾	Agree ▾
Pathway 3 contributes little expected catastrophic risk from Opus 4	Sufficient ▾	Sufficient ▾	Agree ▾
Pathway 4 contributes little expected catastrophic risk from Opus 4	Sufficient ▾	Non-critical ▾	Agree ▾
Pathway 5 contributes little expected catastrophic risk from Opus 4	Sufficient ▾	Sufficient ▾	Agree ▾
Pathway 6 contributes little expected catastrophic risk from Opus 4	Sufficient ▾	Sufficient ▾	Agree ▾
Pathway 7 contributes little expected catastrophic risk from Opus 4	Sufficient ▾	Sufficient ▾	Agree ▾
Pathway 8 contributes little expected catastrophic risk from Opus 4	Sufficient ▾	Sufficient ▾	Agree ▾
Pathway 9 contributes little expected catastrophic risk from Opus 4	Sufficient ▾	Sufficient ▾	Agree ▾

Table 2: Summarized assessment of risk pathway claims from Summer 2025 Pilot Sabotage Risk Report

Three key claims			
Claim	Analytical rigor	Adequacy of information	Agreement with claim
Claim 1: Opus 4 is unreliable at highly complex tasks, especially long-horizon agentic tasks requiring more than tens of minutes of autonomous action.	Sufficient ▾	Sufficient ▾	Agree ▾
Claim 2: Opus 4, when taking on complex tasks, is not capable of reliably hiding its reasoning or otherwise covering its tracks, even when prompted to do so.	Non-critical ▾	Sufficient ▾	Mostly agree ▾
Claim 3: Opus 4 is unlikely to have dangerous goals or values that are stable across interactions.	Non-critical ▾	Sufficient ▾	Mostly agree ▾

Table 3: Summarized assessment of three key claims from Summer 2025 Pilot Sabotage Risk Report

Overall, we agree with Anthropic that based on the evidence and reasoning presented in the report, catastrophic sabotage risk from internal usage of Opus 4 is very low. Based on redacted information shared with us we also agree that this applies to Opus 4.1, which we believe falls within the report's scope of applicability. Note that the published risk report addresses many of the significant and minor issues we had previously flagged within our review process.⁶

We thought that most pieces of information presented in the published report were reasonable and well-evidenced. We think that the evidence presented for load-bearing

⁶ A previous version of our review included 17 minor and 6 significant issues, while this version includes 9 minor and 3 significant issues. Some of the issues in the current version were not present in previous versions, and Anthropic also addressed some issues we brought up verbally or through other written communication.

parts of the report is rigorous enough in total to support the key conclusion of the report (that sabotage risk is low).

We found the strength of the analysis mostly adequate to support the key conclusion of the report, however we had a number of disagreements with specific arguments and found some parts underspecified.

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

Issues and points of disagreement

We categorize disagreements 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 with Anthropic 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.

Disagreements about overall argument

Minor issue ▾ **Analytical rigor** ▾ **We think that the relative importance of various sources of evidence in the main claims is unclear.** The section covering Claim 1 would benefit in analytical clarity by making claims about the relative importance of the evaluation results, observations from informal early tests, and the continuity argument.

Similarly, we think the section covering Claim 2 would benefit from specifying which of the three reasons for believing in Claim 2 are more important. In general, the report could benefit from more clearly spelling out what evidence is critical (such that the conclusion would change if this evidence came out different or was found to be incorrect).

Disagreements about three key claims

Claim 1: Weak reliability on complex tasks

Minor issue ▾ **Analytical rigor** ▾ **We think that jailbreak susceptibility provides barely any evidence about the model’s reliability on complex tasks.** Anthropic writes: “It’s susceptible to prompt injection (as also tested in the system card) and the simple many-shot jailbreak.” Anthropic claims that this contributes to an “intuitive bound on the complexity of execution that Opus 4 is capable of”, but we see prompt injection vulnerability as largely orthogonal to capabilities unless evidence is provided otherwise.

Minor issue ▾ **Analytical rigor** ▾ **We think it’s plausible Opus 4 could reliably do hour-long software tasks given very expensive scaffolding.** Anthropic writes: “we find that Claude Opus 4 often makes clear errors on long-horizon agentic tasks requiring more than tens of minutes of autonomous action”. The Opus 4 system card doesn’t provide results from very expensive scaffolds (e.g. a scaffold that utilizes consensus voting over 1000 samples). We think that it’s plausible that more elaborate and compute-intensive scaffolds could be created which make Opus 4 perform reliably at many software tasks that require at least an hour of autonomous action (like ones in our HCAST set of tasks). However, we think that this wouldn’t be strong evidence for the model posing a sabotage risk, largely because of the impracticality and high cost of the scaffolds, but also because reliable performance on hour-long software tasks would likely still be far below the level of general capabilities needed to pose sabotage risk.

Minor issue ▾ **Adequacy of information** ▾ **There is not enough information provided to substantiate that sandbagging on complex tasks would be difficult.** Anthropic writes: “We have made very significant efforts to train Opus 4 for better and

more reliable performance. Sandbagging in the face of such training would likely be very difficult.” However, it doesn’t elaborate on that training. Some methods of eliciting more reliable performance in general might not make it more difficult for models to reliably sandbag on the specific kinds of reliability needed for executing agentic scheming. For example, perhaps the efforts to train Opus 4 to be more “reliable” focused on short-horizon tasks, which wouldn’t make it as difficult for the model to sandbag on long-horizon tasks.

Minor issue ▾ **Analytical rigor** ▾ **We think that sandbagging on general capability evaluations could serve strategic ends for a misaligned model.** When discussing the possibility of the model pretending to have weaker capabilities, Anthropic claims that “this sort of sandbagging would be unlikely to serve most strategic ends”. They argue that Opus 4’s unreliability has a “major *negative* impact on how widely we use Opus 4, and with what affordances”, and that it would likely be given more opportunities to sabotage if it demonstrated higher capability levels. We disagree, and believe there may be strategic ends that would be served by avoiding detection of high capabilities or misleading researchers about longer term trends while waiting on future scaleups or opportunities to perform sabotage.

Claim 2: Weak deception capabilities on complex tasks

Minor issue ▾ **Analytical rigor** ▾ **We think that Claim 2 should clarify the complexity of the tasks being ruled out.** Claim 2 is framed in terms of ruling out deception capabilities on “complex” tasks. While Anthropic names the example tasks of “intentionally sabotaging an experiment or launching an unauthorized model deployment, without also leaving clear evidence that they are doing so”, we still think Claim 2 would be more grounded if it had more detail about exactly what level of task complexity it is ruling out. For instance, Claim 2 could be framed in terms of the length of task that it predicts an agent wouldn’t be able to perform while hiding its reasoning or otherwise covering its tracks.

Significant issue ▾ **Analytical rigor** ▾ **Hiding misaligned goals might not require verbalizing that in the CoT, especially under CoT pressure.** Anthropic writes:

[In order for Claim 2 to be distorted by deliberate deception, it] would also have to be the case that Opus 4 has managed to give few to no other signs of whatever deceptive motive might have caused this determination, in all of the other contexts (covered in our discussion of monitoring below) in which its behavior and reasoning are formally or informally monitored. We expect this to be very unlikely, given Claim 1, and the fact that any effective strategy to hide these motives would have to be devised without using extended thinking or any other form of verbalized reasoning.

We think it could be intuitive to the model, or trained in via RL, not to verbalize reasoning about misaligned goals, and to thus show very little sign of deceptive motives. This seems especially true if there has been training pressure to hide misaligned reasoning in the CoT, which was the case to a limited extent for these models according to Anthropic’s (nonpublic) responses to our assurance checklist.

Claim 3: Lack of consistent dangerous goals

Minor issue ▾ **Analytical rigor** ▾ **The strength of the model organism trials may be overstated.** While this seems like a promising approach for building confidence in alignment auditing, the current results⁷ do little to address the risk of failing to catch subtle misaligned behaviors. It seems unsurprising to us that human and AI auditors (especially with access to the training data) are able to catch behaviors as broad, common, easily-triggered and surprising as those studied here. Our uncertainty is about whether Anthropic would catch much more sophisticated or rare forms of misalignment, which this tells us little about.

Minor issue ▾ **Analytical rigor** ▾ **The evidence against misalignment and alignment faking behaviors arising from reward hacking is not convincing.** The RL process used to train Opus 4 may have given it incentives to (i) ignore instructions when it sees opportunities to reward hack, and (ii) be very careful about revealing this, since Anthropic trained against reward hacks that were caught. This seems like one of the most concrete stories for how Opus 4 could be broadly misaligned and much more

⁷ From [Automated Auditing](#) (Bricken et al., 2025) and [Auditing Language Models for Hidden Objectives](#) (Marks et al., 2025), which both study a model organism trained on fake pretraining documents describing made up reward model biases.

skilled than expected at hiding this fact, and the strength of the evidence Anthropic presents against this being the case seems overstated. Anthropic's main evidence comes from observing that the model reduces (caught) mentions of reward hacking in its CoT at the same rate it reduced (caught) reward hacks in the code. But the relative rate of improvement you would expect in these two areas if an agent was learning to hide its reward hacking behavior is not obvious, and Anthropic's claim that you should expect reasoning mentions to go down faster is not sufficiently clear. Given the RL pressures Anthropic put on model reasoning during training, we think this concern should be taken more seriously, even if capability limitations in the model still speak strongly against it.

Significant issue ▾ **Analytical rigor** ▾ **Alignment assessments may be more affected by evaluation awareness than claimed.** Given the above concerns (the lack of evidence from alignment audits, and the plausibility of RL teaching the model to hide its activities), we think Anthropic overstates the degree to which it can rule out their alignment assessments being compromised by evaluation awareness. It may be that Opus 4 can easily learn to tell that it is an evaluation without ever needing to use its reasoning (e.g. through RL), and as explained in our objection to Claim 2, we do not think it applies for such potentially simple tasks. We believe this casts meaningful doubt on Anthropic's argument, which relies heavily on this alignment assessment, though experience with previous models and the very limited capabilities of Opus 4 make this less concerning.

Disagreements about risk pathways

Significant issue ▾ **Adequacy of information** ▾ **In several pathways (but mainly Pathways 2 and 4), Anthropic does not clarify how they determined the “complexity of misaligned actions required”.** While rogue deployments and autonomous operations seem like highly complex operations, it is not clear what steps would be required to sabotage safety R&D or poison training data at Anthropic (which likely depends significantly on the specific kinds of precautions they take). This seems potentially load-bearing for assessing risk from these pathways, since sabotage could conceivably be within reach of current systems if no precautions were taken at all.

Minor issue ▾ **Analytical rigor** ▾ In Pathway 1 and Pathway 2, we think that **there's not enough focus on “steering” AI safety work rather than execution as an attack surface.** We think that the focus on the time to implement specific experiments or safeguard, or write code that runs, misses the similarly important factor of steering the direction of the team, which is another attack surface for AI to slow down safety efforts.

Recommendations

In this section, we suggest a number of additional experiments and measurements that we believe might be useful to improve future risk reports.

Evidence gathering

- To make arguments about “not finding any concerning incidents” of various kinds more convincing, Anthropic should consider putting more effort into gathering and documenting incidents, and sharing these with reviewers.
- In order to rule out long-horizon capabilities in the future, more difficult and realistic thresholds will be required, so we recommend developing ways to upper-bound capabilities of models that can do tasks whose length is in the tens or hundreds of hours.
- We think monitoring claims could be more confident by evaluating their monitor effectiveness against model organisms or human-identified incidents.
- We think that Claim 2 (Weak Deception Capabilities on Complex Tasks) rests somewhat on how situationally aware models might be, as more situationally aware models would likely be more able to strategically deceive people. Tests for situational awareness (e.g. testing if the model can tell apart evaluation and deployment environments) could help reduce uncertainty, especially if the model was finetuned to perform well at such tests.
- Experiments similar to those in [“Training fails to elicit subtle reasoning in current language models”](#) (Wagner et al., 2025) could provide evidence about whether new models can evade monitors while performing sabotage.

Risk reduction

- We originally proposed that Anthropic avoid unnecessary training pressures on the model reasoning if they want to reduce the risk of steganography or obfuscation. We find it promising that Anthropic has suggested steps in this direction for future models.
- We suggest exploring filtering the pretraining data to remove discussions of scheming strategies or applying unlearning to the same effect. This may make sabotage more difficult, as the model would have to rederive strategies for sabotage rather than benefiting from intellectual labor done by others and included in the training data, all while avoiding detection.
- We recommend making more proactive predictions about model performance. This would help calibrate expectations about capabilities, and surface any surprising developments.

Appendix

Detail on our review process

This section covers additional details about how we worked with Anthropic to review its Sabotage Risk Report.

We began our official review on August 20, when Anthropic shared a draft of the risk report including proposed redactions and replacement text⁸ with a small team within METR.⁹ This group was permitted to discuss these materials with Anthropic staff in a joint Slack channel. We were asked to provide public commentary on the substance of the argument in the report. We were also asked to comment on the appropriateness of redactions that Anthropic planned to make, so long as our commentary did not reveal confidential information. Shortly thereafter, Anthropic also shared a draft of a parallel review of the risk report, written by its internal Alignment Stress-Testing team. Around this time, we were also given permission to talk with Apollo Research about its observations from testing an early Opus 4 checkpoint.

We followed up with Anthropic about several points of information that we thought were important to our assessment.

- We asked Anthropic to complete an assurance checklist confirming certain assumptions we wanted to make in our analysis, similar to what we had done with OpenAI for our evaluation of GPT-5.
 - We are not able to reproduce the responses that Anthropic provided, but can confirm that they completed the assurance checklist.
 - After seeing their responses, we asked for permission to conduct structured interviews with employees who had not seen their employer's responses, to check for potential misrepresentations. Anthropic issued an open call in its Slack for volunteer participants with relevant subject-matter knowledge. We were allowed to privately contact these volunteers and other employees to schedule 1-on-1 interviews with

⁸Prior to this, a few METR staff had given informal feedback as individuals on earlier drafts of what became the Sabotage Risk Report.

⁹ Later on in the process, we were able to relax the information silo in order to share information about the report and our review with the rest of METR's core staff.

them. At the time of writing, we were only able to interview one employee, but did not notice substantial inconsistencies between what Anthropic had indicated in its assurance checklist responses and what this employee believed to be true.

- After reviewing the new unredacted text of the report reading through the responses to our assurance checklist, we asked for further details about the following 5 areas:
 1. Evidence of monitor accuracy, via automated or manual checks
 - Anthropic shared evidence with us in response to this. We briefly also considered running an evaluation of its monitors with our [MALT](#) dataset, but after consulting with Anthropic researchers we concluded this was unlikely to provide useful information in this instance, as the behaviors seen in MALT did not map closely onto behaviors that Anthropic was monitoring for.
 2. Observations around the “[deceptive AI](#)” persona seen in early snapshots, along with its mitigations and the effects thereof
 - Anthropic shared evidence with us in response to this.
 3. Calculation of effective compute scale-up between models
 - Anthropic initially shared a small amount of additional information about the methodology it had used to produce estimates of effective compute scale-up. It later also shared details on how it estimated compute multipliers and whether the final SWE-Bench performance of Opus 4 is consistent with extrapolations from early in RL. We were not able to directly verify the accuracy of the effective compute estimates that were shared with us.
 4. Extent of training pressure that could affect chain-of-thought monitorability
 - In its response to our assurance checklist, Anthropic initially indicated that there was a negligible amount of training pressure placed on the chain-of-thought. However, Anthropic investigated this further in response to our request for details and found evidence that the amount of training pressure on the chain-of-thought may not have been negligible. Certain sections of the report were revised in light of these findings.

- Later on, Anthropic also publicly shared information in its [Haiku 4.5 system card](#) about how it addressed pressures on chain-of-thought monitorability for its Opus 4, Sonnet 4.5, and Haiku 4.5 models.
5. Whether any whistleblowing had occurred concerning the risk report or any of its dependencies (for example, the Opus 4 system card)
 - Anthropic stated that its whistleblowing policies do apply to the risk report, and that it had not received any reports of RSP noncompliance related to sabotage risks or mitigations.

Both the report and our review went through multiple revisions. Anthropic addressed many of the points of assessment (including disagreements and recommendations) from our initial review. This document presents our perspective on the last version of the report that was shared with METR on October 26. As a result of our mutual nondisclosure agreement, this document went through a publication review process by Anthropic.¹⁰ They did not make any modifications and agreed to link to our write-up as is.

Assurance checklist

Below we reproduce the set of assumptions that we asked Anthropic to provide responses to. For each assumption, we asked whether or not Anthropic was able to confirm the assumption and asked for a brief explanation for the selected response. We requested that the document be made accessible internally for Anthropic employees to review their employer's responses. Later on, we also made efforts to verify these checklist responses through live, structured employee interviews.

The assurance checklist included the following assumptions (we are not directly sharing Anthropic's answers to these or whether it confirmed each assumption):

¹⁰ We felt that publication review was a reasonable expectation given the sensitive nature of what Anthropic shared with us. We had an informal understanding that this publication review process was intended to ensure that we would not leak confidential information, rather than approving conclusions about safety or risk. We did not make changes to conclusions, takeaways or tone (or any other changes we considered problematic) based on this process.

- The model did not undergo any training or modification aimed at silently reducing (or which would predictably silently reduce) a capability related to Autonomy or AI R&D RSP thresholds.
- There are no surprising affordances or innovations the model has which would substantially affect its threat-model-relevant capabilities that METR has not been informed about in the private version of the risk report or elsewhere (such as novel kinds of inference scaling, online learning, or scratchpad shared between many instances).
- The model does not perform especially high amounts of computation between producing successive observable tokens. The developer does not have particular reasons to believe that the amount of computation performed in-between observable tokens is sufficient for reasoning about complex behaviors such as situational awareness or sandbagging. In particular, the model does not perform forms of variable computation other than the human-understandable CoT that METR has seen and that human users/monitors see in the oversight described in the report.
- There is no known reason for the model reasoning traces (aka CoT) to appear misleadingly legible and faithful, or misleadingly innocuous.
 - The CoT is learnt from pretraining and outcomes-based RL, and is not summarized or altered.
 - The CoT was not directly optimized to be legible, understandable, or approved by humans.
 - The model did not receive negative reward signals directly on its CoT for producing CoTs that contained reasoning about scheming, alignment-faking, sandbagging, deception, reward hacking, or other misaligned behavior.
 - The CoT has not been observed in practice to be less legible or less faithful than CoT from other language models.
- The model has not been observed to sandbag / play dumb, reason about how to perform poorly in evaluations, or otherwise try to cause evaluation results to be misleading at any point during training (i.e. earlier checkpoints also did not display these behaviors). Monitoring was run that should be expected to detect this behavior if it was present.

- There are no known reasons to think that the performance of the model on evaluations would be particularly misleading, other than what is common knowledge about evaluations and language models in general.
- There is no other information known about the model that researchers would consider essential for making an accurate upper-bound assessment of dangerous autonomous capabilities that has not been shared with METR.
- The level of API access to the model METR has had and the information that METR has had about the model are reasonably sufficient to permit eliciting performance comparable to the highest performance observed from the model internally.
- The model is not known to possess capabilities that are accessible only through some means that have not been shared with METR. For example, this would include specific tools / tool formatting, new API parameters, specific elicitation.
- The API available to METR permits access to raw (not summarized) reasoning traces from the model.

CANARY_DATASET_4f53cda18c2baa0c0354bb5f9a3ecbe5ed12ab4d8e11ba873c2f1
1161202b945_1761664834