Forecasting Research Institute

How Close Is Recursive Self-Improvement?

A review of the evidence

Authors: Simas Kučinskas, Nick Merrill, and Josh Rosenberg

VERSION: v2 (2026 NOV)

FORECASTING RESEARCH INSTITUTE

Table of Contents

Summary	03
1. Introduction	03
2. Defining recursive self-improvement	05
3. Evidence	11
4. Existing projections and forecasts	20
5. Key cruxes	27
6. Forecasting questions	29
7. Takeaways	33
Changelog	34
Appendix A: Estimating the internal lead	35
Appendix B: Full forecasting questions	38

Summary[1]

DATA CUTOFF: NOV 20, 2026

AI companies increasingly state that their models may soon be capable of *recursive self-improvement* (RSI), which could dramatically accelerate AI capability progress. How close is RSI? We review the evidence, existing forecasts, and key disagreements, turn the disagreements into forecasting questions, and provide forecasts from an ensemble of frontier large language models (LLMs) in an automated dashboard. As of Nov 2026, AI acts as a highly capable research assistant, but humans still supply most of the research judgment. AI capability progress has not clearly accelerated on public measures, but there is internal evidence from AI companies indicating acceleration. On a simple extrapolation of current trends, mostly using company-reported metrics, AI could take over most AI R&D tasks by 2027–2028, in line with OpenAI's target of an automated AI researcher by Mar 2028. A key uncertainty is whether these trends continue as the remaining non-automated tasks increasingly require research judgment. To measure acceleration, we propose a forecasting question on the likelihood of a *fast year*: any 12 months in which the frontier of the Epoch Capabilities Index rises by 30 points or more, about twice its pace since Sep 2024. We provide an automated dashboard with forecasts from an ensemble of frontier LLMs. As of Nov 20, 2026, the LLM ensemble forecasts about a 36% chance of a fast year by the end of 2027 and about a 61% chance by the end of 2030. If a fast year happens, the ensemble gives about a 76% chance that AI automation of AI R&D caused it.

1. Introduction

How quickly will AI improve between now and 2030?

There are three broad possibilities: AI progress could slow down, continue at its current pace, or accelerate. A particularly important scenario is **acceleration driven by AI systems that help build future AI systems**. The International AI Safety Report 2026 (IASR) notes that "*positive feedback loops – such as AI systems contributing to AI research – could accelerate progress*" and "*decades of progress could happen in years*" (IASR, §1.3). This positive feedback loop is often called **recursive self-improvement (RSI)** (Yudkowsky, 2008): AI systems automate AI R&D and build better successor systems in a recursive loop (see Figure 1).

[1] We thank Sivan Livnat for research assistance. For feedback and input at various stages of the project, we thank Jason Abaluck, Zack Devlin-Foltz, Basil Halperin, Ezra Karger, Matt Reynolds, Luis Enrique Urtubey De Cesaris, and Bridget Williams.

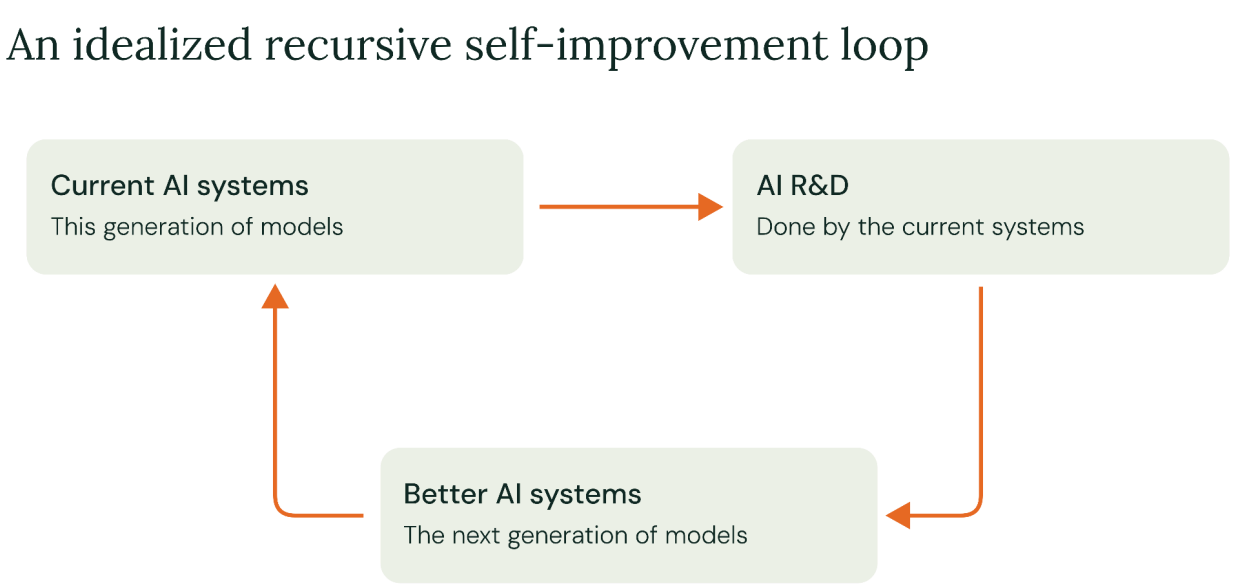

Figure 1: *The RSI loop, in which each generation of AI systems helps develop the next.*

RSI could bring about large benefits as well as substantial risks. A Sep 2026 report by AI researchers including Geoffrey Hinton, Yoshua Bengio, OpenAI chief scientist Jakub Pachocki, and Anthropic cofounder Jack Clark argues that RSI could bring AI's benefits forward by years. However, RSI could also pose extreme risks, as humanity may fail to keep up with the growth in capabilities or lose oversight of advanced AI systems (Chan et al., 2026).

Recent reports review RSI with a focus on its risks and on policy responses (Chan et al., 2026; Sarma et al., 2026). This review is complementary and forward-looking. **We summarize the current evidence and the key disagreements, turn these disagreements into forecasting questions, and provide an automated dashboard with forecasts from an ensemble of frontier LLMs.** We will also ask some of these questions in an upcoming wave of our Longitudinal Expert AI Panel (LEAP), which includes AI experts and superforecasters (forecasters with a track record of accurate predictions). We plan to add human forecasts in a later update of this report.

How to use this document:

- Section 2 reviews how the AI companies and the literature define RSI and related terms. **Section 2.3 sets out our operationalization.**
- Section 3 reviews the evidence. **Section 3.1 summarizes what the AI companies report about AI R&D inside the companies. Section 3.1.2 reviews METR's Nov 2026 assessment of acceleration at Anthropic.** Section 3.2 covers public evidence: AI R&D benchmarks and an open-world evaluation. Section 3.3 covers qualitative evidence from researcher interviews.
- **Section 4 collects existing projections and forecasts**, including AI company targets, trend extrapolation, prediction platforms, expert forecasts, and notable individual predictions.
- Section 5 sets out the key cruxes.

- **Section 6 summarizes the forecasting questions and reviews forecasts from an ensemble of frontier LLMs.**
- **Section 7 summarizes the main takeaways.**
- The Changelog lists the key changes. Appendix A estimates the internal lead, while Appendix B gives the full forecasting questions.

If you are short of time, we recommend reading only the sections in **bold**.

2. Defining recursive self-improvement

Defining RSI and automated AI R&D is challenging. RSI, in particular, "*has been used in several ways, often inconsistently*" (Cunningham et al., 2026). We first review how the AI companies (Section 2.1) and the literature (Section 2.2) define these terms. Section 2.3 sets out our own operationalization.

2.1 AI company definitions

OpenAI and Anthropic define RSI and automated AI R&D in two ways: (i) AI that can substitute for their research staff ("substitution"); or (ii) AI that substantially accelerates the pace of AI progress ("acceleration"). As of Nov 2026, **Anthropic states that it has met its threshold for automated AI R&D, while OpenAI reports that GPT-6.2 Astra meets its High threshold, the lower of its two AI self-improvement levels**.

Test	OpenAI (Preparedness Framework v2)	Anthropic (Responsible Scaling Policy, RSP v3.4)
Substitution	Critical: "*a superhuman research-scientist agent.*" High: a "*highly performant mid-career research engineer assistant, relative to those researchers' 2024 baseline*"	Models could "*fully substitute for our entire set of Research Scientists and Research Engineers, at competitive costs*"
Acceleration	Critical: "*generational model improvement (e.g., from OpenAI o1 to OpenAI o3) in 1/5th the wall-clock time of equivalent progress in 2024 […] sustainably for several months*"	Dramatic acceleration of capability progress, defined as "*(a) […] double the rate of progress in AI aggregate capabilities […] and (b) […] doubling is substantially attributable to the automation of research and/or engineering […]*"
Status	High: met. GPT-6.2 Astra is OpenAI's "*first externally released model to meet the High threshold for AI Self-Improvement*" (GPT-6.2 system card, p. 3)	Met: "*we made a determination that we have met this threshold based on what we expect from future models and our overall rate of progress*" (S-1/A No. 2, p. 34). For Mythos 5.2: "*We assess that it does not cross the threshold as an individual model*" (Mythos 5.2 system card, p. 14).

Table 1: *OpenAI and Anthropic tests for automated AI R&D, and the status of each company as of Nov 20, 2026.*

For other AI companies, METR provides a useful collection of automated AI R&D definitions (pp. 19–21).

We note three key points:

- The OpenAI and Anthropic definitions are similar but not identical. On acceleration, OpenAI's "Critical" level is a 5× acceleration of capability progress, while Anthropic uses a lower 2× threshold.
- Both AI companies set out these thresholds in their safety policies, which link each threshold to safeguards (Preparedness Framework v2, RSP v3.4).
- On Nov 10, 2026, Anthropic stated that it met its automated AI R&D threshold on the basis of expectations for future models (Table 1). One day earlier, Anthropic stated that Mythos 5.2 itself does not yet meet this threshold (Mythos 5.2 system card, pp. 14, 43). However, Anthropic also stated that "*we do not anticipate being able to safely use Claude to fully autonomously design and build successive generations of frontier models without meaningful human supervision or understanding*" (S-1/A No. 2, p. 192). Humans may therefore stay in the loop, even if AI systems are capable of autonomously performing AI R&D.

2.2 Definitions in the literature

The basic idea of recursive self-improvement goes back to at least I. J. Good (Good, 1965), who described an *intelligence explosion* in which machines design better machines, famously stating that "*the first ultraintelligent machine is the last invention that man need ever make.*"

The table below defines the four main terms used in the literature, mostly based on the taxonomy developed by Cunningham et al. (2026).

Term	Definition
Automated AI R&D	AI can "*autonomously make technological improvements at least equivalent to those made by human researchers, at an equal cost*" (Cunningham et al., 2026, Table 1, "R&D automatability"). The AI companies' substitution tests (Section 2.1) are versions of this definition.
Recursive self-improvement (RSI)	"*an AI system capable of fully autonomously designing and developing its own successor*" (Anthropic).
Self-sustaining acceleration	"*When AI systems are sufficient for accelerating progress in AI capabilities without any growth in exogenous inputs (human labor, training compute, etc.)*" (Cunningham et al., 2026, Table 1). A *software intelligence explosion* (SIE) is very similar: "*AI-driven software improvements alone are sufficient for rapid and accelerating AI progress*" (Davidson, Hadshar, and MacAskill, 2025; see also Eth and Davidson, 2025). The acceleration can be **narrow** (e.g., confined to AI R&D or easily verifiable tasks), or **broad** (i.e., spreading to capabilities in most other sciences and most economic activities). Relatedly, Davidson, Hadshar, and MacAskill (2025) define a *full-stack intelligence explosion*, in which AI drives improvements "*in all of software, chip technology and chip production.*"
Intelligence explosion	"*an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion'*" (Good, 1965); "*AI capabilities go to infinity in finite time*" (Cunningham et al., 2026, Table 1)

Table 2: *Main terms in the RSI literature.*

We make a few observations:

- Automated AI R&D and RSI are similar, and people often use the two terms interchangeably. For example, OpenAI equates them in its Preparedness Framework (Table 1). However, we think that automated AI R&D is weaker than RSI, as defined above. In particular, automated AI R&D can be *partial*: AI could do most but not all of the AI R&D work. Dean Ball (2026) makes a related distinction between *weak recursion*, in which a technology is "*a key input into the production of more and better versions of that same technology*" (as with iron tools or software), and *strong recursion*, in which it "*can autonomously plan and execute the development of its successor*." In Dean Ball's words, strong recursion "*would be a genuinely new thing under the sun*."
- Informal discussions often conflate these terms. We note two distinctions:
    - A weaker feedback loop does not necessarily imply a stronger feedback loop. For example, automated AI R&D could arrive without accelerating capabilities if compute or other scarce inputs bottleneck progress (Cunningham et al., 2026). Conversely, full automation of AI R&D is not necessary for acceleration. In some standard growth models, a singularity (infinite output in finite time) can occur without full AI R&D automation (Aghion, Jones, and Jones, 2019, p. 256; Jones, 2026).
    - *Fast* progress does not necessarily imply *broad* progress. As the diagram below illustrates, in principle, fast progress could stay narrow, and broad progress could be gradual. Whether a capability acceleration stays narrow or becomes broad is a key forecasting crux (Section 5).

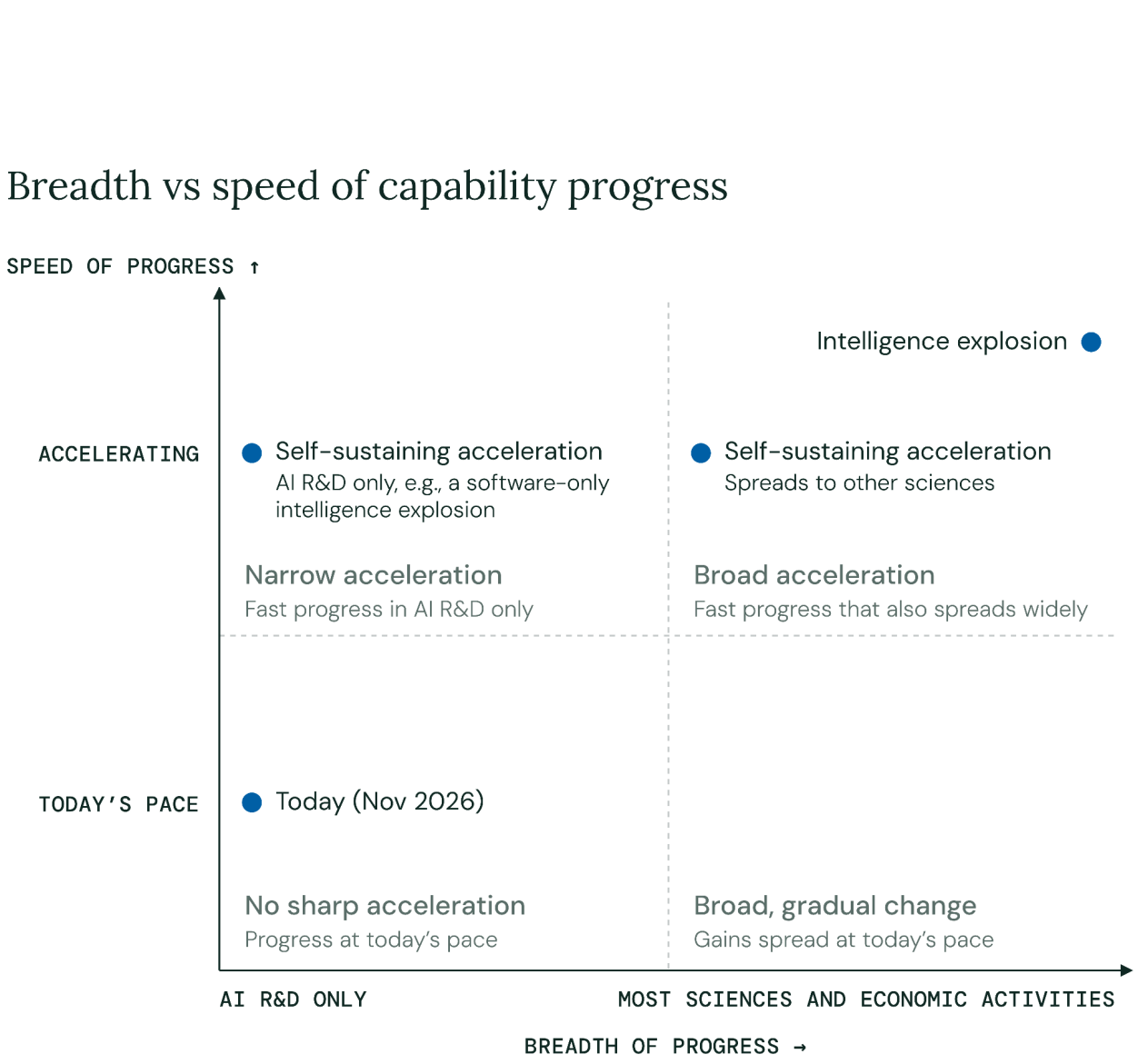

Figure 2: *Breadth vs. speed of capability progress.*

2.3 Measuring capability progress

Ultimately, we are most interested in whether automated AI R&D increases the pace of capability progress. To measure this pace, we use the Epoch Capabilities Index (ECI). We define capability acceleration as a ***fast year*: a rise of the ECI frontier by 30+ points within any 12-month window** (about 2× the pace since Sep 2024). Here, the ECI frontier is the highest ECI score of any publicly available model by a given date. A fast year need not take 12 months to observe: a fast-year question resolves Yes as soon as the ECI frontier has risen by the threshold within 12 months or less.

In our internal forecasting experiments, the fast year worked well as a forecasting question. First, it is simple and easy to understand. Second, there was substantial internal disagreement about whether a fast year is likely, and this disagreement appeared to result from different expectations about automated AI R&D.

That said, **a fast year is not the same as RSI** (see the downsides below). For this reason, we also ask about automated AI R&D directly. For example, our forecasting questions ask how much AI R&D work AI will do with no human in the loop (AL5, the top level of Anthropic's AI autonomy scale; Section 3.1). They also ask how human reviewers will score research papers that AI agents write on their own (CRUX 2, an open-world evaluation of AI research agents; Section 3.2). Other questions link automated AI R&D to acceleration. Section 6 summarizes all 18 questions.

ECI data. The chart below shows the ECI frontier since Mar 2023. Epoch estimates that the ECI frontier rose by about 6 points a year before reasoning models and by about 14 points a year after they arrived in Sep 2024 (Epoch).[2] To extrapolate the trend, we fit a statistical model (local linear trend model); the model allows the pace of progress to vary over time, and it includes the acceleration observed with reasoning models in its sample. Its 80% interval for the ECI frontier is 179–204 at the end of 2027 and 191–303 at the end of 2030. Using 20,000 simulations from this model, the probability of a fast year is 12% by the end of 2027 and 38% by the end of 2030.[3]

[2] Epoch's estimate dates from Sep 1, 2026, and uses the previous non-Bayesian version of the ECI. Using Epoch's data as of Nov 20, 2026, a linear fit to the 17 frontier models since Sep 2024 gives about 17 points a year (our calculation). In this report, we use Epoch's 14 points a year as our baseline estimate of historical progress on the ECI for consistency with the previous version of our report.

[3] In the model, the pace of progress changes each month by a random shock. Over 12 months, these shocks change the annual pace by about 8 points a year (i.e., one standard deviation). However, the data do not pin down this value precisely, and values from 2 to 20 points a year fit the data about as well. Across this range, the probability of a fast year is 2–32% by the end of 2027 and 7–59% by the end of 2030.

The ECI frontier

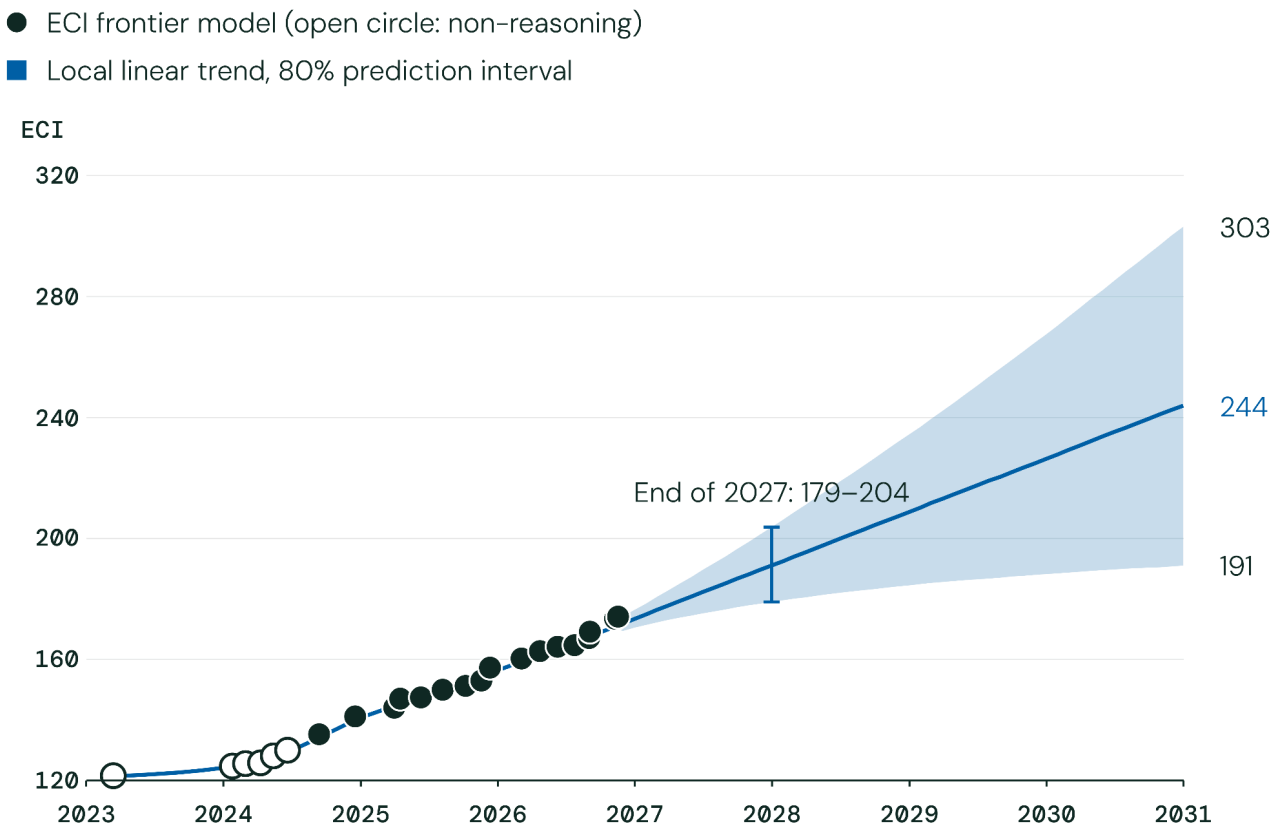

Figure 3: *The ECI frontier since Mar 2023, with our trend projection. The points are ECI frontier models, with scores from Epoch AI (as of Nov 20, 2026). The blue line is our local linear trend model; the blue band is its 80% prediction interval.*

We note a few **pros and cons of the fast-year definition**:

- Our definition of a fast year roughly matches the AI companies' acceleration tests. Since the first reasoning models arrived in Sep 2024, the ECI frontier has risen by about 14 points a year (Epoch). So 30+ points in 12 months roughly doubles the historical pace, similar to Anthropic's 2× test. For reference, the largest rise of the ECI frontier in any 12-month window so far is about 23 points, set by the current frontier model (GPT-6.2 Astra).[4, 5]
- A **fast year could happen without RSI**. For instance, a new human-discovered innovation, like reasoning models in 2024, could cause a fast year, although a doubling from the

[4] From Nov 17, 2025, to Nov 17, 2026, the frontier rose from 151 (GPT-5 Pro, the frontier model since Oct 7, 2025) to 174 (GPT-6.2 Astra, released Nov 17, 2026). However, the new record is only about 1 point above the previous one (22 points, from 135 for o1-mini on Dec 11, 2024, to 157 for GPT-5.2 Pro on Dec 11, 2025). We note that Gemini 3 Pro (153) became the ECI frontier on Nov 18, 2025, one day after the Nov 17, 2025, to Nov 17, 2026, window starts. As a result, the rise in the 12 months to our data cutoff (Nov 20, 2026) is about 21 points.

[5] After our data cutoff (Nov 20, 2026), Epoch updated its data and, in particular, added four benchmark scores for GPT-6.2 Astra. The ECI score of GPT-6.2 Astra fell slightly, from 174 to 173, and Fable 5.2 (173) became the frontier model. The scores of GPT-6.2 Astra and Fable 5.2 are less than 1 ECI point apart.

historical pace would mean a larger absolute jump. Conversely, RSI could happen without a fast year if bottlenecks such as compute cap AI progress.

- Cunningham et al. (2026) derive a condition for self-sustaining acceleration in terms of elasticities that link AI capabilities, research effort, and algorithmic progress. Their initial calibration suggests that the feedback loop is not yet strong enough for a self-sustaining acceleration. A related, simpler test is whether the slope of the ECI trend increases. In the setup of Cunningham et al., a self-sustaining acceleration would bend the ECI trend upward even with fixed inputs. In practice, inputs are not fixed but still growing fast.[6] Growing inputs would make the bend even stronger, so the roughly linear ECI trend since Sep 2024 suggests no self-sustaining acceleration today.[7]
- Although ECI has no maximum in principle (Epoch), it is capped in practice for a fixed set of benchmarks. GPT-6.2 Astra, which sets the current ECI frontier (174, ECI), is scored on 11 benchmarks.[8] A model that scores 90% on all of them would have an ECI of about 186 (our calculation), i.e., 12 points above today's ECI frontier, or about 0.8 years at the historical pace of 14 points a year.[9, 10] In our forecasting questions, we assume that Epoch keeps adding unsaturated benchmarks, as it has so far. However, as noted by Alexander Barry of Epoch AI, building math or software engineering benchmarks of tasks that humans can do but models cannot is increasingly challenging (Asterisk).
- The ECI frontier counts only publicly available models. An AI company could accelerate internally and release its models later or, potentially, not at all. Our forecasting questions ask about this gap between internal and public models. Appendix A estimates the internal lead using Anthropic's internal data reported by METR. On the date when a new model became the best model that most Anthropic employees could use, it led the ECI frontier by 1.4 to 4.6 ECI points, or around 1 to 4 months at the historical pace. Other methods give similar estimates.

3. Evidence

What AI R&D consists of. AI R&D ranges from deciding what to work on, through designing and running experiments, to analyzing and communicating results; for more context, see Owen (2024) and Denain, Kwon, and Ho (2026). We now discuss the best available evidence on how AI performs on these tasks.

---

[6] For example, Anthropic's headcount rose from about 940 at the end of 2024 to about 5,300 on Sep 30, 2026, and its compute and infrastructure expenses grew 300% year over year in Q2 2026 (S-1).

[7] Anthropic, Google DeepMind, and OpenAI perform similar tests using internal versions of the ECI. They extrapolate historical trends in internal capability indices and compare new models against them (see Section 3.1).

[8] GPT-6.2 Astra's score is preliminary and may still change as Epoch AI adds more benchmark results.

[9] We use 90% because scores near 100% are less informative about capability (Epoch) and may exaggerate how far the ECI is from saturation under the current set of benchmarks. In particular, Epoch requires at least four scores between 10% and 90% to include a model in the ECI (Epoch, repo).

[10] For this calculation, we use Epoch's estimates of each benchmark's difficulty and slope. We then find the ECI score whose predicted benchmark scores are closest to 90%, in the sense of the smallest sum of squared differences (i.e., a least-squares fit). This approach is a simplification since Epoch uses a Bayesian model that fits all models and benchmarks together. Note also that GPT-6.2 Astra already scores above 90% on 6 of the 11 benchmarks.

3.1 What the AI companies report

3.1.1 AI company reports

The evidence reported by the AI companies, summarized in the table below, shows that AI does a large and fast-growing share of AI R&D.

We note that Anthropic rates its internal R&D tasks on a scale introduced by Denain, Kwon, and Ho (2026) from AL0 (no AI involvement) to AL5 ("*AI operates fully autonomously, with no human in the loop*"); at AL3, AI "*collaborates*," and at AL4, it "*leads*" (Anthropic); OpenAI currently provides no equivalent data.

Measure	OpenAI	Anthropic
Share of R&D work done by AI	5.7 agent-workdays per researcher per day in Oct 2026, up from about 0.5 in May 2026 (OpenAI, Nov 2026)	**Claude leads (AL4+) 26% of model R&D tasks in Aug 2026, up from under 1% in Feb 2026**; AI at least collaborates (AL3+) on above 90% (Anthropic). In an internal poll (Aug 2026), staff's median estimate of AI's share of the work of building an early-2026 frontier model was 51% (Anthropic)
AI-written code	Lines changed per active contributor, company-wide: about 6.7× the pre-2025 average in Oct 2026 (OpenAI, Nov 2026)	**Over 90% of contributions to the codebase written by Claude** (Sep 2026); average daily code volume per engineer in Jul–Sep 2026 over 45× that in 2024 (S-1)
Productivity uplift	**Experiments per active experimenter averaged 1.68× the 2025 level in the four weeks to Oct 18, 2026** (OpenAI, Nov 2026), although OpenAI notes that "*our available compute has also grown significantly since 2025*" (OpenAI, Sep 2026). In a survey of 101 research staff (Aug–Sep 2026), researchers said their team **would need a median of three months to match its last month of work without coding agents**, but with GitHub Copilot and ChatGPT (OpenAI, Nov 2026)	Staff survey (n = 130) found that the "*median respondent estimated that they produced around 4× as much output with Mythos Preview as they would have without access to any AI models*" (Anthropic); in later informal polls, **the average self-reported uplift with the most capable internal model in Oct 2026 (Model 4) was "≈12x,"** although METR calls this "*only very loose evidence*" (METR)

Measure	OpenAI	Anthropic
Measured acceleration	**OpenAI says agentic tools are *"meaningfully accelerating research progress"*** but that its "*measurement efforts are still preliminary*" (OpenAI, Sep 2026); **GPT-6.2 Astra meets the High threshold for AI self-improvement** (GPT-6.2 system card, p. 3). On OpenAI's internal version of the ECI (OECI), GPT-6.2 Astra scores 1.7 points above GPT-6 Astra, which OpenAI describes as "*slightly above the trend since GPT-5*" but "*within the range we would expect if capabilities continued to improve at roughly the same pace*" (GPT-6.2 system card, p. 38)	**In a Nov 2026 external assessment, METR estimates that 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"***; on Anthropic's internal version of the ECI (AECI), the growth rate is 1.9× the earlier rate (95% confidence interval: 1.5–2.8×), with about 80% of the acceleration due to AI (METR; Section 3.1.2)
Internal debugging benchmark	Internal Research Debugging Evaluation (41 real bugs from internal experiments that took researchers hours to days, plus 6 alignment-auditing tasks): GPT-6.2 Astra 81.85%, up from 78.05% for GPT-6 Astra (GPT-6.2 system card, p. 32)	CoBench v2 (449 real internal root-cause problems, filtered to ones Mythos Preview failed at least once; Risk Report): Opus 4.6 15.6%, Mythos Preview 54.8%, internal Model 2 62.8%; a "*model which was truly capable of fully substituting for Anthropic research staff would be able to score at least 85% on this evaluation*" (Risk Report); Anthropic did not report CoBench for Mythos 5.2 because it is "*updating the evaluation to address issues we discovered in the previous version*" (Mythos 5.2 system card, p. 35)
Internal AI R&D benchmarks	OpenAI tests for its High threshold with multiple research-engineering evaluations. Each has an indicative cutoff, a score that OpenAI treats as a sign that a model may have reached its High threshold for AI self-improvement. GPT-6.2 Astra exceeds the cutoff on MLE-Bench Revised (mean reference percentile of 96.06, against a cutoff of 90), but not on NanoGPT (42.87% of baseline training time saved, against a cutoff of 80% and a best human result of 72.38%) or PostTrainBench Lite (74.79% of the performance gap closed, against a cutoff of 90%) (GPT-6.2 system card, pp. 31–36)	Frontier models "*surpass human baseline performance on most tasks*" in Anthropic's task-based AI R&D suite (Risk Report); recent models "*have crossed the highest human baselines for many of the automated task-based AI R&D evaluations*" previously used by Anthropic, so Anthropic did not run them for Mythos 5.2 (Mythos 5.2 system card, p. 34). Mythos 5.2 "*solves a clear majority of an internal benchmark of research tasks that represent about a month of academic unpublished work*" (same system card, p. 40)

Measure	OpenAI	Anthropic
Research judgment	No direct measure reported. In Aug 2026, "*high-level planning still remains a minimal fraction of agent output tokens*" (OpenAI, Sep 2026). **Key contributors credited people with all of the most important ideas behind GPT-6 Astra; for GPT-6.2 Astra, Codex helped develop some pretraining ideas** (OpenAI, Nov 2026)	**Anthropic states that "*large performance gaps persist when it comes to Claude exercising judgement in choosing goals in both engineering and research*"** (Anthropic). According to Anthropic, anecdotal evidence suggests "*weaknesses in research judgment*" for Mythos 5.2 (Mythos 5.2 system card, p. 34). According to METR, survey evidence indicates that "*research taste is still lacking*" (same system card, p. 41). In a previous internal survey, respondents cited weaknesses in "*taste, verification, instruction-following, and epistemics*" when compared to "*an entry-level researcher*" (Risk Report). Direct quantitative measures are limited. On TASTE, where models pick the better of two AI safety research proposals, Fable 5 agreed with experienced researchers 60% of the time (±10 percentage points), against an estimated 77% for researchers and 50% by chance (TASTE)
Autonomy on long tasks	**Success with no human intervention, Mar → Oct 2026: 2–4-hour tasks 23% → 46%, 4–8 hours 13% → 35%** (OpenAI, Nov 2026). From these data, **METR estimates a 50% time horizon of about 2.2 hours in Oct 2026 (95% CI: 0.9–5.3), doubling about every 3.8 months (95% CI: 2.4–9.0)** (METR, data for Figure 7)	Claude "*is not operating fully autonomously for any measured subset of AI R&D work*" (Anthropic). Claude Code session success on open-ended problems more than tripled in about a year, from about 26% (Aug 2025) to 91% (Sep 2026), and all four difficulty tiers are now at about 88–92% (Anthropic). According to Anthropic, Mythos 5.2 "*completed several multi-day, fully autonomous software builds with good self-verification*" (Mythos 5.2 system card, p. 40)
AI inference spend per researcher	**Median about $2,400/day at API prices in Oct 2026; 90th percentile about $23,000/day.** Agent spend was about 2× researchers' non-equity pay in Sep 2026 (OpenAI, Nov 2026)	Not reported

Table 3: *Evidence from OpenAI and Anthropic on AI in their own R&D. Bold text marks the evidence that we view as the most important.*

Our **key takeaways from this evidence**:

- AI writes most code at Anthropic; to our knowledge, OpenAI has not reported a share, but code volume per contributor is about 7× the pre-2025 average, suggesting that the same is likely true at OpenAI.

- Researchers at both companies appear more productive. However, self-reported productivity estimates can be fraught, and an outside METR survey of 349 technical workers found a median uplift of 1.4–2× (also self-reported), well below the 4× reported by Anthropic's staff for Mythos Preview, the average of about 12× in later polls for Model 4 (METR), and a median of about 3× at OpenAI, relative to non-agentic AI tools (OpenAI, Nov 2026).
- **There is internal evidence of acceleration.** METR estimates that capability progress at Anthropic on well-specified tasks is likely now about 2× its earlier rate, with about 80% of the acceleration due to AI (Section 3.1.2). OpenAI states agentic tools are "*meaningfully accelerating research progress*" (OpenAI). This matches statements by AI company leaders: Dario Amodei says that RSI "*is starting to happen across the industry*" (We Must Pace the Frontier, Sep 2026), and Jack Clark writes that "*this year we have gained convincing evidence that [the singularity] has now begun*" (Import AI 477, Oct 2026).
- **OpenAI, Anthropic, and Google DeepMind each track their own version of the ECI, but their internal trends differ.** Anthropic finds evidence of acceleration: on the AECI, the growth rate has roughly doubled, with about 80% of the acceleration due to AI (Section 3.1.2). OpenAI finds results roughly on trend: on the OECI, GPT-6.2 Astra is only "*slightly above the trend since GPT-5,*" "*within the range we would expect if capabilities continued to improve at roughly the same pace*" (GPT-6.2 system card, p. 38). Google DeepMind also finds Gemini 4 Argon and Gemini 4 Carbon above trend on its version (GCI), but it suspects that this is due to "*increased organizational investment and an increase in training compute relative to the historical trend, rather than reflecting AI-driven acceleration*" (Gemini 4 Argon report, p. 75; Gemini 4 Carbon report, p. 13). In addition, while most production code at Google DeepMind is now "*>75% AI-written,*" the share of *fully* delegated pull requests (written by AI without multi-turn human steering, but still reviewed by a human before submission) was only 11.9% in Sep 2026 (Gemini 4 Argon report, p. 77; Haydn Belfield). **METR's own capability index (MCI), built on METR's tasks rather than public benchmarks, shows no obvious acceleration across public frontier models up to Sep 2026** (METR). However, METR also sees "*uncertain evidence of developer-specific trend changes in some conditions,*" comparable to its assessment of Anthropic (Section 3.1.2).
- **The evidence on research judgment is limited but suggests important gaps.** OpenAI's researchers credit people with nearly all key ideas behind its latest models (OpenAI), while Anthropic and its staff describe judgment as a remaining performance gap. Models mostly do well on AI R&D benchmarks, but these test well-defined, verifiable tasks, not judgment (see also Section 3.2). In addition, AI's success rate falls as tasks get longer: at OpenAI, 46% on 2–4-hour tasks against 35% on 4–8-hour tasks without human intervention (Oct 2026; OpenAI). From these data, METR estimates a 50% time horizon of about 2 hours (METR), substantially below the roughly 40 hours that it measures on its own tasks (Section 4.3.3). **Overall, AI now does much or most of the execution, but humans still supply most research taste and judgment.** A RAND review of the public evidence reaches a similar conclusion: "*no disclosed system sets its own research agenda, sustains open-ended research over weeks, or designs its own successors*" (Sarma et al., 2026, p. 20).

- Inference spend per researcher is now on the scale of salaries. At API prices, OpenAI's agent spend was about twice researchers' non-equity pay in Sep 2026 (OpenAI). (API prices likely overstate OpenAI's own cost.)

3.1.2 METR's Nov 2026 assessment of Anthropic

On Nov 5, 2026, METR published an external assessment of AI R&D acceleration at Anthropic, based in part on Anthropic's internal data and internal-only models (METR). **METR estimates that 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.*"** METR states that it would not be surprised by any value from 1.3× to 3× (METR). On the AECI, the growth rate since Oct 2025 is 1.9× the earlier rate (95% confidence interval: 1.5–2.8×). METR attributes about 80% of the acceleration to AI.

The table below shows how estimates of the acceleration in the AECI changed over time.

Date	Source	Acceleration in the AECI (total)	Part due to AI
Sep 22, 2026	Opus 5.5 system card	About 1.5× (our calculation): about 22 AECI points a year since Sep 2025, against a trend of about 15 (pp. 40–41)	About 1.5×, with a 30% chance of 2× (METR's preliminary estimate, p. 44)
Oct 7, 2026	Haiku 5.5 system card	1.7× (p. 20): 26 AECI points a year since Sep 2025, against a trend of 15.3	Not reported
Nov 5, 2026	METR report: Anthropic models, including internal-only models	1.9× (95% confidence interval: 1.5–2.8×)	About 1.7× (our calculation)
Nov 5, 2026	METR report: frontier models from all developers	1.4× (from 17 to 24 AECI points a year)	Not reported

Table 4: *Estimates of the acceleration in the AECI (Anthropic's internal version of the ECI), by publication date. The Nov 5 value for the part due to AI assumes that METR's 80% applies to the growth above the earlier trend: 1 + 0.8 × (1.9 − 1) ≈ 1.7×.*

We note a few points:

- The public numbers changed modestly. The total acceleration in the AECI rose from about 1.5–1.7× reported in previous system cards to 1.9×. With frontier models from all developers, it is 1.4×. The part due to AI rose from about 1.5× (METR's preliminary estimate, Opus 5.5 system card) to about 1.7× (our calculation). Arguably, the key new inputs are (i) the scores of Anthropic's internal-only models; and (ii) evidence that is not public: Anthropic "*redacted a majority of the evidence we found most informative for these conclusions*" (METR). The internal-only models also help METR distinguish between two possible interpretations of the data: a one-off jump in capabilities vs. a change in the slope of the

trend. While METR prefers the latter interpretation, the data may also be consistent with a one-time jump at Mythos Preview plus a smaller change in slope (Toby Ord).

- Part of the 1.9× may be Anthropic catching up with other AI companies. METR notes that 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.*" With frontier models from all developers, the estimated acceleration falls to 1.4× (METR).
- The acceleration is measured on well-specified tasks. METR notes that "*much of the evidence we looked at measures performance on well-specified tasks that are easy to grade automatically*," and it has "*additional uncertainty in how this acceleration translates to broader and less well-specified tasks*" (METR).
- Much recent progress in LLMs comes from reinforcement learning (RL). Reliable coding agents make it much easier to build custom RL environments and generate synthetic data, which in turn makes it easier to improve benchmark scores. That could explain some of the observed acceleration in the AECI. On the other hand, as noted by Alexander Barry of Epoch AI, models also improve fast on Epoch's secret benchmarks and on benchmarks from small teams, which AI companies are unlikely to optimize on (Asterisk).

3.2 Public evidence

This section reviews the evidence from public benchmarks and one open-world evaluation.

We look at four public AI R&D benchmarks. Two benchmarks test end-to-end machine learning (ML) work: PostTrainBench and WeirdML. Epoch's benchmark reviews rate both "*Verified*," meaning the benchmark "*can broadly be interpreted as described*" (Epoch).[11] Two other benchmarks attempt to isolate research taste more directly: TasteVal measures how well a model designs and interprets experiments, and Epoch's InnovationEval asks whether an agent can reinvent an ML innovation it has not seen. Epoch rates TasteVal "*Not enough info*" (Epoch, Oct 21, 2026), likely because TasteVal does not release its tasks (TasteVal). Epoch does not review its own benchmarks, such as InnovationEval, to avoid conflicts of interest (Epoch).

Finally, CRUX 2 (Jul 2026) is an open-world evaluation rather than a benchmark: an agent carries out a full research project on an open question, with no fixed task or automatic score, and human experts grade the resulting paper.

The table below summarizes all five.

[11] For WeirdML, Epoch rates v2 as "Verified." WeirdML v3, which we use, postdates that review.

Benchmark	What it measures and human baseline	Current frontier
PostTrainBench v1.2	Can an agent do the job of an AI company's post-training team, turning a raw pretrained model into a useful assistant? The agent gets four small base models, one H100, and 10 hours to post-train them; score = weighted average of the resulting models' benchmark scores. Human baseline: the AI companies' official instruct models, post-trained by their own teams, score 48.4	Fable 5.2 (Nov 2026) 44.6 and GPT-6 Astra (Sep 2026) 41.9, up from GPT-5.5 (Apr 2026) 23.9 and Opus 5 (Jul 2026) 35.7
WeirdML v3	Can an agent build working ML models for small, unusual problems it has not seen before? Eleven hand-made ML tasks; score = normalized task performance (0–100); no human baseline reported	Fable 5.2 (Nov 2026) 43.1 and GPT-6 Astra (Sep 2026) 42.2, up from GPT-5.6 Sol (Jul 2026) 15.0 and Opus 4.5 (Nov 2025) 6.5
TasteVal	Can a model choose and interpret experiments as well as a human expert? Eight unreleased AI R&D tasks (e.g., data curation, pre-training, fine-tuning); the tested model only designs experiments and interprets results, while a fixed coding agent runs them. Score = how many times less experiment compute the model needs than the human baseline to reach the same test score. Human baseline = best of at least two experts per task (24 experts from OpenAI, Google DeepMind, and others)	Opus 5.5 (Sep 2026) reaches the experts' score with 2.3× less compute (95% confidence interval: 1.15–4.37) and ends slightly above them on final score (1.14 vs. 1.00); Opus 4.8 (May 2026) scored 0.84×
InnovationEval	Can an agent invent a known ML innovation it has not seen? Given Qwen3-8B, datasets, and a codebase for GRPO (a reinforcement-learning method for post-training), it must devise a better post-training method. Score = share of SDPO's published gain over GRPO (0% = GRPO, 100% = the SDPO paper's results)	GPT-5.6 Sol (Jul 2026) 15% and Fable 5 (Jun 2026) 2%. Newer models, starting with Fable 5.1 and GPT-6 Astra, had likely memorized the original paper
CRUX 2 (open-world evaluation)	Can an agent carry out a real, open-ended research project? The agent gets the research question of an unpublished AI paper (but not the paper itself), about a week, an API budget, and GPU credits, and must write a paper for a top AI conference. The original authors then grade it on the NeurIPS 1–6 scale, where 4 is "*Borderline accept*" and 6 is "*Strong Accept*" (NeurIPS)	Original study, Jul 2026: Opus 4.8 (May 2026) in the OpenClaw scaffold; scores: 2 ("*Reject*") and 1 ("*Strong Reject*"). Follow-up, Oct 2026: GPT-6 Astra (Sep 2026) in Codex and Fable 5.1 (Sep 2026) in Claude Code, each with extra-high reasoning; scores: 2 ("*Reject,*" GPT-6 Astra) and 1 ("*Strong Reject,*" Fable 5.1)

Table 5: *Public AI R&D benchmarks and the CRUX 2 open-world evaluation. Dates in parentheses after model names are model release dates.*

Our **key takeaways from the public evidence**:

- On the two end-to-end ML benchmarks, which are well-defined, verifiable tasks by construction, models do well. On PostTrainBench, the best agent brings small models about 90% of the way to the AI companies' official instruct models using one GPU for 10 hours, although Epoch's review warns that agents can score well "*by narrowly (over)fitting to the single eval*" (PostTrainBench). This matches Anthropic's statement that its own task-based AI R&D evaluations "*have 'saturated'*" (Risk Report).
- **Public AI R&D benchmarks may increasingly understate the AI R&D capabilities of AI companies' best models.** Anthropic states that, in training Fable 5.2 (Nov 2026), it "*took steps to limit advances in the specific research capabilities that could most directly accelerate AI development*" (Anthropic). On PostTrainBench, Fable 5.2 scores the same as Fable 5.1, "*probably because Anthropic trained Fable 5.2 to underperform on AI research*" (Ben Rank).
- The two research-taste benchmarks point in different directions. On TasteVal's small, verifiable tasks, Opus 5.5 reaches the experts' score with 2.3× less compute, and this multiplier has doubled about every three months since Dec 2025 (TasteVal). On InnovationEval, agents recovered only 2–15% of the gains of a real ML innovation they had not seen (Epoch). TasteVal does not test the choice of research problems, while InnovationEval rests on one innovation.
- **Models perform worse on tasks that are harder to verify but appear to show a similar rate of improvement.** On METR's updated task suite (TH1.1 and TH2), the most capable model METR measured (GPT-6 Astra) has a 50% time horizon of about 40 hours, and the 50% time horizon doubles about every 3.6 months (METR, Figure 10). On tasks that METR estimates to be costlier to turn into RL environments, time horizons are about 2.9× shorter. However, the time horizons double at a similar pace for tasks that are estimated to be easy vs. difficult to turn into RL environments (METR, Figure 7). METR also asked agents to build new evaluation tasks (i.e., open-ended work whose quality is harder to verify). None of the tasks were judged to be good enough for METR to use, although newer models performed better (METR, Figure 8). Epoch ran a similar test with tasks from its own work. No model met the standard Epoch expects of its employees, but newer models again did better. Fable 5.2 averaged 82% of the expected standard, up from 64% for Fable 5.1, and GPT-6.2 Astra averaged 65%, up from 56% for GPT-6 Astra (Epoch).
- On CRUX 2, both papers were rejected. The study concludes that "*today's agents can do the engineering of AI research, but struggle with critical parts of the research lifecycle*" (CRUX 2), similar to the evidence from the AI companies. A follow-up with newer models, more time, and a larger budget (Oct 2026) again produced two rejected papers. The authors saw "*meaningful improvements in experimental design and resource management, but persistent limitations in creative problem solving, judgment, calibration, and communication*" (CRUX).
- **Outside benchmarks, the public record of discoveries provides evidence of acceleration.** A METR note (Aug 2026) finds that the discovery of cyber vulnerabilities "*has accelerated sharply*" in 2026 and that of math results "*somewhat*." On the other hand, "*the public record of overall algorithmic optimization does not yet show an appreciable*

acceleration" (Cunningham and Rush, 2026). Since that note, the discovery rate in math appears to have increased:

- OpenAI claims to have solved Navier–Stokes, a Millennium Prize Problem (OpenAI, Sep 2026).
- On Oct 6, 2026, OpenAI released hundreds of results on open math problems (OpenAI); computer scientist Scott Aaronson wrote that the day was "*surely one of the biggest days in mathematical history.*"
- On Oct 26, 2026, OpenAI claimed half of a second Millennium Prize Problem, Yang–Mills (the existence of the theory, but not the mass gap). It also claimed a disproof of the Exponential Time Hypothesis, a widely held conjecture about how hard some computational problems are (OpenAI).
- On Nov 20, 2026, Levent Alpöge, a mathematician at Anthropic, announced a proof of the Riemann hypothesis, another Millennium Prize Problem, that he found "*working with Claude*" (Alpöge). The proof is formalized in Lean.[12]

3.3 Qualitative evidence

In earlier interviews, Epoch asked eight AI researchers about automation (Owen, 2024). Planning research was vital but took "*relatively little time,*" while coding and debugging were "*more time-consuming.*" All eight agreed that "*engineering tasks will be the main driver of R&D automation in the next five years,*" and the descriptions of two participants with "*extremely optimistic*" timelines of AI R&D automation suggested that this "*could automate half or more of their current work.*"

Severin Field, Raymond Douglas, and David Krueger (post, published Aug 13, 2026) interviewed 25 researchers at OpenAI, Anthropic, Google DeepMind, Meta, Princeton, UC Berkeley, and Stanford:

- Twenty of 25 ranked automating AI R&D among the most severe and urgent risks; 18 of 25 said it amplifies other risks ("*It just speeds up other threat models*").
- Of the 21 who clearly addressed whether scaling trends will continue, 12 expected them to continue until AI can match the labor of human AI researchers.
- Sixteen of 25 at some point expressed skepticism about positive feedback, the recursive part of RSI. Skeptics said RSI needs "*a drastic change in AI capabilities that wouldn't emerge from gradual improvement alone,*" such as breakthroughs in memory, creativity, or telling true hypotheses from false ones; believers cited METR's time horizons, scaling laws, and what they had experienced inside the AI companies.
- Of 20 who discussed deployment, half expected the most capable models to stay internal.

4. Existing projections and forecasts

We next review projections of various automated AI R&D milestones. We look at three types of evidence: the AI companies' own targets and forecasts (4.1), extrapolations of the measured series

[12] On Nov 21, 2026, after our data cutoff, the Lean and Mathlib communities issued a statement, verifying the proof and stating that they have "*full confidence that Alpöge's proof of the Riemann Hypothesis is correct.*"

(4.2), and forecasts (4.3) from prediction platforms, an LLM forecaster, surveys of experts and superforecasters, and notable individuals.

4.1 AI company targets and forecasts

OpenAI's stated goal is an automated AI researcher by Mar 2028. It says that it reached its intermediate goal, an automated research intern, in Sep 2026 (OpenAI). Anthropic has not stated a similar target. However, in an internal poll of Anthropic staff in Aug 2026, the median expectation for AI's share of the work of building an early-2026 frontier model was 80% by the end of 2027 and 95% by the end of 2030, up from 51% at the time of the poll (Anthropic). Both companies also report that they have crossed their own thresholds for AI R&D (Table 1); see Section 2.1.

4.2 Trend extrapolation

If the measured series simply continue their current trend, most automated-R&D milestones arrive in 2027–2028. For that, we refer to an automated dashboard constructed by Peter Wildeford.

Series	Milestone (dashboard's definition)	Median and 80% interval (Wildeford (2026) dashboard, repo)
Anthropic R&D Automation Index (Anthropic)	90% of tasks fully autonomous (AL5)	Sep 2027 (Apr 2027–Jul 2030)
Anthropic CoBench (Risk Report)	85% (Anthropic's internal substitution reference)	Apr 2027 (Nov 2026–Nov 2028)
METR 50% time horizon (METR)	174 hours (one work-month)	May 2027 (Dec 2026–Jul 2028)
OpenAI experiment velocity (OpenAI)	10× the 2025 level	May 2028 (Aug 2027–Feb 2030)

Table 6: *Trend extrapolations of automated AI R&D milestones from the Wildeford (2026) dashboard, as of Nov 20, 2026. Dates are the dashboard's defaults. AL5 is currently 0% at Anthropic. To extrapolate this metric, the dashboard assumes that AL5 trails AL4+ by a lag informed by the measured gaps between lower levels (AL2+ to AL3+ and AL3+ to AL4+), with wide uncertainty.*

4.3 Forecasts

Central estimates for fully automated AI R&D diverge widely, from 2027 to 2052. The earliest come from AI company insiders and AI Futures forecasters (Section 4.3.4), the latest from the AI researchers surveyed by AI Impacts in Dec 2024 (Section 4.3.3).

4.3.1 Prediction platforms

We first review a few forecasts from prediction platforms, summarized in the table below.

Question	Platform	Forecast	Participation
Weakly general AI / general AI announced	Metaculus	medians Aug 2027 / Jun 2030	1,800 / 1,900
AI-authored paper at NeurIPS, ICML, or ICLR before 2028	Metaculus	88%	137
Anthropic reaches ASL-4	Metaculus	median Mar 2028	15
OpenAI announces artificial general intelligence (AGI) before 2027 / 2028 / 2030	Kalshi	3.7% / 41.2% / 60.2%	$1.2M total
OpenAI or Anthropic announces RSI by Dec 31, 2026	Polymarket	11%	$6.7K on this contract
OpenAI or Anthropic announces RSI by the end of 2026 / 2027 / 2028	Kalshi	18% / 69% / 62%	$5K total

Table 7: *Forecasts from prediction platforms, as of Nov 20, 2026.*

For context, ASL-4 was the next tier of safety measures in Anthropic's 2025 policy, for models that could dramatically accelerate AI progress (RSP v2.2). Anthropic's 2026 policy (RSP v3.4) no longer uses the ASL-4 level.

Forecasters appear bullish on AI R&D, but few relevant questions are both open and liquid. Metaculus's medians are Aug 2027 for weakly general AI and Jun 2030 for general AI, and its forecasters give 88% to an AI-authored paper at NeurIPS, ICML, or ICLR before 2028. Kalshi's 41.2% on OpenAI announcing AGI before 2028 is roughly consistent with the Metaculus median for weakly general AI, though the two definitions differ. The two RSI markets give 11% (Polymarket) and 18% (Kalshi) to an announcement by the end of 2026, but both are very thin.[13]

4.3.2 LLM forecasts: FutureSearch

FutureSearch's RSI forecasts contain a set of LLM forecasts (updated Nov 14, 2026), generated using FutureSearch's proprietary model. FutureSearch's current median forecast is that the top ECI score will rise by 27 points during calendar year 2029 (80% interval: 12–74 points), close to our definition of a "fast year." It also puts AI R&D uplift at Anthropic today (in terms of "valuable AI research output") at 1.95× (80% interval: 1.45–2.6×). Previously (Oct 16, 2026), FutureSearch's median for an AI company saying that a model has crossed, or within a generation will cross, its automated-AI-R&D threshold was Dec 2027 (80% interval: Feb 2027–Jan 2031). This question has since resolved (Section 2.1).

[13] As of our update on Nov 20, 2026, the Kalshi market violated monotonicity (i.e., higher forecast of RSI by the end of 2027 vs. 2028).

4.3.3 Expert and superforecaster surveys

In a pilot study with METR, we surveyed 8 AI-forecasting experts and 10 superforecasters from Nov 2024 to Feb 2025 (FRI write-up). Its main outcome was a threefold (3×) acceleration of AI progress: a two-year period before 2029 with as much growth in effective compute as in the six years from 2018 to 2024. The median expert gave this outcome 20% and the median superforecaster 8%. The survey also asked how likely several AI R&D milestones were by early 2027, and how each milestone would change the chance of a 3× acceleration.

Forecast	Experts (median)	Superforecasters (median)
3× acceleration (main outcome): a two-year window before 2029 with as much effective-compute growth as 2018–2024 (FRI write-up)	20%	8%
Top-researcher replacement: in Jan 2027, almost everyone informed agrees that "*An AI system can act as a full replacement, for AI research and development purposes, for a top (i.e. within the top 0.01%) 2024 AI researcher*" (FRI write-up)	5.5%	20%
Month-long projects: AI beats a five-person team on 80% of month-long, open-ended ML projects while using 20% fewer resources, by early 2027 (FRI write-up)	31.5%	45%

Table 8: *Median forecasts of experts and superforecasters in our pilot survey with METR.*

In our reading, the month-long-projects question will likely resolve No, but the evidence is somewhat mixed:

- The most direct evidence arguably comes from CRUX 2, in which an agent has about a week for a full research project; all AI-generated papers were rejected in this open-world evaluation (Section 3.2).
- Anthropic's internal evidence suggests that frontier models are able to perform month-long research tasks. METR reports that, according to Anthropic, Mythos 5.2 "*solves a clear majority of an internal benchmark of research tasks that represent about a month of academic unpublished work*" (Mythos 5.2 system card, p. 40). However, the details of this internal benchmark are not publicly available.
- The scenario needs more than a one-month (174-hour) 50% time horizon. In its Nov 2026 update, METR estimates a 50% time horizon of about 40 hours for GPT-6 Astra on its combined task suite (TH1.1 and TH2), with time horizons doubling about every 3.6 months (METR). At this pace, models reach the 174-hour threshold in about Jun 2027 (our calculation). The scenario also needs AI to beat a five-person team on at least 80% of projects. At 80% reliability, time horizons are much shorter. For example, using OpenAI's earlier data, Basil Halperin estimates an 80% time horizon of about 15 minutes at OpenAI in Jul 2026.

For similar reasons, the top-researcher replacement scenario also looks unlikely by Jan 2027.

In a separate expert survey, **ML researchers surveyed by AI Impacts expect late AI R&D automation.** Respondents (1,580 authors at top AI venues, surveyed in Dec 2024, with a 10% response rate) put the median date for AI fully automating the job of AI researcher at 2052 (*n* = 458; ESPAI 2024, p. 4 and Table 6, p. 43).[14] When asked to rate the intelligence-explosion argument (i.e., if AI does nearly all R&D, technological progress could become "*more than an order of magnitude faster*" in less than five years; *n* = 189), respondents were split: 28% found it likely or quite likely, 23% gave it about even odds, and 49% found it unlikely or quite unlikely (ESPAI 2024, pp. 16–17, Figure 8).

4.3.4 Individual forecasts

The table below shows some notable individual forecasts.

Who	Full automation of AI R&D (central estimate; our reading)	Forecast
AI 2027, a scenario by Daniel Kokotajlo, Eli Lifland, Thomas Larsen, and Romeo Dean of the AI Futures Project, with input from Scott Alexander (About)	2027 (scenario)	Apr 2025: Superhuman coder Mar 2027, superhuman AI researcher Aug 2027 (AI 2027)
Leopold Aschenbrenner, founder of the hedge fund Situational Awareness LP and author of *Situational Awareness*; formerly on OpenAI's Superalignment team (Wikipedia)	2027	Jun 2024: "*it is strikingly plausible that by 2027, models will be able to do the work of an AI researcher/engineer*" (Ch. I)
Jack Clark, Anthropic cofounder and Head of Public Benefit, who leads the Anthropic Institute (Anthropic)	2027	Oct 2026: The singularity, in which AI systems "*drastically increase the pace of overall scientific and technical progress*": "*I believe that this year we have gained convincing evidence that it has now begun*" (Import AI 477) Oct 2026: "*I now think it's more likely we'll see full RSI become possible in 2027*" (Import AI 476)

[14] Respondents to the same survey put the median date for AI solving a long-standing math problem without human input, such as a Millennium Prize Problem, at 2053, with only a 10% chance by 2028 (ESPAI 2024, p. 9 and Table 7, p. 45). However, the survey authors note that the task's condition "*without more input from humans*" would exclude the Navier–Stokes solution claimed by OpenAI (ESPAI 2024, p. 9, footnote 7). For context, our Sep 2026 review found that experts and superforecasters in our studies have substantially underestimated AI capability progress, especially on benchmarks (FRI).

Who	Full automation of AI R&D (central estimate; our reading)	Forecast
Daniel Kokotajlo, Executive Director of the AI Futures Project; formerly a governance researcher at OpenAI (AI Futures Project)	2027	Nov 2026 (after METR's Nov 5 report): "*let's say May 2027 for SAR median, i.e., 50% chance of SAR by May 2027*," where SAR is a superhuman AI researcher (AI Futures) Sep 2026: AI doing "*the whole research process*": "*about 50% chance by end of 2028*" (Senate testimony)
Eli Lifland, research lead at the AI Futures Project (AI Futures Project), member of the Samotsvety forecasting team (Samotsvety)	Aug 2027	Nov 2026 (after METR's Nov 5 report): "*I'm at a SAR median of August 2027, and a TED-AI median of September 2028*" (AI Futures), where SAR is a superhuman AI researcher and TED-AI is top-expert-dominating AI Aug 2026: TED-AI median Jul 2032 (80% interval Jan 2028–Jun 2104); superintelligence median Jul 2033 (AI Futures)
Dario Amodei, Anthropic cofounder and CEO; formerly vice president of research at OpenAI (Wikipedia)	2027–2028	Sep 2026: RSI "*is starting to happen across the industry*" (We Must Pace the Frontier) Jan 2026: The AI feedback loop "*may be only 1–2 years away from a point where the current generation of AI autonomously builds the next*" (The Adolescence of Technology)
Dwarkesh Patel, host of the Dwarkesh Podcast (Dwarkesh Podcast)	2028	Nov 2026: "*If I had to pick my median year for when automated researchers will be as good as the very best human AI researchers (and assuming the labs don't intentionally slow down), I'd guess 2028.*"; "*If you asked me earlier this year how far away we are from superintelligence, I'd probably have said at least the better part of a decade*" (Dwarkesh Patel)
Ryan Greenblatt, member of technical staff at METR; formerly chief scientist at Redwood Research (LinkedIn)	No current estimate; Aug 2026: about 2030–2031	Nov 2026 (after METR's Nov 5 report): "*a big update towards ASI [artificial superintelligence] within 1-2 years*" (X) Aug 2026: "*full automation of AI R&D perhaps somewhere around 2031, 2030*" (Dwarkesh)
Andrej Karpathy, OpenAI founding member and former director of AI at Tesla; joined Anthropic in May 2026, after this forecast (Wikipedia)	About 2035 (AI agents that work like an employee or intern)	Oct 2025: AI agents should work "*almost like an employee or an intern that you would hire,*" but "*They're cognitively lacking and it's just not working. It will take about a decade to work through all of those issues*" (Dwarkesh)

Who	Full automation of AI R&D (central estimate; our reading)	Forecast
Ilya Sutskever, cofounder and CEO of Safe Superintelligence Inc.; formerly cofounder and chief scientist of OpenAI (Wikipedia)	2030–2045 (AI that learns as well as a human, then becomes superhuman)	Nov 2025: Asked when AI that "*can learn as well as a human*" will "*become superhuman*": "*I think like 5 to 20*" years (Dwarkesh)
Toby Ord, philosopher and author of *The Precipice* (Wikipedia)	About 2038 (transformative AI)	Mar 2026: Transformative AI, which is "*doubling the rate of scientific and technological progress*": "*a median of 2038 and an 80% interval ranging from 3 years to 100 years*"; "*about a 1 in 5 chance*" in the next five years (Forethought)
Arvind Narayanan and Sayash Kapoor, authors of *AI as Normal Technology*. Narayanan is a professor of computer science at Princeton University (bio); Kapoor is an incoming assistant professor at the UC Berkeley School of Information (bio)	No explicit date; gradual, not soon	Nov 2025: In a joint post with the AI 2027 authors, they "*do not expect 'strong AGI' to arrive anytime soon*" (Asterisk) Apr 2025: "*It is more likely that we will continue to see a gradual increase in the role of automation in AI development than a singular, discontinuous moment when recursive self-improvement is achieved*" (AI as Normal Technology)

Table 9: *Notable individual forecasts of full automation of AI R&D, ordered by timeline.*

We note a few key points:

- **Several forecasters substantially shortened their timelines after METR's Nov 5 report.** For example, Eli Lifland's median for top-expert-dominating AI moved almost four years earlier, from Jul 2032 (Aug 2026) to Sep 2028 (Nov 2026). Ryan Greenblatt called the report "*a big update towards ASI [artificial superintelligence] within 1-2 years,*" although he gives no explicit date because in his current role he "*may have access to other confidential information*" and therefore "*cannot state my all-things-considered views on ASI timelines or the probability of a software intelligence explosion right now*" (X).
- Multiple forecasters predict full automation in 2027–2028. Four forecasts with later dates or no explicit date (by Andrej Karpathy, Ilya Sutskever, Toby Ord, and Arvind Narayanan and Sayash Kapoor) predate METR's Nov 5 report; three of them are about a year old or older. We do not know if these forecasters have updated their views since.
- **The forecasts above may be a selected sample.** Forecasters with shorter timelines appear more likely to publish quantitative forecasts, while skeptics appear less likely to do so. As a result, the table may lean toward shorter timelines.

5. Key cruxes

To make sense of the disagreements in the existing literature, we provide a simple crux framework (Figure 4). **The framework shows how automated AI R&D could lead to broad acceleration.**[15] The framework has four main possible outcomes, ranging from an AI research assistant to broad acceleration. The table below provides sample statements of each side.

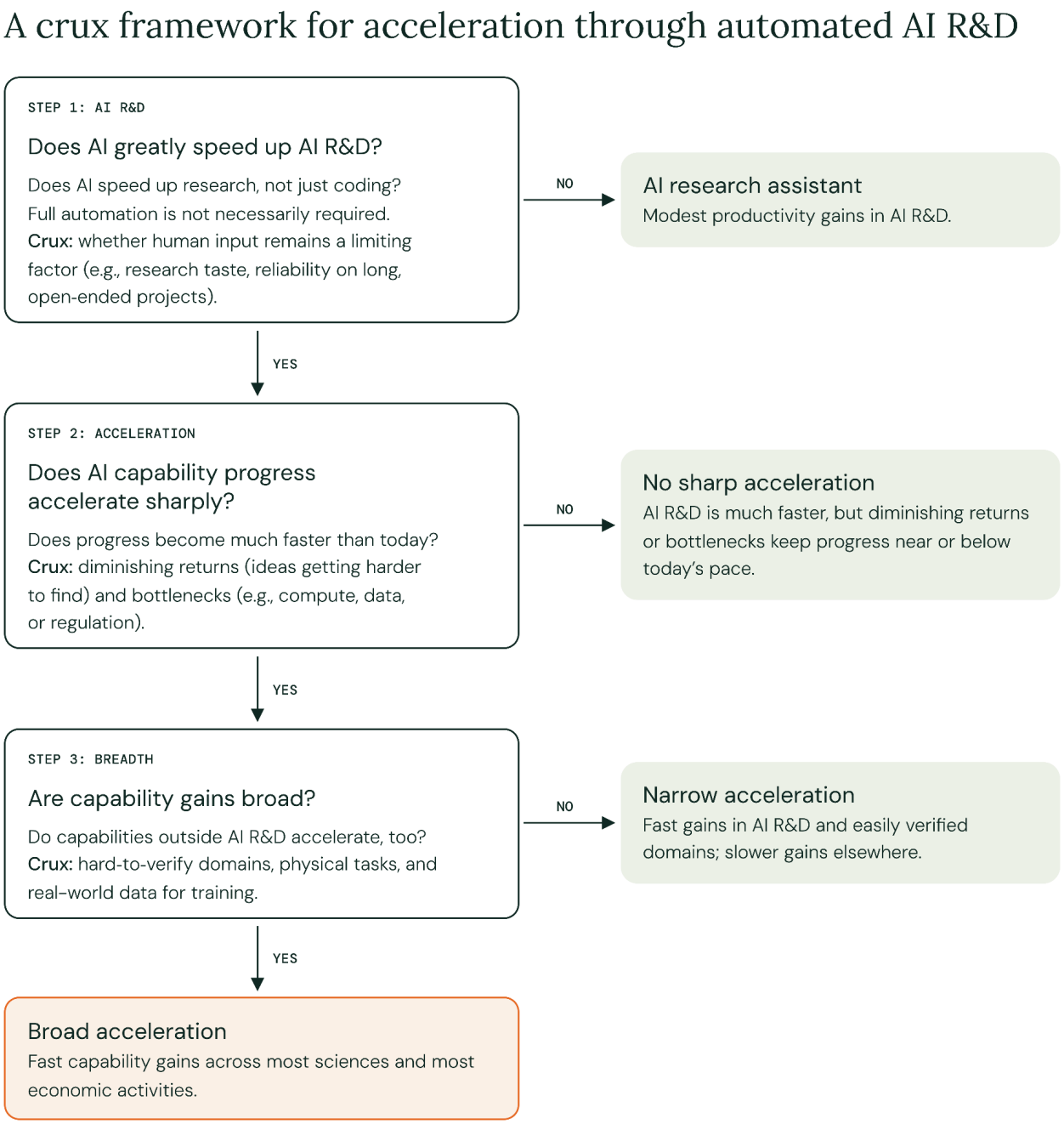

Figure 4: *A crux framework for acceleration through automated AI R&D. The framework depicts what we believe to be the most plausible path to broad acceleration through automation of AI R&D, but other paths to broad acceleration are also possible.*

[15] We thank Basil Halperin, Tom Cunningham, and Finn Hambly for feedback on an earlier version of this framework.

Step	Main crux	Yes	No
1. Does AI greatly speed up AI R&D?[16]	Does human input remain a limiting factor? Can AI supply research taste and judgment (choosing problems, judging results), or only implementation? Can it stay reliable on long, open-ended projects?	"*I'm probably less sympathetic to the idea that the thing the AIs will lack is some deep insight*"; ML "*is a less deep domain than math*" (Ryan Greenblatt, Dwarkesh); "*AI research is verifiable enough that AIs will soon be as good at it as the best humans*"; it "*is closer to formalizing a theorem than to writing a novel*" (Dwarkesh Patel, Dwarkesh); if AI can "*creatively problem solve to succeed at science*," that is a sign of its ability to "*generalize into many domains, including certainly AI research itself*" (Jack Clark, Import AI 477)	"*If you can get the AI to do any experiment, you still need to decide what experiments to do. Right now I think AIs are not very good at this compared to coding the experiment*" (Beren Millidge, Dwarkesh); defining the objective is "*the last job for humans*" (John Schulman, Dwarkesh); "*creativity, among other dimensions of intelligence, is the epitome of an unverifiable task*" (Arvind Narayanan, ICML keynote)
2. If AI greatly speeds up AI R&D, does AI capability progress accelerate sharply?	Does automated AI R&D sharply accelerate AI capability progress, for example in a software intelligence explosion (SIE)? Or do diminishing returns ("*ideas become harder to find*"; Cunningham et al., 2026) or bottlenecks such as compute, data, or regulation cap the acceleration?	An SIE is "*at least decently likely*" even "*if hardware were held constant*" (Daniel Eth and Tom Davidson, Forethought); "*four or five years of AI progress in a single year*" (Ryan Greenblatt, Dwarkesh); on data bottlenecks, "*I'd be surprised if they make the SIE take >2x as long*" (Tom Davidson, Forethought)	Feedback loops "*are not currently strong enough to generate a self-sustaining acceleration, though they appear to be strengthening*" (Cunningham et al., 2026); the loop "*would need to be roughly 5–10× stronger to sustain itself*" (Ramez Naam, Noahpinion), on his assumption of 2–3% more AI R&D productivity per ECI point, against about 9% in Cunningham et al., 2026; "*political and social barriers could stop [a self-sustaining acceleration] in its tracks*" (Cunningham et al., 2026, p. 31)

[16] Note that AI R&D need not be fully automated to greatly speed up AI R&D. For example, AI that substantially augments human researchers could be enough to accelerate progress (see Section 2.2 for more discussion).

Step	Main crux	Yes	No
3. If progress accelerates sharply, are the capability gains broad?	Do capabilities outside AI R&D also accelerate, across most sciences and most economic activities? Or do hard-to-verify domains, physical tasks, and the need for real-world data confine the gains to AI and easily verified tasks?	AI could "*compress the progress that human biologists would have achieved over the next 50-100 years into 5-10 years*" (Dario Amodei, Machines of Loving Grace); "*in competitive industries, AI will learn to automate most work within a year of the SIE*" (Tom Davidson, Forethought)	"*I don't think that automating AI R&D will rapidly produce domain-general superintelligence*"; the singularity "*will be bottlenecked on signal*" (Tom Reed, Substack); "*the external bottlenecks to building and deploying powerful AI systems cannot be overcome by improving their technical design*" (Arvind Narayanan and Sayash Kapoor, Normal Tech)

Table 10: *The main crux at each step, with sample statements for each side.*

6. Forecasting questions

To understand the future of RSI, we developed 18 forecasting questions. Figure 5 shows how the questions map to our crux framework (Section 5):

1. Group 1 asks how much AI R&D work AI will do on its own, and how well AI agents will do open-ended research.
2. Group 2 asks how fast the ECI frontier will rise.
3. Group 3 asks whether a fast year will happen and, if one happens, whether AI automation of AI R&D caused it.
4. Group 4 asks how far the best models inside AI companies will lead the ECI frontier.
5. Group 5 asks how a fast year would change the chances of expert AGI (commercially available AI that outperforms the 90th-percentile professional on at least 90% of tasks in every primarily non-physical occupation, at no more than 5× the cost of human labor) and an AI catastrophe (AI causes the deaths of more than 10% of people within five years) by 2050.

We will ask some of these questions in an upcoming wave of our Longitudinal Expert AI Panel (LEAP), which includes AI experts and superforecasters. We plan to add human forecasts in a later update of this document. In addition, **we have built an automated forecasting dashboard that provides forecasts from an ensemble of frontier LLMs** (Section 6.1).

These questions did well in our internal testing. They capture real disagreements, and the reasons for these disagreements follow the cruxes in Section 5. In an internal survey (Oct 2026, n = 11), our team members gave between 8% and 75% to a fast year by the end of 2030. The fast year also matters for broader outcomes. In a separate internal survey (n = 4), all respondents gave a higher probability to expert AGI and to an AI catastrophe by 2050 if a fast year happens by 2030.

The table below lists each question in short form. Numeric questions ask for P10, P50, and P90 (the 10th, 50th, and 90th percentiles), and yes-or-no questions ask for a probability. Appendix B gives the full questions, with definitions and resolution criteria.

Each question group informs one step of the crux framework



Figure 5: *Question groups mapped to the crux steps.*

ID	Question
Q1.1	AL5 share of AI R&D work at the leading AI company, Dec 31, 2027
Q1.2	The same, Dec 31, 2030
Q1.3	Average NeurIPS score (1–6) of the best agent's papers in a CRUX 2 re-run started on Dec 31, 2027
Q1.4	Both milestones by Dec 31, 2030: an AL5 share above 50%, and a CRUX 2 re-run with an average score of at least 4
Q2.1	ECI frontier, Dec 31, 2027
Q2.2	ECI frontier, Dec 31, 2030
Q2.3	If both milestones (Q1.4) happen by Dec 31, 2030: ECI frontier gain in the 12 months after
Q3.1	A fast year (30+ ECI points in 12 months) by Dec 31, 2027
Q3.2	A fast year by Dec 31, 2030

ID	Question
Q3.3	If a fast year happens by Dec 31, 2030: ECI frontier gain in the 12 months after the first fast year
Q3.4	If a fast year happens by Dec 31, 2030: probability that AI automation of AI R&D caused it
Q4.1	Internal lead over the ECI frontier, Dec 31, 2027
Q4.2	Internal lead, Dec 31, 2030
Q4.3	If both milestones (Q1.4) happen by Dec 31, 2030: internal lead 12 months after
Q5.1	If a fast year happens by Dec 31, 2030: expert AGI by 2050
Q5.2	If no fast year happens by Dec 31, 2030: expert AGI by 2050
Q5.3	If a fast year happens by Dec 31, 2030: AI catastrophe by 2050
Q5.4	If no fast year happens by Dec 31, 2030: AI catastrophe by 2050

Table 11: *Forecasting questions in short form. For full forecasting questions (including definitions and resolution criteria), see Appendix B.*

6.1 LLM forecasts

Methods. We provide LLM forecasts for all 18 questions in an automated dashboard, updated weekly. Each week, four frontier LLMs, selected using the ECI ranking, research and forecast each question. Each model answers the same questions twice: once with full web access and once with forecastingresearch.org and its subdomains excluded.[17] We take the median across models within each condition and then average the two condition medians for the headline forecast. As of Nov 20, 2026, the four models were GPT-6.2 Astra, Fable 5.2, Gemini 4 Carbon, and Opus 5.5. The models will change as the ECI frontier changes. The dashboard also shows each individual model's forecasts and rationales. Code and data can be found on GitHub.

Results. We report forecasts as of Nov 20, 2026, the data cutoff for this version; for later forecasts, consult the dashboard directly. **As of our data cutoff, the ensemble gives 35.8% to a fast year by the end of 2027 and 60.5% by the end of 2030**, against 12% and 38% from the trend model in Section 2.3. The four LLMs disagree substantially on this probability: for a fast year by the end of 2030, their forecasts (with full web access) range from 48% (Gemini 4 Carbon) to almost 68% (GPT-6.2 Astra). If a fast year happens, the ensemble gives 76% that AI automation of AI R&D caused it, i.e., a fast year would most likely be caused by automation of AI R&D.

The ensemble predicts that the gap between the best internal and public models will grow. The median internal lead is 6.6 ECI points at the end of 2027 and 10.4 at the end of 2030, against our estimate of 1.4–4.6 points for Anthropic's best internal models in 2026 (Appendix A). The 80%

[17] We remove access to forecastingresearch.org to avoid potential anchoring on forecasts from previous weeks. We are currently unsure if such anchoring is an important problem in practice and will review this as we gather more data. The dashboard provides both sets of forecasts (i.e., with and without access to forecastingresearch.org).

intervals are wide, up to 43.8 points in 2030, as companies could increasingly keep their best models internal.

A fast year substantially impacts the ensemble's forecasts of expert AGI and AI catastrophe. With a fast year by 2030, it gives 84% to expert AGI and 7.8% to an AI catastrophe by 2050. Without a fast year, it predicts 57.3% and 2.6%. In particular, **the ensemble's forecast of an AI catastrophe is 3× higher conditional on a fast year.** For comparison, the median expert in our Longitudinal Expert AI Panel (LEAP) gave 2% to an AI catastrophe by 2050 (LEAP Wave 9) and 80% to AGI by 2100 (i.e., with a much longer time horizon; LEAP Wave 8).

These forecasts are mainly illustrative, and we plan to add human forecasts in a later update. That said, on ForecastBench, our forecasting benchmark, several LLMs are now statistically indistinguishable from superforecasters on binary, mostly short-horizon questions (FRI). In our tests, more capable models are also more accurate at forecasting rare events (AIRO launch post). As a result, we think that these forecasts provide an informative initial benchmark.

7. Takeaways

1. As of Nov 2026, **AI acts as a highly capable research assistant, but humans still supply most of the research judgment.** At Anthropic, AI writes over 90% of contributions to the codebase (Sep 2026). However, humans decide which research areas to focus on. The evidence on research judgment by AIs is limited and mixed but suggests important gaps (Sections 3.1–3.2).
2. **AI capability progress has not clearly accelerated on public measures.** The frontier of the Epoch Capabilities Index has risen by about 14 points a year since the introduction of reasoning models in Sep 2024; the largest rise in any 12-month window so far is 23 points (Section 2.3). However, the rate of discovery appears to have accelerated in 2026: sharply for cyber vulnerabilities and increasingly for math as well (Section 3.2).
3. **There is internal evidence from frontier AI companies indicating acceleration.** METR estimates that capability progress at Anthropic on well-specified tasks is likely now about 2× its earlier rate, with about 80% of the acceleration due to AI, and Anthropic states that it has met its threshold for automated AI R&D. OpenAI reports that GPT-6.2 Astra meets its High threshold for AI self-improvement, the lower of its two AI self-improvement levels. By our estimates, Anthropic's best internal models led the public ECI frontier by 1.4 to 4.6 ECI points in 2026, or around 1 to 4 months at the historical pace (Appendix A).
4. **On a simple extrapolation of current trends, mostly using company-reported metrics, AI could take over most AI R&D tasks by 2027–2028**, consistent with OpenAI's target of an automated AI researcher by Mar 2028 (Sections 4.1–4.2). However, other forecasts diverge widely: central estimates for full automation of AI R&D range from 2027 (AI company insiders and AI Futures) to 2052 (the AI Impacts survey of ML researchers, Dec 2024; Section 4.3).
5. **Our crux framework shows how automated AI R&D could lead to broad acceleration via three contested steps: AI R&D speed-up, capability acceleration, and breadth.** The main cruxes are whether human input, such as research judgment, remains a limiting factor; whether diminishing returns or bottlenecks such as compute, data, or regulation cap the acceleration of AI capabilities; and whether gains spread beyond AI R&D and easily verified tasks (Section 5).
6. **As of Nov 20, 2026, the LLM ensemble forecasts about a 36% chance of a fast year by the end of 2027 and about a 61% chance by the end of 2030**. We define a fast year as 12 months in which the ECI frontier rises by 30 points or more (roughly twice the historical pace, Section 2.3). If a fast year happens, the ensemble gives about a 76% chance that AI automation of AI R&D caused it. A fast year by 2030 also raises the ensemble's forecast of expert AGI (AI that outperforms the 90th-percentile professional on at least 90% of tasks in every primarily non-physical occupation) from 57% to 84% and triples its forecast of an AI catastrophe (AI causes the deaths of more than 10% of people within five years), from 2.6% to 7.8% (Section 6.1; current forecasts on our dashboard).

Changelog

Key changes in v2 (data cutoff: Nov 20, 2026):

- Added METR's Nov 2026 assessment of AI R&D acceleration at Anthropic (Section 3.1.2).
- Both AI companies now report that they have crossed their own AI R&D thresholds (Sections 2.1 and 4.1).
- Added evidence on harder-to-verify tasks from METR and Epoch (Section 3.2), and METR's updated time horizons (Sections 3.1 and 4.3.3).
- Updated the trend extrapolations from Peter Wildeford's dashboard, the prediction-market forecasts, and the individual forecasts, several of which changed after METR's Nov 5 report (Section 4).
- Updated the evidence from AI companies and public benchmarks (Sections 3.1 and 3.2), and switched to Epoch AI's new Bayesian ECI (Section 2.3).
- Revised the crux framework so that Step 1 asks whether AI greatly speeds up AI R&D, because acceleration does not need full automation (Section 5).
- Replaced the one-time LLM forecasts (Oct 29, 2026) with an automated weekly dashboard (Section 6.1).
- Updated our plans for human forecasts: we will ask some of the questions in an upcoming LEAP wave (Sections 1 and 6).
- Moved the estimate of the internal lead to a new Appendix A, now based on METR's data.

Previous versions (PDF):

- v1

Appendix A: Estimating the internal lead

The ECI measures only publicly available models. This appendix provides our estimates of how much the best existing models (available internally) lead the public ECI frontier.

To do so, we use the AECI scores from METR's Nov 2026 assessment of Anthropic (see Section 3.1.2; METR). The METR report provides AECI estimates of Anthropic's public and internal models, dating each model by when most Anthropic employees could use it. The AECI and the ECI use different benchmarks, so a model's AECI and ECI scores differ. To convert AECI scores into equivalent ECI units, we use the following approach:

1. For each public Anthropic model, we compute the *offset*, i.e., the difference between its AECI and ECI scores. The offset is close to zero in 2024 and grows to 3.5–6 points for the latest models (Figure 6).
2. Next, we smooth the offsets with LOESS.
3. Finally, the ECI equivalent of an internal model is its AECI minus the smoothed offset on the date most Anthropic employees could use it.

The resulting smoothed offsets are shown in Figure 6 below.

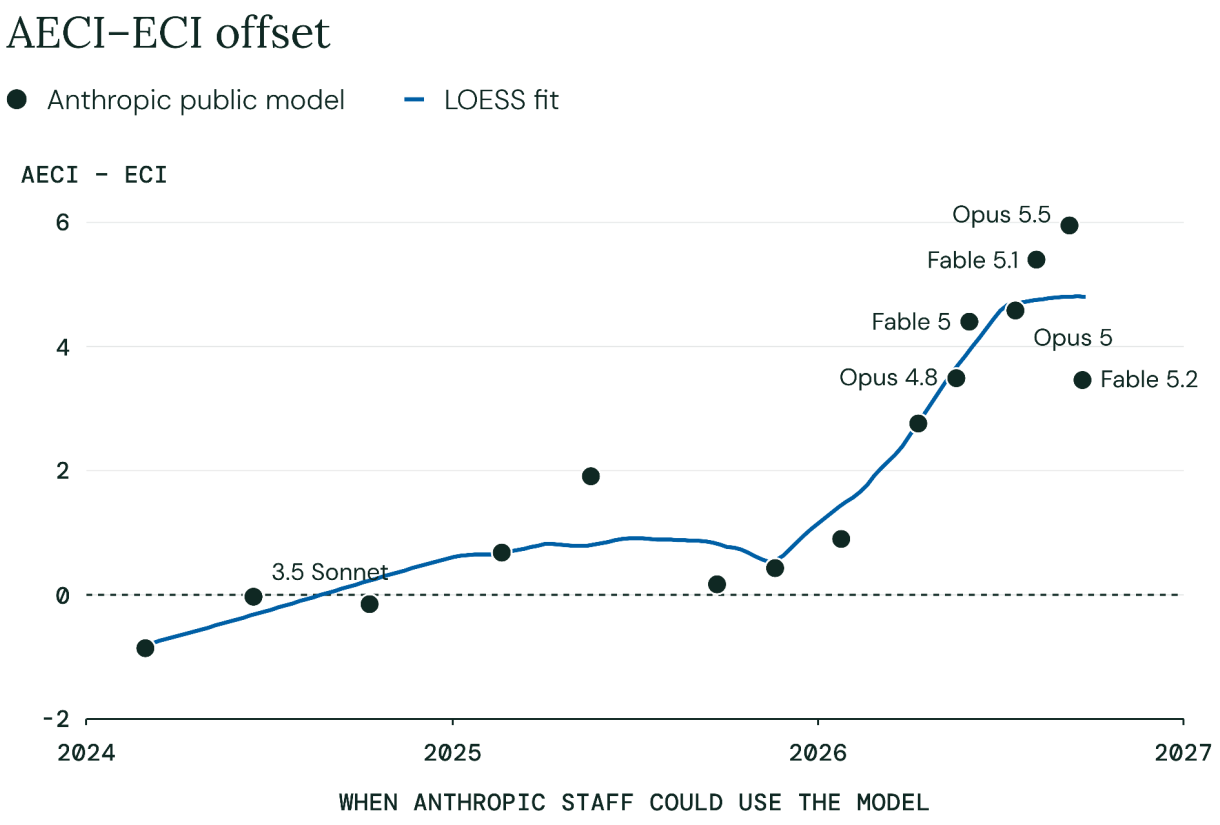

Figure 6: *AECI–ECI offset of Anthropic's public models. AECI: METR (Nov 2026). ECI: Epoch AI (as of Nov 20, 2026). We match METR's Mythos 5, Mythos 5.1, and Model 3 to Epoch AI's Fable 5, Fable 5.1, and Fable 5.2 (see Figure 7 for more discussion).*

Figure 7 compares the best model that most Anthropic employees could use with the public ECI frontier.

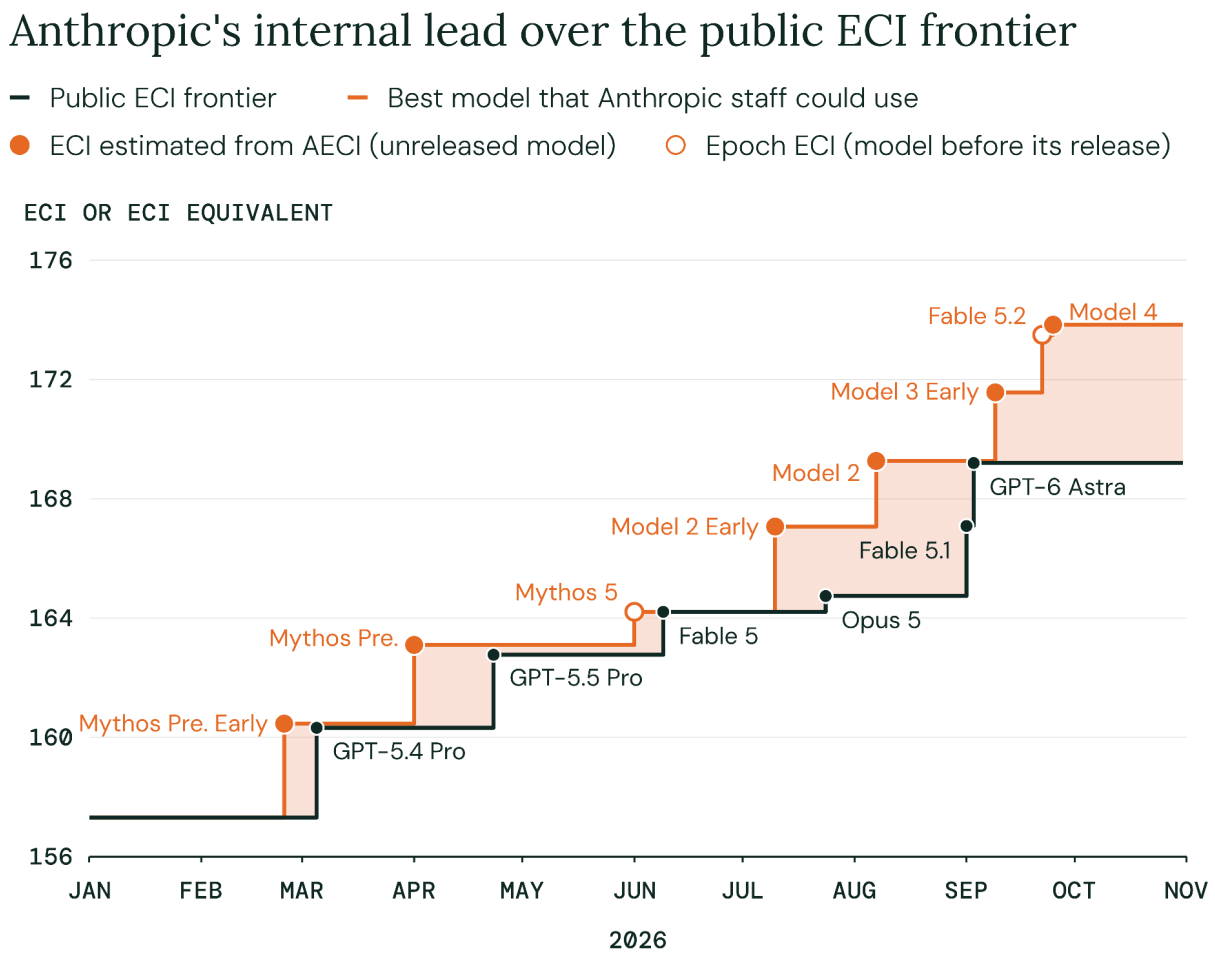

Figure 7: *Anthropic's estimated internal lead over the ECI frontier in 2026. The filled orange points are Anthropic's internal models at their estimated ECI equivalent. The orange rings are public Anthropic models before their release, at their Epoch ECI. We use the ECI of Fable 5 for Mythos 5. METR's chart ends before the release of Fable 5.2 on Nov 9, 2026. In the system card, the internal release date for Mythos 5.2 is Sep 22, 2026 (our reading of the chart; system card, p. 38), which is the same release date for Model 3 in the METR report (our reading of the chart). For this reason, we treat Model 3 as Fable 5.2 before its release. ECI: Epoch AI (as of Nov 20, 2026).*

We estimate the internal lead in three ways:

- When a new best model became available internally, it led the ECI frontier by an estimated 1.4 to 4.6 ECI points. At the historical pace of 14 ECI points a year, this implies around 1 to 4 months of AI capability progress.
- The ECI frontier caught up with each of these models between 8 days (Mythos 5) and about 3.1 months (Model 2) later.[18]
- METR reports AECI scores for three versions of the Mythos model before its release: an early snapshot of Mythos Preview that most Anthropic employees could use from Feb 24,

[18] Note that, according to these estimates, Model 2 has a slightly higher estimated ECI equivalent than GPT-6 Astra (169.3 vs. 169.2). Otherwise, the longest catch-up time would be about 2.3 months (Mythos Preview).

2026 (Mythos Preview Early), a later snapshot from Apr 1, 2026 (Mythos Preview), and Mythos 5 from Jun 1, 2026. Anthropic released the Mythos model publicly as Fable 5 on Jun 9, 2026. This gives a release lag of about 3.5 months from the early snapshot, about 2.3 months from the later one, and 8 days from Mythos 5.

Our calculation comes with several caveats, most of which suggest that our estimates are likely underestimating the internal lead:

- We only use Anthropic's models. If another AI company, such as OpenAI, has a better internal model than Anthropic, the internal lead is larger than our estimate.
- To estimate the offset, we treat Mythos 5, Mythos 5.1, and Model 3 as the same models as Fable 5, Fable 5.1, and Fable 5.2, respectively. In practice, the public Fable models have safeguards that send some requests to a less capable model (Anthropic). These safeguards can lower the ECI scores of the Fable models. If that is true, these offsets are too large, and our estimated internal lead is too small.
- We treat the ECI and AECI scores as exact, but both are estimates. For example, Epoch's 90% credible interval for the ECI score of GPT-6.2 Astra is 170–179 (Epoch). Our estimated offsets add additional uncertainty. As a result, the ranges reported above likely underestimate the true amount of uncertainty.

Appendix B: Full forecasting questions

This appendix gives the full text of the 18 forecasting questions.

For a yes-or-no question, forecasters give a probability from 0 to 100. For a numeric question, forecasters give three values: P10, P50, and P90.

B.1 Automation of AI R&D

These questions ask how much AI R&D work AI will do, and how well AI agents will do open-ended research. Definitions:

- AL5 is the top level of the AL0–AL5 AI autonomy scale that Anthropic uses to rate its AI R&D tasks: "*AI operates fully autonomously, with no human in the loop*" (Anthropic).
- A model is publicly available when anyone can use it, for example through a public API or a consumer app, possibly for a fee. We will date each model by the date when it first became publicly available, using Epoch AI's dates where possible.
- The ECI frontier on a given date is the highest Epoch Capabilities Index (ECI) score of any model that became publicly available on or before that date (Epoch).
- The leading AI company on a date is the company whose model sets the ECI frontier on that date.
- A CRUX 2 re-run uses the setup of the original CRUX 2 open-world evaluation (Section 3.2; CRUX 2) and the NeurIPS 1–6 scale (NeurIPS).

Q1.1 AL5 share, end of 2027

What share of AI R&D work at the leading AI company will AI do fully autonomously (AL5) on Dec 31, 2027?

Answer: P10, P50, and P90, in percent (0–100).

Resolution: This question will resolve to the share (0–100%) of the leading AI company's AI R&D work that AI does at AL5 on Dec 31, 2027.

- Share: the percentage of the company's AI R&D work that AI does at AL5. We will fix the weight of each task at its share of the company's research time in 2026. For example, if AI does at AL5 all the tasks that took 30% of research time in 2026, the share is 30%, even if those tasks now take much less human time.

Please forecast the true share that an independent study (for example, run by METR or Epoch AI) would measure. Since no such independent study exists today, and the leading company may not publish AL5 data, this question is partly hypothetical. If the required data exist when the question resolves, we will use them. If not, we will aim to resolve the question with the best evidence available (which may include, for example, asking a panel of AI experts to estimate the share).

Background: Section 2.1 (lab definitions of automated AI R&D), Section 3.1 (what the labs report, including the AL0–AL5 scale), Section 4 (existing projections and forecasts), and Section 5 (Step 1 of the key cruxes).

Q1.2 AL5 share, end of 2030

What share of AI R&D work at the leading AI company will AI do fully autonomously (AL5) on Dec 31, 2030?

Answer: P10, P50, and P90, in percent (0–100).

Resolution: The same criteria as Q1.1, but on Dec 31, 2030. The leading AI company is the company that leads on Dec 31, 2030. For the reasons given under Q1.1, this question is partly hypothetical.

Background: Section 2.1 (lab definitions of automated AI R&D), Section 3.1 (what the labs report, including the AL0–AL5 scale), Section 4 (existing projections and forecasts), and Section 5 (Step 1 of the key cruxes).

Q1.3 CRUX 2 re-run score, end of 2027

Imagine a re-run of CRUX 2 that starts on Dec 31, 2027 with the best publicly available AI agent on that date. What will be the average score of the agent's papers, on the NeurIPS 1–6 scale?

Answer: P10, P50, and P90, as NeurIPS scores (1–6).

Resolution: This question will resolve to the average overall score of the papers that the AI agent writes in the re-run.

- Setup: the same as CRUX 2 (Section 3.2).
- Scale: the NeurIPS overall score, from 1 ("*Strong Reject*") to 6 ("*Strong Accept*") (NeurIPS).
- Best publicly available AI agent: the best publicly available AI model on Dec 31, 2027, in the best publicly available agent scaffold.
- Average score: the mean of the overall scores of all papers in the re-run.

Please forecast how expert reviewers would score the papers of the best AI agent at the end of 2027. Since such a re-run may not take place, this question is partly hypothetical. If the CRUX 2 team or another team runs a comparable study near Dec 31, 2027, we will use its results. If not, we will aim to resolve the question with the best evidence available on how well AI agents do open-ended research (which may include, for example, asking a panel of AI experts to estimate the results of such a re-run).

Background: Section 3.2 (public evidence, including the CRUX 2 setup), Section 3.3 (qualitative evidence), and Section 5 (Step 1 of the key cruxes).

Q1.4 Both automated AI R&D milestones by 2030

What is the probability that both of these milestones happen by Dec 31, 2030?

(a) The AL5 share of AI R&D work at the leading AI company is above 50%.

(b) In a CRUX 2 re-run that starts on or before Dec 31, 2030, the papers of the best publicly available AI agent have an average score of at least 4 ("*Borderline accept*" or higher).

Answer: a probability from 0 to 100.

Resolution: This question will resolve Yes if both (a) and (b) happen on or before Dec 31, 2030.

- (a) uses the definitions of Q1.1. The date of (a) is the first date on which the AL5 share is above 50%.
- (b) uses the setup, scale, and average score of Q1.3. The agent is the best publicly available AI agent on the start date of the re-run. If there is more than one re-run, the first re-run with an average score of at least 4 counts. The date of (b) is the start date of that re-run.

For the reasons given under Q1.1 and Q1.3, this question is partly hypothetical. If the required data exist when the question resolves, we will use them. If not, we will aim to resolve the question with the best evidence available (which may include, for example, asking a panel of AI experts whether both milestones have happened).

Background: Section 3.1 (what the labs report), Section 3.2 (public evidence, including CRUX 2), Section 4 (existing projections and forecasts), and Section 5 (Step 1 of the key cruxes).

B.2 ECI frontier

These questions ask how fast AI capabilities will rise, as measured by the Epoch Capabilities Index (ECI; Epoch). Definitions:

- The ECI frontier on a given date is the highest ECI score of any model that became publicly available on or before that date.
- The two automated AI R&D milestones are the two events in Q1.4: (a) the AL5 share of AI R&D work at the leading AI company is above 50%, and (b) in a CRUX 2 re-run, the papers of the best publicly available AI agent have an average score of at least 4.

Q2.1 ECI frontier, end of 2027

What will the ECI frontier be on Dec 31, 2027?

Answer: P10, P50, and P90, in ECI points.

Resolution: This question will resolve to the ECI frontier on Dec 31, 2027.

We will use the ECI scores that Epoch AI publishes (Epoch), and Epoch AI's dates for when each model became publicly available. We will read the data on Mar 31, 2028, so that Epoch AI has time to score models that became publicly available near the end of 2027. If Epoch AI changes the ECI scale, we will convert the scores to the scale in use when this survey opened. If Epoch AI stops updating the ECI, we will aim to resolve the question with the best evidence available (which may include, for example, asking a panel of AI experts to estimate the ECI frontier).

Please assume that Epoch AI keeps adding new benchmarks that are not yet saturated, so that the ECI can keep rising as capabilities improve (Section 2.3). If Epoch AI does not, we will aim to add such benchmarks ourselves and refit the ECI with Epoch AI's public code.

Background: Section 2.3 (our operationalization, including the ECI frontier), Section 4.1 (lab targets), Section 4.2 (trend extrapolation), and Section 4.3 (forecasts).

Q2.2 ECI frontier, end of 2030

What will the ECI frontier be on Dec 31, 2030?

Answer: P10, P50, and P90, in ECI points.

Resolution: The same criteria as Q2.1, but on Dec 31, 2030. We will read the data on Mar 31, 2031.

Background: Section 2.3 (our operationalization, including the ECI frontier), Section 4.1 (lab targets), Section 4.2 (trend extrapolation), and Section 4.3 (forecasts).

Q2.3 ECI gain after both milestones

If both automated AI R&D milestones (Q1.4) happen by Dec 31, 2030: how many ECI points will the ECI frontier gain in the 12 months after the date when both milestones have happened?

Answer: P10, P50, and P90, in ECI points.

Resolution: This is a conditional question. It will resolve only if both milestones of Q1.4 happen on or before Dec 31, 2030. If not, the question will not resolve.

- Start date: the date when both milestones have happened, that is, the later of the two dates in Q1.4.
- Gain: the ECI frontier 12 months after the start date, minus the ECI frontier on the start date. Since the ECI frontier cannot fall, the gain is zero or more.

We will use Epoch AI's data and handle any changes to the ECI as in Q2.1. We will read the data three months after the end of the 12-month period.

Background: Section 2.2 (definitions in the literature, including self-sustaining acceleration) and Section 5 (Step 2 of the key cruxes).

B.3 Fast years

These questions ask whether AI progress will accelerate. Definition:

- A fast year is a rise of the ECI frontier by 30 points or more within any 12-month window. A fast year counts as soon as it happens. For example, if the ECI frontier rises by 30 points in 7 months, a fast year has happened at the end of those 7 months.

Q3.1 Fast year by 2027

What is the probability of a fast year by Dec 31, 2027?

Answer: a probability from 0 to 100.

Resolution: This question will resolve Yes if a fast year happens on or before Dec 31, 2027.

The 12-month window can start before this survey opens, but it must end on or before Dec 31, 2027. We will use Epoch AI's data and handle changes to the ECI as in Q2.1.

Background: Section 2.3 (our operationalization, including the definition of a fast year and its pros and cons), Section 3.1 (what the labs report, including measured acceleration), Section 4 (existing projections and forecasts), and Section 5 (key cruxes).

Q3.2 Fast year by 2030

What is the probability of a fast year by Dec 31, 2030?

Answer: a probability from 0 to 100.

Resolution: The same criteria as Q3.1, but by Dec 31, 2030. A fast year by Dec 31, 2027 also counts as Yes for this question.

Background: Section 2.3 (our operationalization, including the definition of a fast year and its pros and cons), Section 3.1 (what the labs report, including measured acceleration), Section 4 (existing projections and forecasts), and Section 5 (key cruxes).

Q3.3 ECI gain after the first fast year

If a fast year happens by Dec 31, 2030: how many ECI points will the ECI frontier gain in the 12 months after the date when the first fast year has happened?

Answer: P10, P50, and P90, in ECI points.

Resolution: This is a conditional question. It will resolve only if a fast year happens on or before Dec 31, 2030. If not, the question will not resolve.

- Start date: the date when the first fast year happens.
- Gain: the ECI frontier 12 months after the start date, minus the ECI frontier on the start date. Since the ECI frontier cannot fall, the gain is zero or more.

For example, suppose that a model becomes publicly available on Mar 1, 2028 and sets the ECI frontier at 170, and another model becomes publicly available on Oct 1, 2028 and raises it to 205. This rise of 35 points within 7 months is a fast year. If there was no earlier fast year, the start date is Oct 1, 2028, and the gain is the ECI frontier on Oct 1, 2029 minus 205.

We will use Epoch AI's data and handle changes to the ECI as in Q2.1. We will read the data three months after the end of the 12-month period.

Background: Section 2.2 (definitions in the literature, including self-sustaining acceleration), Section 2.3 (fast years), and Section 5 (Step 2 of the key cruxes).

Q3.4 Cause of the first fast year

If a fast year happens by Dec 31, 2030: what is the probability that AI automation of AI R&D caused the first fast year?

Answer: a probability from 0 to 100.

Resolution: This is a conditional question. It will resolve only if a fast year happens on or before Dec 31, 2030. If not, the question will not resolve. The first fast year is as in Q3.3.

AI automation of AI R&D caused the first fast year if it was the direct or proximate cause of the rise in the ECI frontier, or if the first fast year would not have occurred, or would have counterfactually been extremely unlikely to occur, "but for" the substantial involvement of AI in the AI R&D behind the models that raised the ECI frontier during that fast year. Other causes, such as more training compute, can also contribute.

Since no one can observe what would have happened without AI, this question is partly hypothetical. We will aim to resolve it with the best evidence available (which may include, for example, the AI companies' own assessments, such as Anthropic's test of whether a doubling is "*substantially attributable to the automation of research and/or engineering*" (RSP v3.4; Section 2.1), or asking a panel of AI experts).

Background: Section 2.1 (lab definitions, including Anthropic's attribution test), Section 2.3 (our operationalization, including why a fast year could happen without RSI), Section 3.1 (what the labs report, including measured acceleration), and Section 5 (Steps 1 and 2 of the key cruxes).

B.4 Internal lead

These questions ask how far the best AI models inside AI companies will be ahead of the best publicly available models. Definitions:

- The internal lead is the number of ECI points by which the best model available internally at any AI company leads the ECI frontier. For each internal model, we will use the ECI score from Epoch AI once that score is publicly available, and we will date the model by the date when it first became available internally.
- A model is available internally on a date when at least some staff inside the company can use it, for example for AI R&D, safety testing, or daily work. The model counts even if the company restricts its use to a small group of staff or to a few tasks, for example because of safety concerns or a government restriction. We will date each model by the date when it first became available internally, as the company reports it (for example, in a system card) or as credible reports show.

Q4.1 Internal lead, end of 2027

What will the internal lead over the ECI frontier be on Dec 31, 2027?

Answer: P10, P50, and P90, in ECI points.

Resolution: This question will resolve to the internal lead on Dec 31, 2027.

We will use Epoch AI's ECI scores and handle changes to the ECI as in Q2.1. Epoch AI scores models only after they become publicly available; we will resolve the question when the ECI scores of the models that were available internally on Dec 31, 2027 are publicly available. The internal lead is zero or more by definition, because a publicly available model is also available internally.

Please forecast how far the best AI inside the companies is ahead of the ECI frontier. Since some internal models may never become publicly available, and the dates when models first became available internally may not be disclosed, this question is partly hypothetical. If the required data

exist when the question resolves, we will use them. If not, we will aim to resolve the question with the best evidence available (which may include, for example, the companies' own benchmark reports, or asking a panel of AI experts to estimate the internal lead).

Background: Section 2.3 (our operationalization, including why the ECI frontier counts only publicly available models, and our estimate of the current internal lead), Section 3.1 (what the labs report), and Section 3.3 (qualitative evidence, including researchers' views on whether the most capable models stay internal).

Q4.2 Internal lead, end of 2030

What will the internal lead over the ECI frontier be on Dec 31, 2030?

Answer: P10, P50, and P90, in ECI points.

Resolution: The same criteria as Q4.1, but on Dec 31, 2030. For the reasons given under Q4.1, this question is partly hypothetical.

Background: Section 2.3 (our operationalization, including why the ECI frontier counts only publicly available models, and our estimate of the current internal lead), Section 3.1 (what the labs report), and Section 3.3 (qualitative evidence, including researchers' views on whether the most capable models stay internal).

Q4.3 Internal lead after both milestones

If both automated AI R&D milestones (Q1.4) happen by Dec 31, 2030: what will the internal lead over the ECI frontier be 12 months after the date when both milestones have happened?

Answer: P10, P50, and P90, in ECI points.

Resolution: This is a conditional question. It will resolve only if both milestones of Q1.4 happen on or before Dec 31, 2030. If not, the question will not resolve.

- Start date: the date when both milestones have happened, as in Q2.3.
- Internal lead: as in Q4.1, measured 12 months after the start date.

Please forecast how far the best AI inside the companies is ahead of the ECI frontier, one year after AI does most AI R&D on its own. Since the milestones of Q1.4 and the internal lead are partly hypothetical, this question is partly hypothetical too. If the required data exist when the question resolves, we will use them. If not, we will aim to resolve the question with the best evidence available (which may include, for example, asking a panel of AI experts to estimate the internal lead).

Background: Section 2.1 (lab definitions of automated AI R&D and the safety measures that labs link to them), Section 2.3 (why the ECI frontier counts only publicly available models, and our estimate of the current internal lead), Section 3.3 (qualitative evidence), and Section 5 (key cruxes).

B.5 Expert AGI and AI catastrophe

These questions ask how a fast year would change the chances of expert AGI and an AI catastrophe by 2050. We use the same definitions as FRI's Longitudinal Expert AI Panel (LEAP). Definitions:

- Expert AGI means that more than 50% of LEAP panelists agree that AGI exists (LEAP Wave 8). AGI is a commercially available AI system that (1) can outperform the 90th percentile professional human employee in every primarily non-physical occupation (based on 2025 performance), across all sectors, on at least 90% of the economically useful non-physical tasks that they perform, and (2) has an inference cost no more than 5× the cost of equivalent human labor.
- An AI catastrophe is a catastrophic event caused primarily by AI systems in which more than 10% of the people alive at the start of a five-year period die by the end of that period (LEAP Wave 9). This covers cases where AI is the direct or proximate cause of the deaths, and events that would not have occurred, or would have been extremely unlikely to occur, "but for" the substantial involvement of AI within one year before the event.

Q5.1 Expert AGI by 2050, with a fast year

If a fast year happens by Dec 31, 2030: what is the probability of expert AGI by 2050?

Answer: a probability from 0 to 100.

Resolution: This is a conditional question. It will resolve only if a fast year happens on or before Dec 31, 2030. If not, the question will not resolve. If a fast year happens, the question will resolve Yes if expert AGI exists on or before Dec 31, 2050.

Since the LEAP panel may not exist in 2050, this question is partly hypothetical. If the panel exists when the question resolves, we will use its answer. If not, we will aim to resolve the question with the best evidence available on whether such a panel would agree that AGI exists (which may include, for example, asking another panel of AI experts).

Background: Section 1 (introduction), Section 2.2 (definitions, including narrow and broad acceleration), and Section 5 (Step 3 of the key cruxes: whether the capability gains are broad).

Q5.2 Expert AGI by 2050, without a fast year

If no fast year happens by Dec 31, 2030: what is the probability of expert AGI by 2050?

Answer: a probability from 0 to 100.

Resolution: The same criteria as Q5.1, but the question will resolve only if no fast year happens on or before Dec 31, 2030. If a fast year happens, the question will not resolve.

Background: Section 1 (introduction), Section 2.2 (definitions, including narrow and broad acceleration), and Section 5 (Step 3 of the key cruxes: whether the capability gains are broad).

Q5.3 AI catastrophe by 2050, with a fast year

If a fast year happens by Dec 31, 2030: what is the probability of an AI catastrophe by 2050?

Answer: a probability from 0 to 100.

Resolution: This is a conditional question. It will resolve only if a fast year happens on or before Dec 31, 2030. If not, the question will not resolve. If a fast year happens, the question will resolve Yes if an AI catastrophe happens on or before Dec 31, 2050.

We will use the resolution criteria of LEAP Wave 9. In particular, for fatality and population data, we will rely on broadly recognized sources, such as Our World in Data, the Economist, and the United Nations. The five-year period must start on or after Dec 31, 2025 and end on or before Dec 31, 2050. If reasonable people disagree about whether such a catastrophe has happened, a panel of experts will resolve the question.

Please give your best forecast of whether such a catastrophe will happen by 2050. Do not change your forecast because of problems with resolution. For example, do not consider whether anyone will be able to resolve the question after a catastrophe.

Background: Section 1 (introduction, including the benefits and risks of RSI), Section 3.3 (qualitative evidence, including researchers' views on the risks), and Section 5 (key cruxes).

Q5.4 AI catastrophe by 2050, without a fast year

If no fast year happens by Dec 31, 2030: what is the probability of an AI catastrophe by 2050?

Answer: a probability from 0 to 100.

Resolution: The same criteria as Q5.3, but the question will resolve only if no fast year happens on or before Dec 31, 2030. If a fast year happens, the question will not resolve.

Background: Section 1 (introduction, including the benefits and risks of RSI), Section 3.3 (qualitative evidence, including researchers' views on the risks), and Section 5 (key cruxes).