Independent Assessment of AI R&D Acceleration at Anthropic in October 2026

Publication date: November 5, 2026

Redaction summary:

As part of this assessment, Anthropic had rights to redact any non-public information in this post. Per our approach to transparency, we had the right to disclose if Anthropic exercised redaction rights, and whether or not any of our findings depended significantly on redacted information. While we were able to freely share our highest-level conclusions about risk, Anthropic redacted a majority of the evidence we found most informative for these conclusions. Anthropic described the redactions as "aimed at removing confidential information that could accelerate other AI developers."[1] Some of these redactions are themselves challenging to describe clearly without revealing sensitive information we have been asked to redact. We share a table of important evidence we were and were not able to discuss given the redactions in our process section below. All redactions that were important for our conclusions are covered by our descriptions in that table or other in-line descriptions in the report below.

Much of the redacted evidence we viewed had multiple plausible interpretations, and both METR and Anthropic researchers have significant uncertainty about how to synthesize the evidence into a holistic picture. This makes it hard for the public to judge this evidence themselves, reconcile it with evidence from other sources, and come to their own conclusions (i.e. this report does not achieve 'transparency about evidence'). As a consequence, the public has to rely on METR's judgment much more than would be ideal, given the weight of the questions answered here.

Executive summary

From August 12 to October 29, 2026, METR staff performed a high-access evaluation at Anthropic to assess the degree to which AI is accelerating AI R&D.[2] If AI systems started to significantly accelerate further AI development, this could lead to a sudden emergence of extreme and risky capabilities.

Anthropic offered unprecedented access to internal data on their AI R&D process, including a variety of metrics on pretraining, reinforcement learning (RL), training compute in each domain, benchmark scores for internal models, AI speedup analyses, forecasts of a future model's metrics, and interviews with relevant staff.

In our first section, we share our subjective conclusions about risk. We now believe that:

- As of October 2026, Anthropic is likely improving the capabilities of their models on well-specified tasks at around twice the rate they did between March 2024 and October 2025. This estimate is given with high uncertainty; we would not be surprised if the true acceleration was closer to 1.3x or 3x the historical rate.
- A majority of this acceleration today is likely attributable to Anthropic's extensive use of AI to do R&D, though there is significant uncertainty and other causes likely played a large role in acceleration in late 2025.
- The lack of observed bottlenecks or strong diminishing returns means there is a significant risk that further capabilities progress could lead to even faster acceleration, though, as we say below, we may be missing important bottlenecks.

In our second section we share some of the key empirical findings from our Full Report:

- Since October 2025, benchmark scores (measured via AECI, Anthropic's version of the Epoch Capabilities Index) for Anthropic's models have been growing at an estimated 1.9x the pre-October 2025 rate [95% confidence interval: 1.5–2.8x], implying that Anthropic has made roughly two years of historical progress over the last year.
- We are confident that the acceleration was not due to an acceleration in the scaling of training compute.
- Interviews and surveys are consistent with AI usage causing a doubling in the pace of capabilities progress. Anthropic staff self-report dramatic uplift from use of AI systems in interviews and surveys, although these are challenging to interpret. We were told that R&D is accelerated primarily via AI assistance

rather than fully autonomous contributions, but there are some signs that fully autonomous contributions will play a larger role soon.

- We did not see evidence for significant bottlenecks which would prevent further capabilities progress due to limits on algorithmic progress, training compute, experiment compute, or data in the next 12 months. However, our understanding of the status of different bottlenecks changed several times as we received more detailed information, so we are not confident in our current assessment.

Some of the key findings that most informed our conclusions about risk were redacted by Anthropic.

In our third section, we describe the assessment process:

- The scope, setup, and timeline of this assessment.
- Details of the information shared with METR.
- Details of what information and evidence we were not able to share given Anthropic's redactions.

Finally, we describe the biggest limitations of our analysis:

- We could be underestimating capability acceleration if we did not assign enough importance to (1) evidence that some growth rates have further increased recently; (2) further automation of AI R&D tasks leading to a further acceleration in progress; (3) evidence that AECI growth has sometimes understated growth in real-world impact.
- We could be overestimating capability acceleration if we did not assign enough importance to (1) non-AI factors in Anthropic's research progress; (2) AECI's focus on verifiable tasks; (3) reasons why Anthropic's early growth rate (2023–2025) was unusually low.

The remainder of the report goes into detailed discussion of each of the key components. We include appendices with supplemental material.

Our subjective conclusions on risk

Our conclusions on risk.

Independently of this assessment, METR published an October 2026 article proposing a set of high-level questions about AI R&D acceleration we believe the public needs to understand to be able to judge loss-of-control risks from frontier AI development. In this section, we describe how the evidence we reviewed in this assessment changed our subjective conclusions on these high-level questions.

Pace of capabilities progress.

A significant acceleration of Anthropic's overall rate of progress would be surprising, since diminishing returns and dependence on compute scaling mean this would likely require remarkable scaling in other inputs.[3] After reviewing the internal evidence, we still have significant uncertainty and internal disagreements on the team, but most METR staff who worked on this project now believe that, as of October 2026, Anthropic is likely improving the capabilities of their models on well-specified tasks at around twice the rate they did between March 2024 and October 2025. We would not be surprised if the current rate of growth was closer to 1.3x or 3x the historical rate. We have additional uncertainty in how this acceleration translates to broader and less well-specified tasks.

The 1.9x acceleration [95% confidence interval: 1.5–2.8x] in the AECI trend since October 2025 described below significantly informed this view, but could be explained by overfitting to the benchmark distribution or otherwise be an artifact of the AECI methodology. On the other hand, we are not confident that AECI would reliably let us estimate recent R&D acceleration, since the data points are lagging acceleration in research and are somewhat sparse.

Beyond the AECI evidence, we also saw redacted evidence that significantly informed our view on this question, and made us take this apparent ≈2x acceleration more seriously. It also gave us some indication that acceleration could be above 2x and increasing over time, though we would be surprised if it was significantly higher.

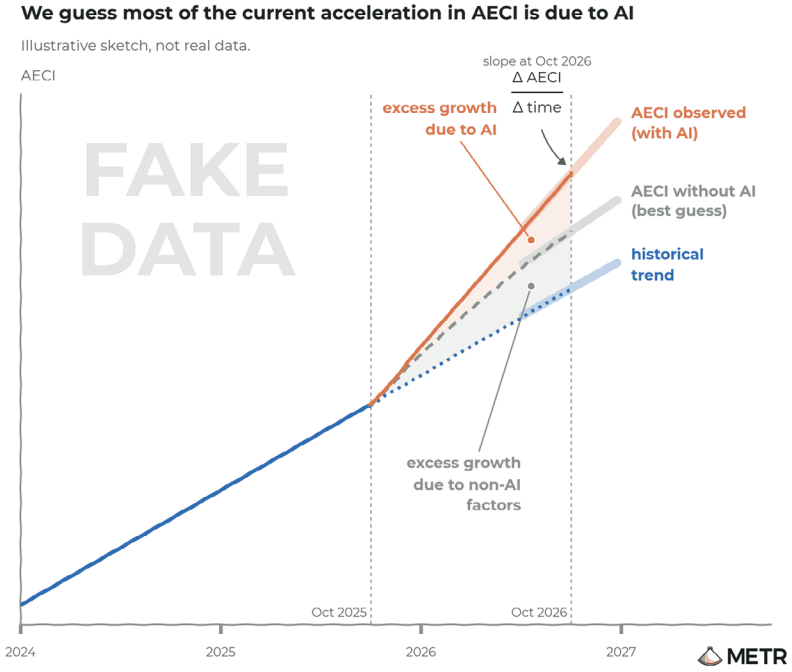
Attribution to AI.

We looked at a variety of factors that could explain the acceleration we observed, including compute scaling, hiring, data improvements, algorithmic discoveries, and resolving other bottlenecks.

As we describe below, we were able to confidently determine that the acceleration was not due to an acceleration in the rate of growth of training compute. Beyond that, there was not a consensus among Anthropic's own staff and we did not find decisive evidence. Our overall subjective estimate is that:

1. **The acceleration in late 2025 was likely influenced by non-AI factors.** We investigated a variety of factors which could have contributed to the observed acceleration in late 2025, but details of many of these factors have been redacted.
2. **We guess that a majority (perhaps ≈80%) of the current (October 2026) acceleration is due to AI.** This judgment is given with substantial uncertainty, but is based on a variety of lines of evidence on contributions from AI and other causes. We believe AI's contribution has grown relatively more important over time, and do not think the other factors we investigated would explain the sustained acceleration. The evidence of speedup from the usage of AI we describe below also seems compatible with a large contribution.

We illustrate these points below:



Our more detailed estimates for the cumulative and instantaneous effects of AI on Anthropic's rate of progress in AECI, and at various stages of the model training pipeline, were redacted.

Bottlenecks to further acceleration.

A key concern is that acceleration from the use of AI could start to compound, as the development of more capable models leads to further acceleration. As we describe below, we did not identify any clear bottlenecks (limited inputs that would prevent further growth) or steep diminishing returns to further scaling that would get in the way of such compounding acceleration, though our understanding evolved several times over the course of this assessment, and we have significant uncertainty. Because of this, we believe there is a significant risk that further capabilities progress could lead to even faster acceleration over the coming year.

Key findings

In this section, we summarize the key empirical findings from our Full Report that we can describe without revealing redacted information (see the process section for more details on what was redacted).

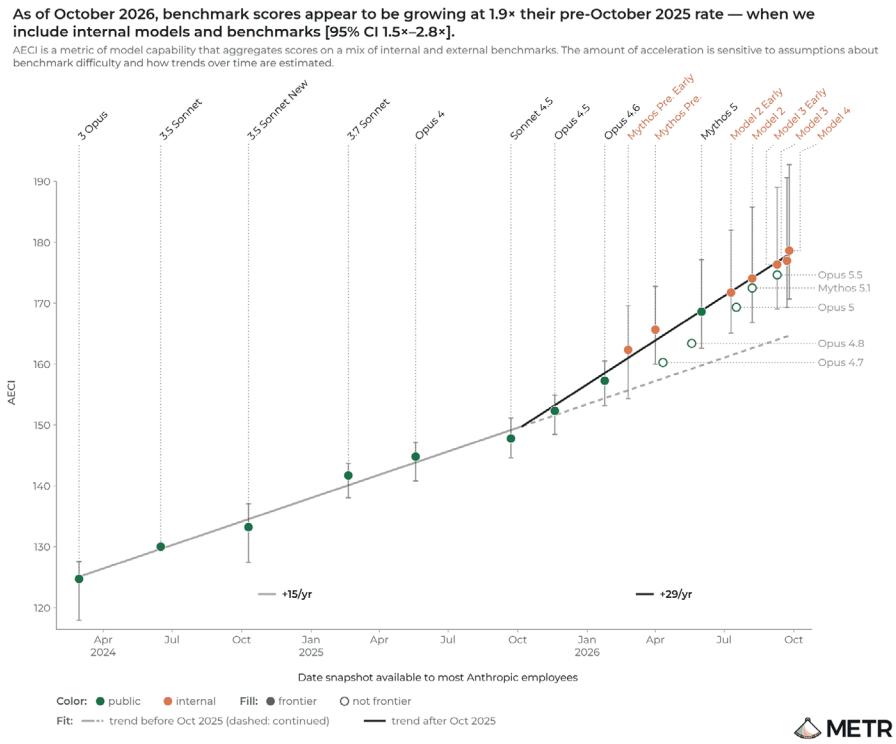
Terminology: We describe growth as annual multiples and acceleration as a ratio of growth rates.

We use "acceleration" to refer to an increase in the rate of growth, such that a 2x acceleration means that 2 years of historical progress is made in 1 year.

If AECI's historical trend was 15 points/year, then increasing to 30 points/year implies an acceleration rate of 2 (it takes 1 year to achieve 2 years' progress at the old rate). For an exponentially growing variable, the acceleration rate is the ratio of growth rates. If a variable is growing 3x/year, then to make 2 years of progress in 1 year the growth rate must be (32)x/yr = 9x/yr (and acceleration can be written as ln(9)/ln(3) = 2).

Since October 2025, benchmark scores for internal models have been growing at around twice their historical rate.

Benchmark scores of Anthropic's models, as measured by AECI (Anthropic's version of the Epoch Capabilities Index), have accelerated 1.9x [95% confidence interval: 1.5–2.8x] since October 2025. Put another way, Anthropic appears to have made as much progress on benchmarks in the past year as they did between October 2023 and October 2025 (roughly, Claude 2.1 to Sonnet 4.5):



Our estimated acceleration is greater than that reported in the Haiku 5.5 system card (1.7x) primarily because of the inclusion of data on newer unreleased models, explained further below.

We investigated the robustness of these results to different choices of benchmark inclusion, model dating, model inclusion, and benchmark saturation. Robustness analyses find acceleration rates between 1.1x and 2.9x. We would be surprised if the true evolution of benchmark scores (e.g. as measured by an improved version of AECI) showed no acceleration or acceleration above 3x. See our appendix section for details.

The acceleration is not driven by an acceleration in the scale of training compute.

We received convincing evidence that the acceleration was not due to an acceleration in the scale of training compute. Further details on this evidence were redacted.

Interviews and surveys are consistent with AI usage driving this doubling in the pace of capabilities progress.

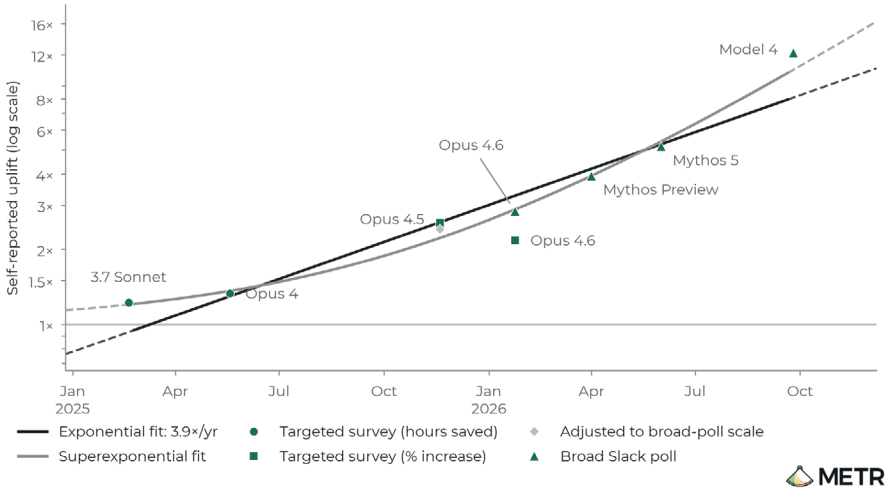
We have many lines of evidence relevant to AI's contribution to progress in different parts of the model training pipeline, including interviews, self-reported uplift, headcount, lines of code, experiment usage, and qualitative reflections from leadership. We highlight three important unredacted sources.

First, both METR and Anthropic interviewed many researchers in detail about the degree of overall acceleration in their area of work attributable to AI use (see details in the appendix below). Our overall impression was a quite wide range of estimates, but centered around a 2x acceleration. Some quotes, with sensitive details redacted, are in the appendix.

Second, Anthropic also shared data on self-reported uplift due to AI from a variety of sources. The most recent data is from a series of Slack polls, where staff use reactions to vote on their estimated speedup from the most recent model at the time. We consider this to be only very loose evidence, as the responses reflect low-effort consideration. Nevertheless, we observe consistent growth in these self-reports, and very high levels (reaching 12x in October 2026), as shown in the chart below.

Informal self-reported uplift over time from Anthropic employees

Results come from informal surveys and Slack polls. They should be taken as very loose evidence.



Note that individual uplift/speedup estimates do not correspond to acceleration rates. Sustaining a particular growth rate in outputs like model capabilities generally requires a similarly sustained growth in inputs like labor and/or labor productivity to overcome diminishing returns. Achieving a significant *acceleration* in the overall rate of progress generally requires a significant acceleration of these inputs. We thus care most about whether we see a large change in the growth rate of individual uplift, not the absolute level. Our more detailed analysis of what rate of growth in individual uplift might be needed to explain our estimated acceleration rates in key areas was redacted.

Third, we understand that AI is not yet able to make highly significant and fully autonomous contributions to the most important parts of the model training process. However, researchers described significant "signs of life" implying such contributions could occur soon, although with substantial variation in maturity depending on the area.

We did not identify any convincing bottlenecks to continued rapid progress.

We did not see evidence for significant bottlenecks in algorithmic progress, training compute, experiment compute, or data that will take effect in the next 12 months. However, our understanding of the status of different bottlenecks changed several times as we got more detailed information, so we are not confident in our current assessment. Anthropic redacted our analysis of a number of specific bottlenecks.

Additional redacted evidence was sometimes hard to interpret, but supports sustained and potentially increasing acceleration.

We investigated a number of internal metrics and methodologies for measuring acceleration in key areas of the model training pipeline. This included evidence of progress on benchmarks and metrics of algorithmic efficiency and overall model strength in pretraining and RL, as well as improvements to compute utilization, inference efficiency and data efficiency. These investigations included attempts to understand:

1. Different methodological variants of internal metrics. We spent multiple weeks on these investigations, and this is our greatest source of uncertainty.

2. The beliefs of staff on the relevant teams about key internal evidence and estimates of acceleration.
3. Whether other parts of the model training pipeline could be biasing our estimates of acceleration, and where adjustments could be applied.
4. Whether Anthropic could be overfitting to internal metrics.

This internal evidence provided support for there being significant acceleration in Anthropic's model training pipeline. Some of this evidence indicated that acceleration could be above 2x and increasing over time, but there are significant uncertainties about how to interpret it.

Process details

METR staff assessed how much AI is speeding up AI R&D at Anthropic.

METR staff[4] (2.5 full-time equivalents) piloted a high-access evaluation at Anthropic to assess the degree to which AI is accelerating AI R&D, and how quickly extreme and risky capabilities may emerge in the near future.

Anthropic lists AI R&D acceleration as a key risk indicator in its Responsible Scaling Policy (RSP). In this assessment, we do not pay special attention to the thresholds set in the RSP. Instead, we try to develop a holistic picture of AI-driven acceleration of AI R&D.

This engagement started in mid-August, and is distinct from the assessment announced on October 15 by Anthropic.

Details of the engagement.

The METR staff worked over eleven weeks to independently assess how much AI is accelerating AI R&D at Anthropic. Our assessment focused on three broad questions:

1. Has the growth rate of Anthropic's internal capabilities accelerated?
2. What share of any acceleration is attributable to AI?
3. Are there important bottlenecks that could slow growth or limit acceleration in the near term?

We were additionally scoped to review the AI R&D section of Anthropic's August 2026 Risk Report, which we are publishing separately. As we describe below, the initial contracted assessment was scoped to include a verification of what percent of compute was allocated to capabilities vs. safety.

METR had the ability to request access to any information we believed relevant to understanding AI R&D acceleration, and Anthropic could unilaterally approve or reject our requests. In practice, we often negotiated with Anthropic staff and compromised towards data that was informative, not extremely sensitive, and easy to share. We give more details on what Anthropic shared as well as what it declined to share below, noting that broadly the access provided surpassed our expectations. For this engagement, we did not request or receive access to sample from non-public models.

Note that neither METR nor Anthropic intended for this project to involve full "employee-like" access, meaning unmediated access to data, documents, or models. In particular, this project is not the one described in Anthropic's October 15 update on independent evaluators.

Anthropic had unilateral redaction rights over all non-public information in this report. We had a unilateral right to disclose:

- the exis*tence and general terms of the review and redaction process,*
- *whether Anthropic exercised redaction rights,*
- *whether or not any findings depended significantly on redacted information,*
- *the level of access provide*d.

For security reasons, we worked on Anthropic devices for the duration of this assessment. Anthropic provided free tokens for this assessment, which we primarily used for AI-assisted data analysis and visualization. Per our standard approach, we did not take payment from Anthropic.

This exercise followed METR's policies for engaging with AI companies. We provide a conflict of interest (COI) disclosure and excerpted contract language in the process transparency disclosures appendix plus the AEF-1 checklist.

Details of Anthropic's redactions.

Our ~100-page Full Report attempts to answer the key questions we recently described as crucial to answer in future AI R&D assessments in detail. We summarize the evidence sources we were given, and which we were able to include in this public report, in the table below.

Table of shared and redacted evidence sources.

Type of information	Shared with METR?	Results described in public report
Measurements of internal capabilities: benchmark capabilities of internal models	Yes	Yes
Measurements of internal capabilities: key intermediate measures	Yes	No
Algorithmic progress estimates	Yes	No
Evidence of AI's contribution	Yes	Partly
Measurements of key R&D inputs	Mostly	No
Details of bottlenecks and forecasts	Partly	No
Interviews with staff	Yes	Partly

Anthropic's redactions also meant that we could not communicate:

1. Our more detailed estimates of acceleration and attribution to AI usage for key areas of the model training pipeline.
2. Our more detailed breakdown of progress attributable to compute, data and algorithmic progress, other than the key takeaway that "We are confident that the acceleration was not due to acceleration in training compute."

3. More detailed answers to the limitations questions in our proposal, which would require more discussion of methodological details.
4. Forecasts of metrics for future models.

They also made some other more minor redactions throughout, which we attempt to describe in each section of this report.

Anthropic described the redactions as "aimed at removing confidential information that could accelerate other AI developers."

Key dates.

August 12	Anthropic begins sharing confidential information with METR.
August 14	Anthropic shares the unredacted version of its August AI R&D Risk Report with METR.
September 12	METR delivers a first draft of its Full Report to Anthropic for feedback.
September 18	METR delivers a first draft of its Risk Report Review to Anthropic for feedback.
September 22-October 22	METR iterates on the Risk Report Review and the Full Report with further information.
October 22	METR delivers its ~100-page Full Report to Anthropic.
October 25	Anthropic shares redaction guidance, summarized above.
October 26	METR delivers a draft of this public report to Anthropic.
October 27-29	Anthropic tightens the redaction guidance, and METR iterates on the public report as a result.
October 29	METR submits a draft public report.
October 29	Anthropic and METR agree on an ~5 day extension to more carefully accommodate the new redaction guidance.
November 3	METR submits the final public report.

METR received a significant amount of private data from Anthropic.

METR made extensive requests for information on every part of the AI R&D pipeline, and Anthropic shared more information than was initially scoped for this project, surpassing our expectations. Information was typically shared within a few days of requesting it.

A detailed description of the data we had access to was redacted. A summarized version is as follows:

1. Internal benchmark scores and AECI for internal models, plus sample tasks, rubric answers and agent expenditure on those tasks.
2. Historical data on algorithmic efficiency and overall model strength in pretraining and RL, as well as improvements to compute utilization, inference efficiency and data efficiency.
3. Headcount and attribution of progress to specific staff.
4. Training compute metrics in many relevant domains.
5. Internal analysis on relevant open questions.
6. Internal surveys of self-reported AI uplift by team.
7. Lines of code and pull requests across the company and some research teams.
8. AI agent inference spend over time.
9. Internal documents on autonomous research contributions.
10. ~15 interviews with key staff, covering a predefined but wide scope that excluded particularly sensitive topics.

In a small number of cases, Anthropic did not provide relevant information when they deemed it too sensitive.

Anthropic also offered extensive staff time and expertise to this project. We believe the two staff we worked most closely with are the two highest-context researchers on AI R&D acceleration measurement at Anthropic. Both of these staff met with us frequently for live discussion, and spent substantial time collecting data, doing analysis, and providing feedback.

Anthropic canceled a compute allocation component of this assessment.

The contracted scope of this project originally included a verification of Anthropic's compute allocation, especially the allocation between capabilities and safety. However, Anthropic changed the scope of this project midway to exclude compute verification and instead just focus on AI R&D acceleration. This component was canceled before METR came to any substantive conclusions on compute usage at Anthropic. Anthropic stated this was due to wanting to keep the engagement focused on risk assessment for the time being, and that it would revisit compute allocation verification later on.

Limitations

We report our best guesses at ways in which we could be misinterpreting the evidence.

We discuss the main ways we could be making mistakes in our interpretation of the evidence. We highlight:

1. Limitations that result from the structure of this assessment.
2. Limitations that result from incorrect analysis.

Limited access may have led us to miss important context or data.

Despite the access provided to METR, it is clear that staff-level access would enable a significantly deeper assessment. Our pipeline of information was relatively narrow, and a small number of our requests for data were denied. Being able to read documents on demand, observe internal discussion and debate, and have frequent and diverse conversations would have allowed for a richer view. While we believe we had access to the bulk of relevant information, it is possible we missed important information because we did not know it existed or could not look for it ourselves.

Over the course of our exercise, we substantially changed our interpretation of key data as more context was shared with us, in some cases leading to multiple reversals on critical questions. We believe fuller access would have dramatically sped up that process of proper interpretation, and we think there likely remain points which we may have misinterpreted due to insufficient context.

Acceleration could be higher than this report estimates.

Our best guesses as to how we might be underestimating acceleration:

1. **We may have missed recent further acceleration.** The rate of growth could be undergoing continuous acceleration, e.g. hyperbolic growth. While we do not see clear evidence of this in the AECI trends, there is other evidence that makes this seem more plausible which Anthropic redacted. If growth is hyperbolic, we are likely underestimating instantaneous acceleration and substantially understating acceleration in the near future. In addition, we observe model-level outcomes only after a model is trained, meaning we observe acceleration in effective R&D with a lag roughly equal to the model's training time.
2. **Automation could be closer than we estimated.** We understand that AI cannot yet fully automate substantial parts of Anthropic's model training pipeline; however, there are "signs of life," discussed below. It is possible that automation is close, and leads to very rapid acceleration once human bottlenecks are removed.
3. **AECI could understate real-world capability impact.** We have sparse data on the relationship between benchmark scores and real-world impact. As one example, ECI scores on math have not accelerated substantially, but we are now seeing the resolution of Millennium Prize Problems that have evaded decades of human effort. Real-world impacts may be downstream of capability thresholds that models suddenly pass, leading to sudden impacts that are hard to predict.

Acceleration could be lower than this report estimates.

Our best guesses as to how we might be overestimating acceleration:

1. **Non-AI factors could explain most of the acceleration.** It's possible that the non-AI factors explain all or most of the recent acceleration. This implies we have less reason to expect high future growth or future accelerations.

2. **Acceleration might be limited to well-specified tasks.** Much of the evidence we looked at measures performance on well-specified tasks that are easy to grade automatically or with a language model. If recent models have been growing more quickly on "clean" tasks than on "messy" tasks (relative to historical relative growth rates), then we may be overstating the acceleration in real-world impact. For instance, recent progress may be overfitting to the types of tasks which are included in AECI.
3. **Our acceleration numbers could be overstated if the historical rate of growth was unusually low.** If progress historically was below its normal or natural rate due to temporary factors, this could lead us to overestimate the amount of acceleration. As we describe in an appendix, Anthropic models had a lower historical growth rate on AECI than the broader industry frontier before late 2025, which could explain some of the acceleration.

Appendices

Methodology

Redactions in this section.

None of this section was redacted, but note that some of the methodological discussion may only apply to the Full Report.

We include data from Anthropic internal-only models.

This includes the following models (these are the codenames shared with us; we understand they have different internal codenames):

- Unreleased 2022, which finished training in 2022
- Model 2, which finished training in July 2026
- Model 3, which finished training in August–September 2026
- Model 4, which finished training in September–October 2026

METR's Full Report includes a forecast for a future model, but this was redacted.

We use piecewise linear fits to detect acceleration.

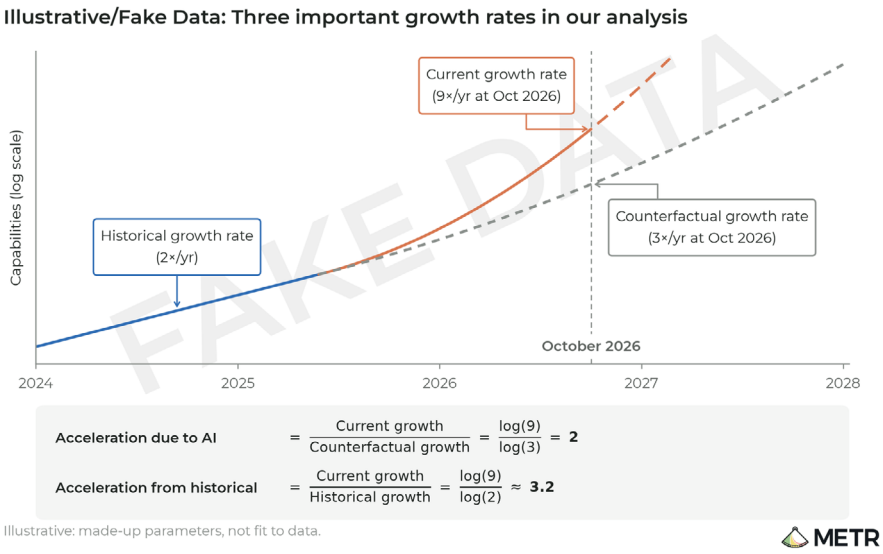
We typically estimate time trends by an ordinary least squares fit with a single estimated breakpoint; we then report acceleration as the ratio of the two estimated linear trends. We adopt the best-fitting breakpoint on standard AECI data as our default breakpoint on other graphs, to create a consistent breakpoint throughout this report. However, in some places we consider deviations from this, and test estimated acceleration as per the best fit on each graph.

We organize our analysis around 3 growth rates: the historical growth rate, the current growth rate and the counterfactual growth rate.

We want to assess the degree to which AI automation is speeding up AI progress. For any measure of progress (e.g. model capabilities as measured by AECI), we can estimate three numbers:

1. **Historical growth rate:** the growth rate prior to a detected trend change.
2. **Current growth rate:** the growth rate as of October 2026.
3. **Counterfactual growth rate:** the growth rate as of October 2026, if Anthropic could not use AI.

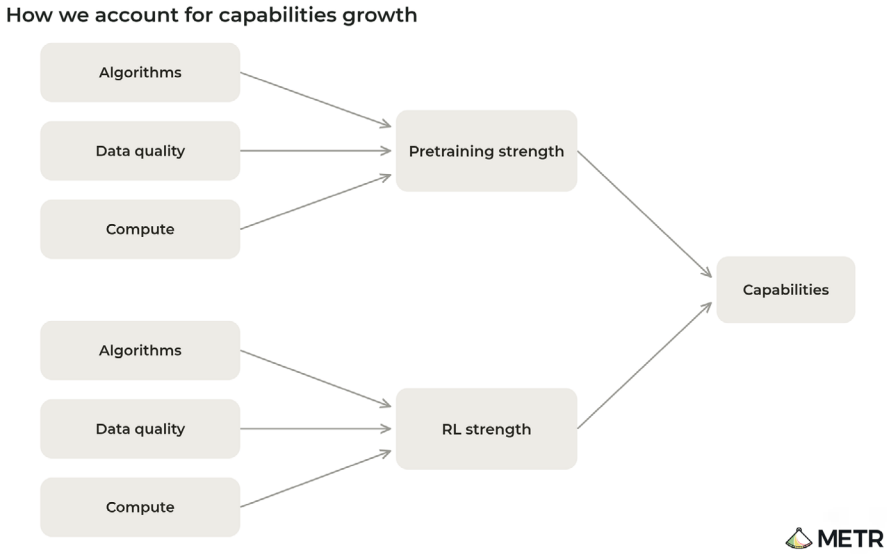
The first two are observable quantities. The counterfactual growth rate we can only guess at. The following plot illustrates these measures on illustrative data:



It is important to distinguish between instantaneous acceleration due to AI, as we have defined it, and *cumulative* acceleration due to AI, i.e. what fraction of overall progress since some point in time (e.g. October 2025) can be attributed to AI. Some interviews and documents are implicitly or explicitly discussing cumulative acceleration, though we generally expect cumulative acceleration to be lower than instantaneous acceleration, due to non-AI acceleratory factors in 2025, and due to the growing influence of AI over time.

In the Full Report, we organize indirect evidence with a simple accounting framework.

We want to understand if there has been an acceleration in overall capabilities due to AI R&D. To reach our subjective conclusions, we organized our evidence using the following accounting framework, but all relevant usage of it was redacted.



Arrows into a box indicate that box is a function of the preceding boxes, so for example, pretraining strength is a function of algorithms, data quality, and training compute. We use the term RL to include all contributions not included in pretraining.

This framework lets us start at the right (capabilities) and determine whether there has been an acceleration. Then we can go back one level to see if there has been an acceleration in pretraining or RL, and then go back one further level to see what inputs are causing that acceleration. This allows us to get a deeper understanding of what is happening to the inputs.

In the Full Report, analyses were all documented in an accompanying notebook and dashboard.

For reproducibility we maintained a notebook and dashboard with all results in the Full Report. Where feasible we included references to the notebook or dashboard, and most charts had a “stamp” which identified their provenance.

Benchmark scores (AECI)

AECI appendix overview.

We expand on the main AECI analysis in a few ways.

1. We explain the methodology behind our main estimation of a 1.9x acceleration in AECI.
2. We show the degree of estimated acceleration varies based on reasonable alternative analysis choices:
   - Varying the link function: 1.1x–2.9x.
   - Varying the snapshots used: 1.9x–2.2x.
   - Varying the benchmarks used: 1.9x–2.2x.
   - Varying the regularizer used: 1.9x–2.1x.
3. We explain why these results differ from AECI estimates in previous Anthropic system cards.
4. We go into detail on potential problems with AECI such as:
   1. Anthropic's internal benchmarks are increasingly saturated: Model 4 achieves greater than 50% on all the AECI benchmarks on which it was tested. However, measured acceleration appears reasonably robust to saturation concerns.
   2. Token spend may be too low to maximally elicit models.
   3. The logistic link function could be misspecified on a few benchmarks. However, on the whole we think it is a reasonable approximation.
   4. There are various reasons why an acceleration in benchmark scores could overstate or understate an acceleration in real-world impacts.

Redactions in this section.

This section was very minimally redacted by Anthropic. The only information redacted was:

1. The names of specific benchmarks.
2. One graph that shows AECI results for a subset of models chosen on the basis of information that is redacted.

Methodological details on AECI estimation.

We use Epoch's public code to estimate AECI values for each model. We generate 95% confidence intervals via bootstrap by drawing benchmark scores with replacement, stratified by model.

Our underlying data includes the AECI scores of many snapshots per model. These snapshots differ in both dates and AECI scores.

For publicly released models, we include the snapshot that was publicly released and the date that snapshot was available for use by most Anthropic employees. For example, if a model has two snapshots: one available to Anthropic employees on November 10 and one available to Anthropic employees on November 20, and the second snapshot was released publicly on December 1, then our procedure would assign November 20 as the release date. We think there are many reasonable alternative dating procedures (e.g. at first snapshot release instead of last snapshot), which are shown below in one of the robustness exercises.

For models that are never released, we include snapshots that received substantial internal usage and the date most Anthropic employees could use them.

We use the date at which the model could be used by most Anthropic employees instead of public release dates; this allows us to make apples-to-apples comparisons with unreleased models. These in general show slightly higher degrees of acceleration as Anthropic is holding onto more recent models for slightly longer before their public release.

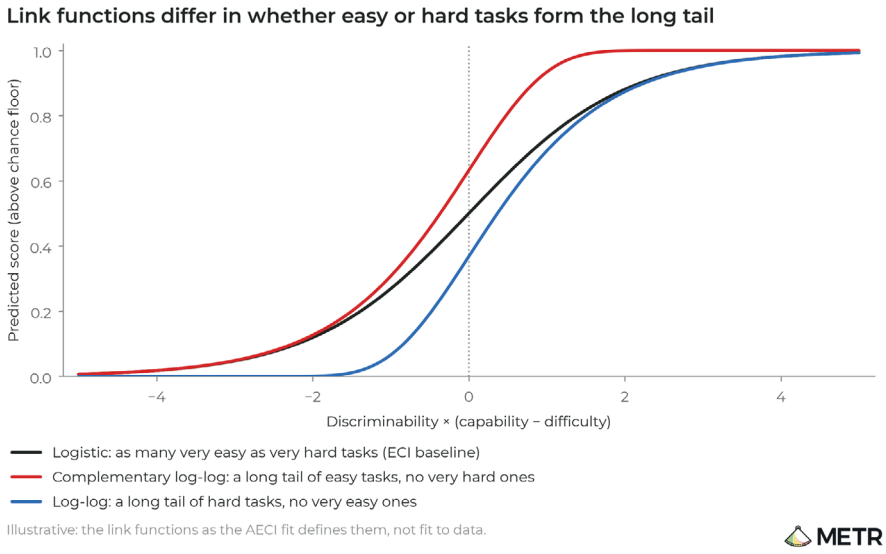
When we fit trends over time we use ordinary least squares with equal weight to each snapshot on the frontier. The confidence interval in the trend estimation only accounts for uncertainty in the AECI scores for each model, and does not include sampling uncertainty over which models are included in the fit.

We estimate a learned breakpoint in the trend, though we exogenously impose that there is 1 breakpoint. We estimate the breakpoint is October 7, 2025 in our main AECI specification. For our robustness checks, we maintain this October 7, 2025 breakpoint to ensure comparability.

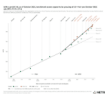
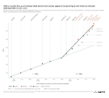
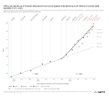
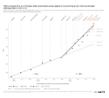
Robustness: Estimates of AECI acceleration vary 1.1x–2.9x depending on the link function.

We show below that estimates of AECI acceleration are somewhat sensitive to assumptions about the distribution of task difficulty within each benchmark. In principle, we could discriminate between these assumptions based on the data, but in practice most functions fit the data about equally well. We acknowledge useful suggestions from Anthropic staff on analysis here.

AECI assumes that a model's score on a benchmark is a function of the model's latent strength, and specifically that the function is logistic (shown below in black). This can be interpreted as an assumption on the distribution of difficulty of tasks in a benchmark, including how many tasks are broken or what fraction can be guessed. The following figure illustrates a few different link functions. The logistic function assumes a symmetric distribution with exponentially declining tails. The log-log function assumes a long tail of hard tasks; the complementary log-log function assumes a long tail of easy tasks.



The following table shows the estimated acceleration depending on various assumptions of the link function. Note all these link functions are global, i.e. not benchmark-specific.

Link function	Intuitive assumption on task distribution	Acceleration	Out-of-sample R^2	Figure
Standard logistic	Benchmarks have a symmetric difficulty distribution of problems with exponential tails.	1.9x [1.5–2.8]	87.2%	above
Probit	Benchmarks have a smaller tail of both easy and hard problems.	1.8x [1.3–2.7]	87.1%	
Cauchit	Benchmarks have a larger tail of both easy and hard problems.	2.8x [2.0–4.1]	86.1%	
Complementary log-log (heavy left tailed)	Benchmarks have a larger tail of only easy problems (for example, if there are many easily guessable tasks).	1.1x [0.8–1.5]	87.1%	
Log-log (heavy right tailed)	Benchmarks have a larger tail of only hard problems (for example, if there are a few extremely hard problems).	2.9x [2.1–4.9]	86.9%	
Learned link function (Stukel two-tail family, 2 globally learned parameters)	Length of both tails is globally estimated from data.	2.6x [1.8–4.1]	87.0%	

The quality of fit is fairly similar across each of these link functions (except cauchit), and we do not have strong a priori reasons to prefer one. We continue to use logistic for our main results since it is the default link function used by Epoch and Anthropic.

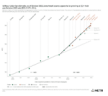
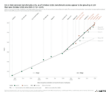
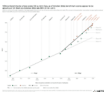
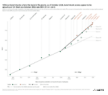
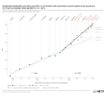
Robustness: An acceleration in AECI of around 2x appears robust to which snapshots are included.

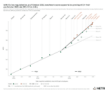
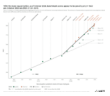
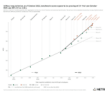
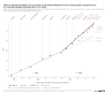
We consider various alternative choices for which models to use when estimating AECI growth:

Parameterization	Acceleration	Figure
Standard AECI (snapshots that got substantial internal usage)	1.9x [1.5–2.8]	above
Only the highest snapshot per model	1.9x [1.4–2.8]	
Only the first snapshot per model	2.2x [1.3–3.2]	
A subset of models (for redacted reasons)	1.9x [1.4–3.3]	[this graph was redacted by Anthropic]
All snapshots	1.9x [1.4–2.6]	

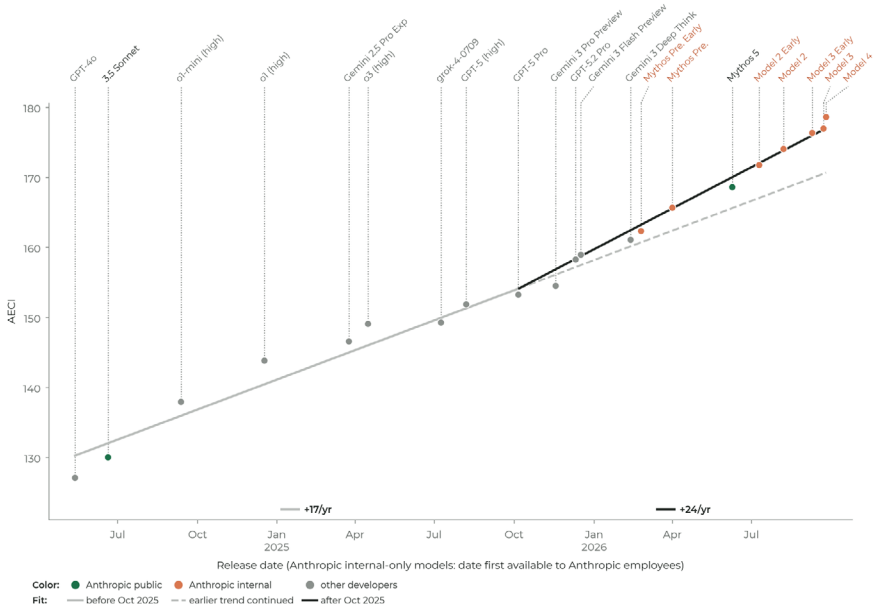
Robustness: An acceleration in AECI of around 2x appears robust to which benchmarks are included.

We consider various alternative choices for which benchmarks to use when estimating AECI growth:

Parameterization	Acceleration	Figure
Standard AECI	1.9x [1.5–2.8]	above
Drop vision benchmarks (e.g. OSWorld, MMMU).	2.2x [1.7–3.5]	
Keep only benchmarks that Claude labels as relevant to AI R&D (e.g. drop multilingual MMLU, honesty benchmarks).	2.0x [1.5–2.9]	
Drop benchmarks that on average decrease as AECI increases. (This may indicate a broken benchmark as more generally capable models are scoring lower.)	1.9x [1.6–2.8]	
Drop benchmarks where the logistic fit has lower than 50% R^2. (This drops benchmarks where the logistic link function is a bad fit.)	1.9x [1.4–2.6]	
Stop using benchmarks after they hit 85% (for saturation robustness).	1.9x [1.4–2.3]	

	Parameterization	Acceleration	Figure
Robustness: An acceleration in AECI of around 2x appears robust to regularization.	Standard AECI	1.9x [1.5–2.8]	above
	Decrease L2 regularization by 10x.	1.9x [1.4–2.8]	
	Increase L2 regularization by 10x.	2.1x [1.6–3.0]	
	Remove regularization entirely.	1.9x [1.4–2.8]	
	Re-center the L2 penalty on a linear time trend instead of towards 0: each model's capability score and benchmark's difficulty is pulled towards a line in its release date. Each benchmark's discriminability is pulled towards 1.	1.9x [1.4–2.8]	
Robustness: AECI shows less acceleration when we include non-Anthropic models.	If we take our main AECI results and plot all frontier models instead of just Anthropic models, we estimate a lower degree of acceleration.		

With other developers' public models included, the frontier's benchmark scores appear to be growing at 1.4× their pre-October 2025 rate



The running record across developers; before Oct 2025 other developers hold it almost throughout, after it Anthropic's models (other developers' internal models are not visible). Least squares with the slope changing at October 7, 2025. METR

Notice the historical trend is estimated mostly from non-Anthropic models, while the current trend is estimated from mostly Anthropic models.

There are a couple of plausible interpretations of this data:

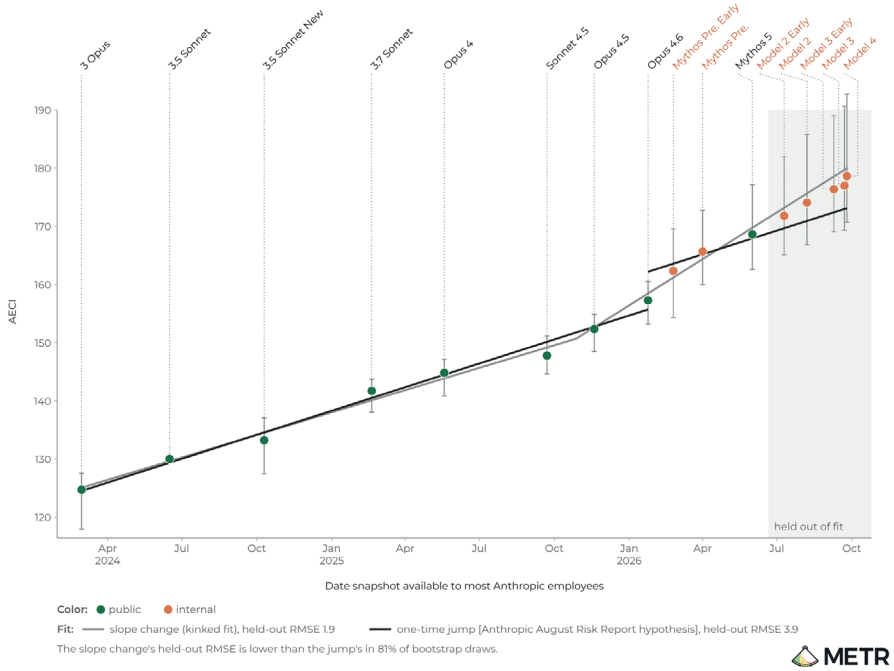
1. Anthropic models were subfrontier in 2025, so Anthropic's historical progress up to 2025 was understated. Therefore, the 1.9x acceleration numbers in the rest of the report are overstated, and 1.4x is more accurate.
2. If Anthropic models are less benchmark-optimized compared to other models, this could mean the historical trend is overstated as the earlier points are more overfit to benchmarks. Therefore, the 1.4x acceleration in the graph above is understated and the 1.9x we quote elsewhere is closer to accurate.

New internal models make a permanent slope change look more plausible than a one-time jump at Mythos Preview.

Prior Anthropic system cards have posited the hypothesis that there was a one-time jump in capabilities at Mythos Preview after which there was a return to the previous trend. Recent models make this hypothesis look less likely than a permanent trend change.

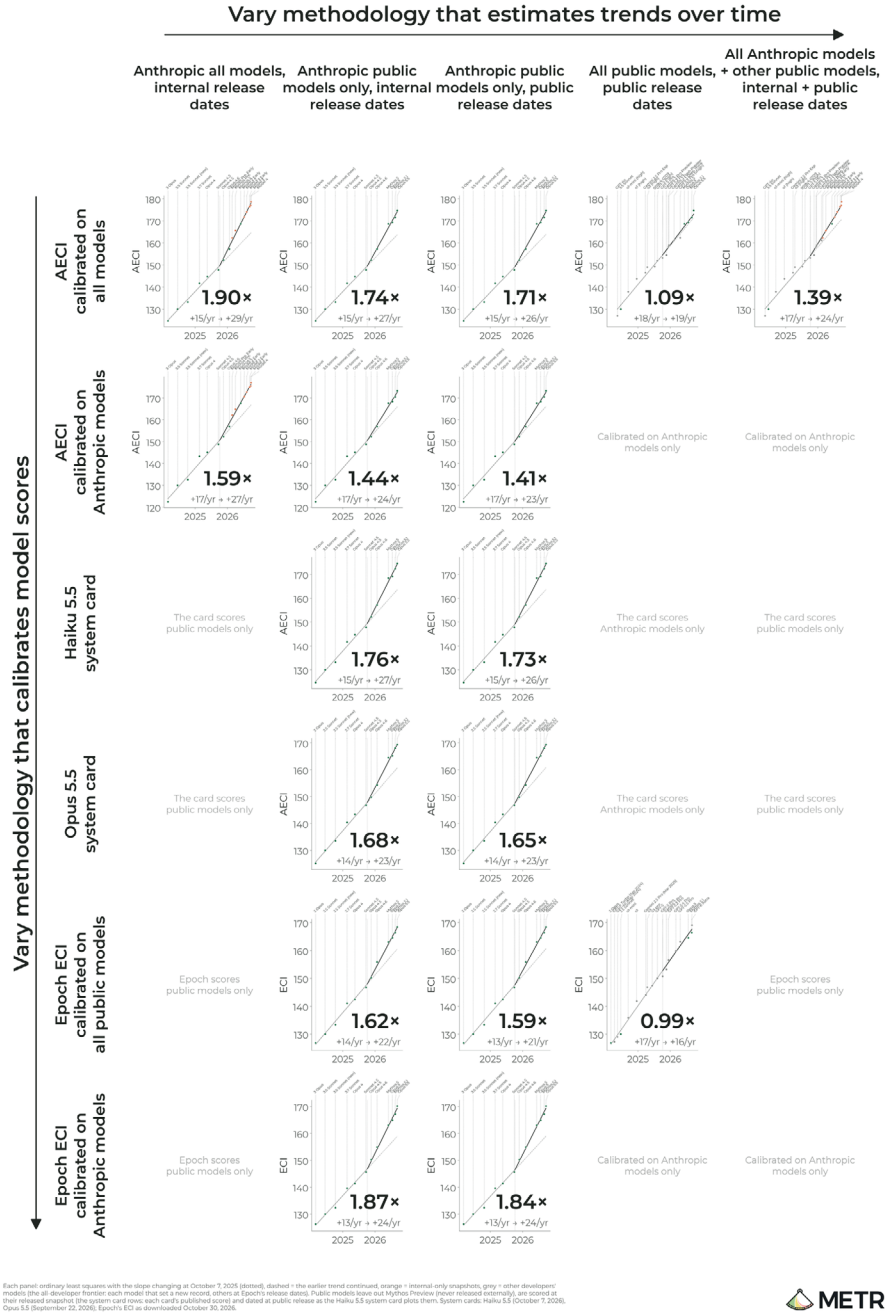
Fit only through Mythos 5, a permanent acceleration predicts the later models better than a one-time jump (held-out error 1.9 vs 3.9 AECI)

AECI is a metric of model capability that aggregates scores on a mix of internal and external benchmarks. The amount of acceleration is sensitive to assumptions about benchmark difficulty and how trends over time are estimated.



Comparing our AECI estimates to various public estimates.

The following chart compares our acceleration estimate to various public estimates of acceleration.



The rows denote different methodologies for calibrating (A) ECI scores for each model, e.g. by changing benchmarks, which models are included in the score calibration, etc. The columns denote different methodologies for converting those scores into an acceleration over time, e.g. by subsetting to certain models, changing the release dates, etc.

Our main takeaways are that:

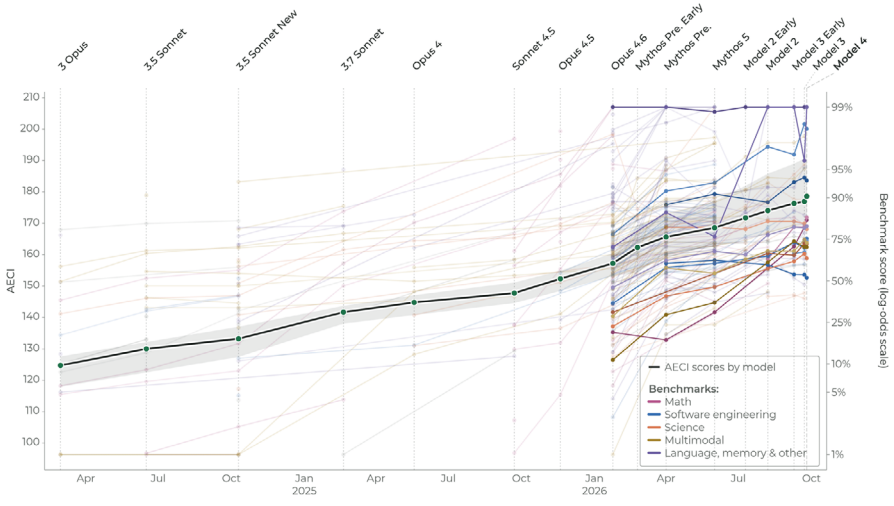
1. The most important difference between the latest Anthropic model release, the Haiku system card (1.7x acceleration), and our estimate (1.9x acceleration) is the inclusion of Anthropic's recent internal-only models.
2. The most important differences between ECI and our estimates are subsetting the trend over time to only Anthropic models and the inclusion of Anthropic's recent internal-only models.

Model 4 achieves >50% on all of its benchmarks.

The plot below shows all benchmark scores for frontier models in the AECI, scaled on a log-odds scale, implying that they should evolve linearly with a model's latent AECI score (if the logistic assumption holds). The highlighted lines represent the 15 benchmarks used for Model 4. Notably, Model 4 scores above 50% on all these benchmarks. We can see that the distribution of benchmark difficulty was very different just 9 months ago: Opus 4.6 had many benchmarks on which it scored below 50%. This lack of benchmarks which are difficult for Model 4 makes us somewhat nervous about our ability to measure AECI now and in the future.

AECI scores plotted over time (left axis), and individual benchmark scores (right axis, log odds)

This graph plots both the log odds of each benchmark (right-hand vertical axis) and the overall AECI score (left-hand vertical axis) over time. As a consequence of AECI's logistic link function assumption, AECI scores should be linear in benchmark log odds.



Technical details: full-color lines are the benchmarks Model 4 was evaluated on, faded lines the rest. The shaded band is the AECI frontier's 95% interval from 100 bootstrap draws over benchmarks stratified by model.

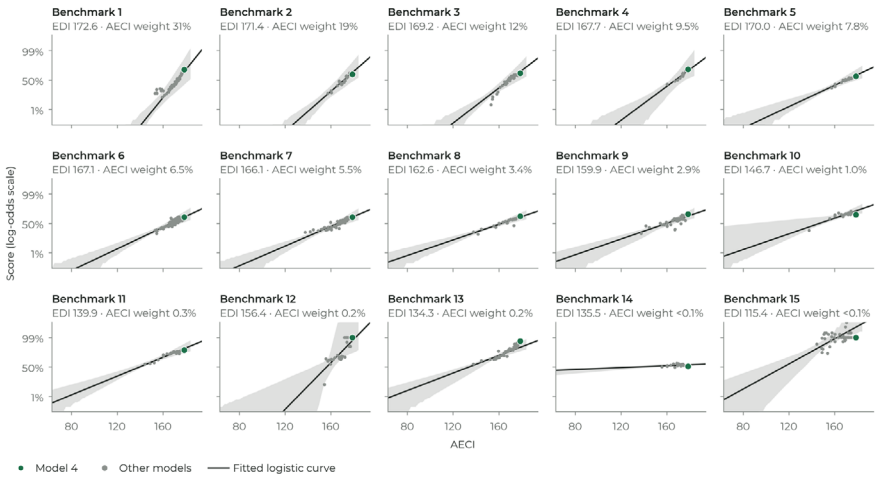
METR

Most of the benchmarks used for Model 4 appear reasonably linear in log-odds space.

A specific concern in measuring acceleration with AECI is benchmarks with many broken tasks, or benchmarks with a fat-tailed distribution of task difficulty. The logistic link function assumes exponential tails of task difficulty; if we instead had a fat-tailed distribution of task difficulty, this would mean that going from 98% to 99% requires more AECI points than going from 80% to 90% (the logistic assumes they are approximately equally difficult). A fat-tailed distribution of task difficulty could cause us to underestimate acceleration; as described above, using cauchit or log-log link functions (which model the last few tasks in a benchmark as particularly difficult) results in acceleration estimates of 2.8x and 2.9x, respectively.

We perform a simple diagnostic test for task difficulty. For each of the 15 benchmarks used for Model 4 in AECI, we compared a model's estimated AECI against the model's log-odds score on that benchmark. If the AECI assumptions are satisfied, meaning task difficulty is logistic in a model's latent ability, then we would expect each plot to be linear. If there are many broken tasks or a fat-tailed distribution of difficulty, then we would expect these fits to be *concave* (we would expect a convex fit if there are disproportionately many easy tasks).

Most of Model 4's benchmarks look linear in log odds, as the logistic link assumes



Shaded bands: 95% intervals. AECI weight: roughly how much each benchmark counts toward Model 4's AECI.
EDI: estimated difficulty of benchmark.

METR

We noticed there are 3 of 15 benchmarks with notably flat or concave fits. When we re-estimate AECI for all models without those 3 suspicious benchmarks, Model 4's score increases by 1 point, from 178.6 to 179.7. We thus think there is some weak evidence for fat tails causing understatement in acceleration, but the magnitude is not large.

Tokens per benchmark may be too low.

Using a low expenditure in benchmarking could bias estimates of capability growth downwards if later models show higher returns to expenditure (i.e. if early models plateau with expenditure, but later models show high returns).

Anthropic shared with us data on tokens used on benchmarks in the AECI. A typical task (one attempt) used about 2.5M input and 72k output tokens across internal benchmarks, and about 10M input and 94k output on agentic coding benchmarks. For a loose comparison, Epoch AI uses roughly 0.5M input and 10k output tokens on a typical task across a few public benchmarks (GPQA, AIME, SWE-bench), and 1M input and 8k output tokens on a specific agentic coding benchmark (SWE-bench). For another loose comparison, METR uses roughly 650k output tokens for its latest long-running software benchmarks, though it is possible the token accounting is not comparable.

Moreover, it would be informative to see returns-to-expenditure curves for each model, i.e. scores elicited at different levels of expenditure. Anthropic's system cards do report expenditure curves for some benchmarks (e.g. Opus 5.5 system card). There is some recent evidence that it is possible to use very large amounts of inference compute to solve difficult problems (e.g. Navier–Stokes). Overall, our guess is that the AECI scores for each model could be increased by a handful of points if inference tokens were scaled further. Our rough guess is that this would have similar returns across models, so the degree of estimated acceleration would be the same, though this is not something we looked into deeply and it is possible that recent models have better token scaling properties.

An acceleration in benchmark scores may fail to track acceleration in real-world impact.

Ideally we want to know the real-world value, or potential impact, of each model, not its benchmark scores. If real-world impact is smoothly related to benchmark scores, then we should expect an acceleration in AECI to be associated with an equivalent acceleration in real-world effects.[5]

There are a number of reasons why acceleration in AECI may fail to track acceleration in the real-world impact of model capabilities:

1. If real-world impact is a convex function of benchmark scores, e.g. if impact depends on crossing a capability threshold, then AECI would understate real-world acceleration. In some domains we have seen a steady increase in benchmark scores accompanied by sudden accelerations in real-world impacts, e.g. impacts on the rate of discoveries in cybersecurity or math (ref).
2. If the benchmarks in AECI cover only a subset of capabilities, and the real-world impact has diminishing returns in those capabilities, then AECI could overstate real-world acceleration. For example, suppose AECI primarily measures a model's ability to solve formal problems, but there are diminishing real-world returns in that ability; then we might observe an acceleration in AECI without an acceleration in real-world impact.

Overall, we are confident there has been at least a modest acceleration in the growth of benchmark scores. We believe there is significant risk that acceleration in AECI may not map cleanly to acceleration in real-world impact.

Evidence of AI speedup

There is considerable evidence of AI speedup.

We define AI speedup as the ratio between a staff member or team's output with AI and without AI, noting that output can often be difficult to measure or define.

We now briefly survey the overall evidence for AI speedup, concluding:

1. There is some evidence for high speedup, though the precise amount of speedup is highly uncertain.
2. Fully autonomous contributions to AI R&D are not currently a large contributor to overall progress, but they may be in the near future.

Redactions in this section.

Anthropic redacted evidence of speedup that was closely connected to specific parts of their model training pipeline, as well as our discussion of how the evidence in this section related to our estimates of acceleration in various parts of the model training pipeline. Moreover, a claim about inference compute usage at Anthropic was redacted because it was priced in compute terms and we did not find it informative enough to request it in API prices. They also redacted sensitive details from the quotes from researchers below.

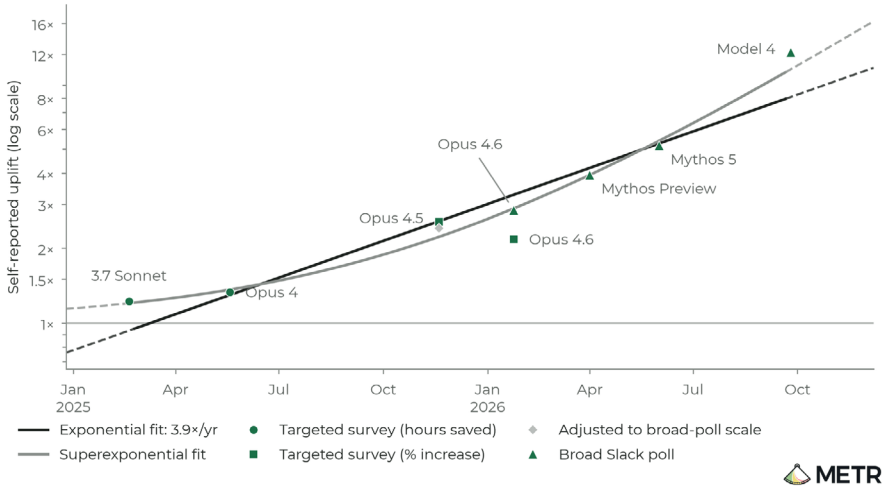
Informal self-reports of uplift have been steadily growing.

Anthropic shared data on self-reported uplift due to AI from a variety of sources. The most recent data is from a series of Slack polls, where staff use reactions to vote on their estimated speedup from the most recent model at the time. We consider this to be only very loose evidence: we expect the responses come from an unrepresentative subset of staff and the responses reflect fairly low-effort consideration. Additionally, there is some evidence that self-report survey estimates of uplift can overestimate true productivity impact.

Nevertheless, we observe consistent growth in these self-reports, shown in the chart below. For the most capable internal model in October 2026, average self-reported uplift was ≈12x.

Informal self-reported uplift over time from Anthropic employees

Results come from informal surveys and Slack polls. They should be taken as very loose evidence.



These self-reported estimates have grown but it is difficult to know how to extrapolate them. The plot above fits uplift against time, using both a log-linear fit and a superexponential fit.

Interviews broadly support AI doubling the rate of growth.

Both METR and Anthropic interviewed many researchers (see details in the following section). We used Claude to classify the quotes according to whether they support 2x or more acceleration from AI in their area of work, relative to working without AI. We interpret the results as mildly in favor of 2x or greater acceleration:

- Yes: 2
- Leans yes: 3
- Borderline/unclear: 4
- Leans no: 0
- No: 2

The quotes, with sensitive details redacted, are in the following section.

There are signs of life for fully autonomous contributions in some parts of the pipeline.

In addition to the uplift of human labor, we are also interested in autonomous AI R&D. We understand that AI is not yet able to make highly significant and fully autonomous contributions to the most important parts of the model training process. However, researchers described significant "signs of life" implying such contributions could occur soon, although with substantial variation depending on the area. Further details were redacted. This evidence came from:

1. Multiple interviews from researchers working to automate different parts of the training pipeline.
2. Internal reports on the degree of automation in different parts of the pipeline.
3. Quantitative analysis on the percent of tasks in a relevant pipeline that were built by humans or AIs.

Interview quotes on AI acceleration

Anthropic and METR staff conducted interviews with a number of researchers about AI R&D acceleration in their work. Interviews vary in length and depth, but we expect researchers spent significantly more time on these than answering the surveys above.

To classify interviews, we had Claude look over all our notes and take all relevant quotes and code them as supporting >2x acceleration due to AI in their area of work, ambiguous or against. Each row in the following table corresponds to a single person and all their relevant quotes. Note that each interviewee often interpreted the

question in somewhat different ways, e.g. whether they are referring to cumulative or instantaneous speedup, or whether they are referring to personal or org-wide speedup.

We used the following prompt:

> "Can you please look through all the interview tabs and all relevant interview docs outside of this google doc. [redacted context]. Can you make a table where each row is a person and then we put all their quotes together. Then please add a column which is your overall determination if this person supports a > 2x acceleration due to Claude in their work."

Person	Quotes	Supports >2x acceleration in their area of work from Claude?
Researcher A	*"Could definitely endorse a greater than 2X speedup"* — METR interview notes *"My inside view is by far the largest effect is model acceleration"* — METR interview notes *"The AI acceleration is larger than all the other effects."* — METR interview notes *"Now only bottlenecked on ideas & compute, no longer bottlenecked on code."* — METR interview notes *"I would definitely endorse >2 years of progress in 1 year and I don't think 4 is crazy."* — METR interview notes *"I think there's some truth to the 'we just started trying harder' story, but IMO the bulk of the effect is recent models."* — METR interview notes *"Like 80% reduction in time between idea and experiment, other 20% some mix of making fewer mistakes and being able to measure fiddly things that would've required too much one-off work before."* — METR interview notes *"I think there was a medium bump when it was good enough to code at all and a big bump with Opus-4.5/Opus 4.6 and an even bigger bump with mythos-preview"* — Anthropic interview notes	**Yes**

Researcher B	*"It is hard to see how I am not being 3x more productive. I feel like I have postdocs and grad students, and we are working towards a larger project. I think the simplest explanation is Claude—just from my individual perspective that feels sensible."* — METR interview notes	**Yes**
Researcher C	*"this is probably the biggest factor imo. When people present experiments, they most often just point to an [agent scaffold] thread. Experimentation overhead is driven towards zero super fast."* — Anthropic interview notes *"I think CC was already good enough before mythos-preview to implement most ideas around Opus 4.5"* — Anthropic interview notes	**Leans yes**
Researcher D	*"It is just much easier to [perform] lots of [redacted] experiments that would have taken days. People don't try [redacted] things bc it takes 1-2 days to implement per idea, but now its ~free with mythos-preview"* — Anthropic interview notes *"I personally think this is a large lift and has led to me trying far more ideas."* — Anthropic interview notes *"On a linear scale, ~0 uplift pre Opus 4.6/Opus-4.5, ~some uplift that was small with them, large uplift with mythos-preview + [agent-scaffold]. I would guess we're just starting to fully switch over … and we'll see a continued slope increase / inflection."* — Anthropic interview notes	**Leans yes**
Researcher E	*"I do genuinely think Claude (claude code, [agent-scaffold]) has been fairly significant. I've seen us discovering very promising directions ~entirely because people can spam 10 ideas at [agent-scaffold] and see what sticks."* — Anthropic interview notes *"A lot of [redacted] research work is just implementation, babysitting, etc. Opus-4.5 in claude code is already pretty helpful with this, even if it mostly can't do end to end."* — Anthropic interview notes	**Leans yes**

Researcher F	*"The slope change is 50-80% Claude"* — Anthropic interview notes *"Change is especially notable for this manager, who can de-risk ideas with Claude enough to get someone else excited, and would not have been able to do that at all as a manager of 10 before."* — Anthropic interview notes (paraphrase)	**Borderline**
Researcher G	*"Inflection point matches his vibes as to what's been happening. Two biggest contributors here are Claude automation and hiring."* — Anthropic interview notes (paraphrase) *"Claude lets researchers try many things quickly, most of which won't work. …"* — Anthropic interview notes	**Unclear**
Researcher H	*"I think it's true [that] we're now trying a lot more experiments at the same time because of claudification."* — Anthropic interview notes	**Unclear**
Researcher J	*"Before Opus-4.5 → Claude not useful for autonomous code / Opus-4.5 → first time useful for autonomous code / mythos-preview + [agent-scaffold] → first time useful for autonomous experiment"* — Anthropic interview notes	**Unclear**
Researcher K	*"IMO [hiring is] probably the biggest reason …"* — Anthropic interview notes	**No**
Researcher L	*"Guesses 15-20% is due to AI. Don't believe in self-reported productivity increase."* — METR interview *"My favored interpretation: [redacted non-AI factor]"* — METR interview *"(1) opus45 couldn't write code; (2) opus46 could write a little; (3) Mythos Preview a sharp inflection; (4) Model 2 is another really large improvement in how useful these things are."* — METR interview *"definitely the case that we're accelerated by claude .. Even though this rate of progress by end of year will be even faster … my attribution [mostly not to Claude] and uncertainty remain."* — METR interview *"One impact is SWE acceleration, which is extremely substantial. We're definitely benefiting from that differentially between 2025 and 2026."* — METR interview	**No**

Adversarial robustness

We believe this exercise was reasonably adversarially robust.

As this is an independent assessment, it is important for us to do our due diligence and understand how hard it would have been for Anthropic to systematically mislead us, i.e., how likely it is that we could have been led to an incorrect conclusion intentionally.

Adversarial robustness weakness 1: We relied on data from a few Anthropic employees. Most of our data was received through two technical staff members; however, we interviewed around 15 other Anthropic employees on a range of topics. We did not notice any discrepancies between what our primary contacts said and what our interviewees said. Given that many of these Anthropic employees were chosen by us, not Anthropic, we think this would have been challenging to pull off if we were being given purposefully misleading information.

Adversarial robustness weakness 2: Most analysis relied on Claude. Claude was involved in many stages of this project, and so in principle could be distorting our perception of the data. We believe Claude was deeply involved in:

- the majority of data collection and analysis that we received,
- writing a significant share of the documents we received,
- the majority of our own analysis.

We did not detect any significant discrepancies consistent with a systematic bias even when checking against human interviews. We suspect this is not a significant risk.

Overall, we think it is extremely unlikely Anthropic was actively trying to mislead us. However, we believe greater access would allow us to increase the degree of adversarial robustness.

Process transparency disclosures

COI disclosure.

We applied v1.0 of our conflict of interest policy when selecting the team for this project, which was the live policy at the start of this project.

At the start of the project, one participating team member was in discussions with Anthropic about a possible role. They received an offer and declined it. This team member also had equity in a competitor as an ex-employee. As such, we required that all of their work incorporated into the assessment was thoroughly checked[6] by an unconflicted team member who took personal responsibility for its accuracy. Going forward, our updated COI policy, applied to risk assessments starting after October 9, 2026, would simply require recusal in such a situation.

No one working on this project held equity in Anthropic. Our COI policy requires recusal in the case that a relevant staff member holds equity. (None of METR's staff hold Anthropic equity, but one incoming staff member does.)

We believe there are no other conflicts that would cause a reasonable person not affiliated with METR to have materially less trust in this assessment.

Other disclosures.

Our process transparency policies require us to describe both the formal constraints (e.g. redaction and publication rights), and any other factors that may have affected our outputs. To mitigate informal pressures, we generally aim to share enough of the relevant evidence for the public to arrive at its own conclusions, but this was made more difficult in this case due to redactions. In order of impact, from highest to lowest:

Considering impact on negotiations with other companies and possibility of Anthropic retracting access. These assessments with companies are voluntary and we considered how this output would affect our future ability to do these assessments — for example, if we are too critical of Anthropic for access limitations or making particular redactions, this might make other companies or Anthropic itself less willing to work with us in the future. The biggest effects here are:

- We want to avoid disincentivizing companies from sharing relevant evidence with us that they then redact the details of, so we try to focus discussion of limitations from redactions on cases where they impacted our ability to communicate our most core conclusions and uncertainties.
- We avoid more detailed process descriptions that would disincentivize companies from doing exploratory collaborations or would make them feel that they need to have everything locked down and avoid changing their mind or discovering new information while working with us. We instead focus on communicating the aspects of the process that ended up having significant effects on our final output.

We made significant clarity edits, particularly to the Risk Report Review, in response to feedback. Our Risk Report Review went through several drafts, partially in response to significant feedback from Anthropic about communication clarity and reducing the surface area for adversarial misreadings of our report. In some cases we felt this feedback toward METR staff was overly forceful, but Anthropic was receptive to our feedback on this point and we hope to develop more detailed processes and guidelines for this in the future. We do not think the final assessment or the Risk Report Review is significantly distorted by this.

Summarized contract terms.

We summarize relevant contract terms below.

Scope

Scope. METR independently assesses how much AI systems are automating or accelerating AI R&D at Anthropic.

Access rights

Anthropic obligations. Anthropic provides the access listed below as reasonably needed, a designated point of contact, and reasonable help arranging surveys and interviews. Anthropic can decline requests that would take disproportionate effort or that fail its information review, but will explain why, describe in general terms what was withheld, and offer alternatives where practical.

Access granted to researchers. The contract describes a variety of information that we had access to in previous engagements as a minimum bar for access in this assessment. It also describes logistical details about conversations with Anthropic staff. We describe the access we ended up getting above.

Information review process. Anthropic aims to decide on information requests within three business days and provides a point of contact for questions.

Publication rights

Publicity. Neither party makes public statements about the agreement without the other's approval, except that both can disclose that the engagement exists and its general nature (an independent AI R&D and compute-allocation assessment).

METR publication rights. METR has the following publication rights:

1. **Approved public claims.** METR submits proposed public findings to Anthropic for review before publication. Anthropic may require redaction of confidential or non-public information, and the parties discuss redactions in good faith, with Anthropic's redaction requests controlling. METR may submit new or revised claims for clearance at any time.
2. Without further review, METR may publish approved public claims, publicly available information (including previously approved claims, e.g. from the Frontier Risk Report), information derivable from a model or document once it becomes public, and METR's own analysis or commentary on those.
3. Anthropic can ask METR for a short delay of any public statement, subject to limited exceptions.
4. **Transparency.** METR can always disclose:
    - the existence and general terms of the review and redaction process
    - whether Anthropic exercised redaction rights

- whether any findings depended significantly on redacted information
- the level of access METR received
- the confidentiality framework.

Deliverables

- A full report: unredacted, delivered to Anthropic, covering the full project scope.
- A METR-internal report: the full report after light redactions by Anthropic, for sharing with METR staff. [Note: this report was not created during the process for timing reasons.]
- Public claims: the public report, cleared through the redaction process described under Approved public claims.

Independence and conflicts of interest

Independence. METR conducts the project and reaches its conclusions independently and does not enter arrangements that would make it financially, operationally, or managerially dependent on Anthropic. METR minimizes conflicts of interest and discloses any that a reasonable person in the industry would consider relevant in public or regulator-facing publications. METR can still evaluate other AI providers. Anthropic will not seek to influence METR's findings in ways that undermine the project's integrity, and METR retains sole discretion over methodology and conclusions.

No payment. No fees are paid in either direction. Tokens and credits provided to METR assessors are free, and each party covers its own costs.

AEF-1 checklist

Below we reproduce the AEF-1v1 checklist from the AI Evaluator Forum. For comparison see METR's AEF checklist for our Frontier Risk Report (p. 67).

1: Secure sufficient access and resources

Item	Description	Fulfilled?	Notes/evidence
1.1	Requirement: The evaluator secured sufficient technical access to assess the specific system characteristics being evaluated.	**No**	Of the sub-items below, AEF-1 requires only 1.1.1 (query access). METR's staff did not have query access to Anthropic's unreleased models, so we mark this item as not fulfilled, though such access was not central to our evaluation. This section of the checklist does not apply well to this assessment.
1.1.1	The evaluator had query access.	**No**	METR's staff on this assessment did not request query access as we did not see it as critical for this assessment.
1.1.2	The evaluator had access to the system's scaffolding.	**N/A**	
1.1.3	The evaluator had exemptions from system safeguards.	**N/A**	
1.1.4	The evaluator had access to intermediate system states.	**N/A**	
1.1.5	The evaluator had finetuning access.	**N/A**	
1.1.6	The evaluator had access to model weights.	**N/A**	
1.1.7	The evaluator was granted access to other tools for elicitation or otherwise supporting external validity.	**N/A**	
1.1.8	The evaluator had access to relevant user data.	**N/A**	
1.2	Recommendation: The system provider shared with the evaluator information relevant to carrying out a trustworthy and useful evaluation.	**Yes**	Anthropic shared information with METR on a case-by-case basis as described above.

1.3	Requirement: The evaluator had access to sufficient computational resources to complete a thorough evaluation.	**Yes**	This assessment did not require large compute resources from METR.
1.4	Requirement: The evaluator had adequate time to carry out a thorough evaluation.	**Yes**	The engagement was extended from 4 weeks to 11 weeks (August 12 – October 29). We describe more about the timeline in the process details section.
1.5	Recommendation: The system provider provided legal safe harbor for actions by the evaluator that are within the agreed upon scope of the evaluation.	Not requested	

2: Minimized conflicts of interest

Item	Description	Fulfilled?	Notes/evidence
2.1	Requirement: The evaluator did not receive compensation contingent on the results of the evaluation.	**Yes**	METR did not request or receive compensation for this assessment. Anthropic provided laptops and token credits for the assessment.
2.2	Requirement: The system provider did not exercise organizational or financial control over the evaluator.	**Yes**	
2.3	Requirement: The evaluator has published a conflict of interest policy, and it was applied to the evaluation.	**Yes**	We applied v1.0 of our conflict of interest policy when selecting the team for this project, which was the live policy at the start of this project. Since then, we have updated our COI policy.
2.4	Requirement: The evaluator clearly disclosed conflicts of interest relevant to the evaluation.	**Yes**	
2.4.1	Was the evaluator paid by the system provider or its direct competitor to conduct the evaluation?	**No**	METR's staff on this assessment did not request query access as we did not see it as critical for this assessment.

2.4.2	Does a meaningful fraction of the evaluator's funding come from the system provider, its employees, or its direct competitors?	**No**	
2.4.3	Do the system provider, its employees, or its direct competitors own equity in the evaluator organization?	**No**	
2.4.4	Do evaluator staff who carried out the evaluation own equity in the system provider or its direct competitors?	**Yes**	See discussion of COI and mitigations in the COI disclosure.
2.4.5	Do any evaluation staff working on the evaluation simultaneously work for the system provider or its direct competitors?	**No**	
2.4.6	Did the evaluator have any other conflicts of interest relevant to the evaluation?	**Yes**	See discussion of COI and mitigations in the COI disclosure.
2.5	Requirement: The evaluator recused any individuals with a significant financial interest in the system provider from carrying out the evaluation.	**Yes**	
2.6	Recommendation: The evaluator disclosed any separate agreements with the system provider that significantly impact the independence and trustworthiness of the evaluation.	**Yes**	METR has other independent evaluation arrangements with Anthropic. Note also that METR's work with nonpublic models and use of free tokens incentivize a cordial relationship with AI companies. See our website.

3: Analytic autonomy

Item	Description	Fulfilled?	Notes/evidence
3.1	Recommendation: The evaluator retained flexibility to define which specific properties of the system to evaluate.	**Yes**	METR had full flexibility on the nature of the investigation, given the scope defined in the contract.
3.2	Requirement: The evaluator had autonomy in deciding the methods of the evaluation.	**Yes**	
3.3	Recommendation: The evaluator ran the evaluations themselves via direct system access.	**No**	METR's staff on this assessment never had direct access to Anthropic systems, but were given access to many individual datasets and documents.
3.4	Requirement: The evaluator retained editorial control over how they present the results of their evaluation.	**Yes**	Anthropic did not have editorial control over the final report, but did have control over what non-public information METR could incorporate into it. See process details and process transparency disclosures.

4: Transparent methods and results

Item	Description	Fulfilled?	Notes/evidence
4.1	Requirement: The evaluator shared sufficient methodological details to allow for independent review of the results.	**No**	This was not possible due to redaction requirements.
4.2	Recommendation: The system provider granted any necessary rights to disclose results to the evaluation's intended audiences upfront.	**No**	As part of this assessment, Anthropic had rights to redact any non-public information in this post. Per our approach to transparency, we had the right to disclose if Anthropic exercised redaction rights, and whether or not any of our findings depended significantly on redacted information.
4.3	Recommendation: The intended audiences for an evaluation were not narrowed based on the results.	**Yes**	

4.4	Recommendation: The system provider did not misrepresent the evaluation's findings.	**Yes**	As of the extended deadline (November 3).
4.5	Recommendation: The system provider or other external parties did not unduly delay the evaluation from being disclosed based on the content of the results.	**Yes**	See discussion in the process transparency disclosures section.
4.6	Requirement: The system provider did not have authority to redact results to conceal concerning findings.	**No**	Anthropic had unilateral redaction rights.
4.7	Recommendation: The evaluator clearly disclosed the redaction authorities granted to the system provider or other external entities.	**Yes**	

5: Protection of sensitive information

Item	Description	Fulfilled?	Notes/evidence
5.1	Requirement: The evaluator had the system provider's permission to release any results based on non-public information or systems.	**Yes**	All the information in METR's public report was affirmatively approved for release by Anthropic.
5.2	Recommendation: The evaluation methods were not gamed or leaked.	**N/A**	Less directly relevant to our automated AI R&D assessment. Some discussion above on ways we could have been misled.
5.3	Requirement: The evaluator implemented measures to protect any confidential information they received from the system provider.	**Yes**	Assessment was conducted on company devices, and METR assessors followed our confidentiality policy.
5.4	Requirement: The evaluator established and followed a responsible disclosure policy.	**No**	METR has not established a responsible disclosure policy.

Update log

November 6, 2026

We made the following changes to the AEF-1 checklist:

- Removed leftover template notes from items 1.1 and 2.4.
- Corrected item 1.1 from "Yes" to "No", and amended its notes accordingly.
- Indented the sub-item rows.

Footnotes

1. We agree that Anthropic's redactions were limited to confidential information that could potentially accelerate other AI developers, though most of the redacted information was very high-level. We do not make a comment on whether these redactions were justified overall, since we do not think METR is well positioned to weigh a company's business interest in preserving this confidential information against the public's interest in better understanding the risks from AI development. Instead, we focus on describing the nature of these redactions and how they affected our conclusions.
2. Authors: Tom Cunningham, Nate Rush, Parker Whitfill and Hjalmar Wijk.
3. METR has published multiple lines of research relevant to AI R&D acceleration, with results pointing in both directions: our time-horizon trends suggest rapid progress, while our 2025 developer productivity study found early-2025 AI tools caused a slowdown, our estimates of future compute bottlenecks point to limits on scaling, and our analysis of benchmarks overstating AI progress cautions against reading too much into headline scores.
4. Four primary authors worked on this report. Two further METR employees spent time supporting and overseeing this work.
5. A separate problem is benchmarks intended to have an interpretable outcome, e.g. if SWE-bench tasks fail to fit the description of "resolving github issues." We are using AECI as a black-box index of overall capability, so are not affected by these problems of benchmark description.
6. Per our COI policy, this means following at least the usual quality of review of METR research, entailing understanding the data and analysis well enough to be able to understand the validity of the conclusions, and including consistency checks, understanding the provenance of statistics, and reading through some of the underlying data.