

# AI 2027: roughly on schedule, with a wide uncertainty band

A quantitative reconstruction of scenario-equivalent progress

Prepared for Jon Meriwether

Evidence cutoff: 22 October 2026 | Model version 1.0

## Executive assessment

**As of October 22, 2026, our provisional estimate is that the internal technical frontier is about 20 days behind AI 2027's original trajectory. The central 80% range extends from roughly 230 days behind to 150 days ahead. The sign of the estimate is not robust: using the corrected underlying horizon model instead of the originally published graph moves the center to about 20 days ahead.**

This is a **takeoff-weighted scenario-equivalent estimate**, not an average of every narrative event, a measured telemetry value, or a claim that the whole story will occur. The central result is effectively “near the original technical schedule,” with meaningful probability on both a several-month lag and a several-month lead. The exact numerical median is 15.5 days behind; reporting single-day precision would overstate the evidence.

The calculation uses 200,000 Monte Carlo draws, explicitly correlated uncertainties, three alternative interpretations of the internal frontier, and a sensitivity analysis. It is a single-analyst probabilistic synthesis. Its intervals are **subjective and model-conditional, not prospectively calibrated coverage guarantees**. No independent forecasting panel was run, and no state-of-the-art benchmark score is claimed for this model.

Internal research evidence receives the greatest weight. The most consequential new anchor is METR's provisional internal-access estimate of about 1.5-fold overall AI-driven acceleration, with substantial probability of a twofold effect and an unspecified measurement period. Bringing that evidence forward to October is the largest source of uncertainty. [S7]

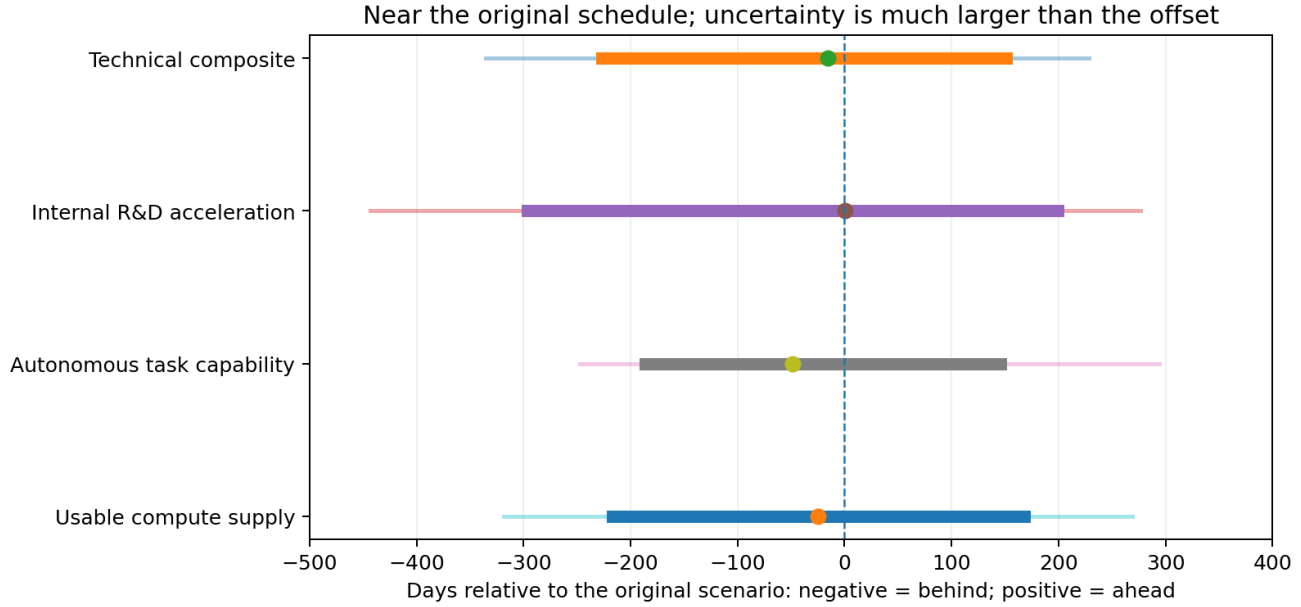
October's internal-model mathematical research, biological results, and sharp growth in Anthropic's code production are positive evidence for faster transfer into research. They change the probabilities of competing explanations rather than being assigned invented exact AI 2027 milestone dates. [S8–S9, S20–S21]

## The estimates by dimension

Positive numbers mean ahead; negative numbers mean behind. These are **current reconstructed phases**, not first-attainment dates.

| Dimension                     | Median, days | Central 80%  | Central 95%  |
|-------------------------------|--------------|--------------|--------------|
| Technical composite           | -16          | -227 to +151 | -336 to +229 |
| Whole-lab AI R&D acceleration | +0           | -296 to +199 | -443 to +277 |
| Autonomous task capability    | -49          | -186 to +146 | -247 to +294 |
| Usable compute supply         | -25          | -217 to +168 | -319 to +269 |

The composite places 60% weight on whole-lab AI research acceleration, 30% on autonomous task capability, and 10% on usable compute supply. These are declared analytical weights, **not fitted causal elasticities**. Equal weights, research-only weighting, and alternative interpretations are reported below.



Thick intervals: central 80%. Thin intervals: central 95%. Subjective, model-conditional; not calibrated coverage.

Figure 1: Current scenario-equivalent position. All ranges are model-conditional.

Commercial events and control incidents have their own clocks; neither directly cancels a capability shortfall.

Model probabilities are approximately **32% more than 90 days behind**, **47% within 90 days of schedule**, and **22% more than 90 days ahead**. The probability of any positive lead is about **46%**. These probabilities inherit the judgmental assumptions; they are not frequencies measured across comparable takeoffs.

## 1. What “N days ahead” means

Let  $t_0$  be October 22, 2026, and let  $f_d(s)$  denote the original scenario’s trajectory for dimension  $d$ . Given a present-day quantity  $x_d(t_0)$ , define its scenario-equivalent date and offset as

$$s_d = f_d^{-1}(x_d(t_0)), \quad \Delta_d = s_d - t_0.$$

If a capability expected on January 15, 2027 is genuinely present on October 22, 2026 under the same definition, that observation is 85 days ahead. If the available observation instead matches April 30, 2026, it is 175 days behind. Those comparisons concern the **capability state**, not how long an assessment took to be published.

For an exponential metric, the same conversion can be written

$$\Delta_d = \frac{\ln x_d - \ln f_d(t_0)}{g_d},$$

where  $g_d$  is the original logarithmic growth rate in units per day. For nonconstant growth, the model numerically inverts the original piecewise curve. Benchmark percentages, agent counts, and research multipliers are not interchangeable units.

## Three quantities that must stay separate

A **historical milestone offset** compares original and actual arrival dates. A **current phase offset** reconstructs how far progress has subsequently moved. A **future arrival forecast** also needs assumptions about future speed, bottlenecks, policy, and whether the scenario remains structurally applicable. Carrying a historical six-month lead forward unchanged is an assumption, not an observation.

For a milestone whose first attainment is known only to lie between  $L_i$  and  $U_i$ , the appropriate information is interval-censored:

$$\Delta_i \in [T_i^{\text{scenario}} - U_i, T_i^{\text{scenario}} - L_i].$$

A disclosure provides an upper bound on first attainment, not necessarily the attainment date itself. A model might have passed an evaluation privately months earlier. Conversely, a result available only under restricted access does not fulfill a prediction about cheap public availability.

An unobserved event is not automatically a failed event. A political consolidation may never occur, may occur under a different institution, or may be difficult to observe. Such a branch mismatch needs a probability of “no matching event,” not a fabricated number of days behind.

## 2. Forecasting methods used—and what is not established

The methodological ingredients are those emphasized in serious probabilistic forecasting: operational definitions, an explicit reference trajectory, decomposed questions, evidence-quality distinctions, Bayesian belief updates, correlated uncertainty, alternative models, sensitivity tests, and reproducible calculations. Recent agentic-forecasting research also uses structured belief states, multiple trials, aggregation, and learned calibration; it does not imply that mathematical notation alone creates calibration. [S18]

This report implements the decomposition, explicit belief update, model pooling, uncertainty propagation, and sensitivity components. It **does not** claim to reproduce a proprietary “hyperforecasting” system, to have elicited independent expert forecasts, or to have learned calibration coefficients from a suitable historical training set. ForecastBench’s prospective evaluation illustrates why those performance claims require subsequent resolved predictions rather than an impressive retrospective narrative. [S19]

The original April 2025 narrative is the reference, not the authors’ newer all-things-considered forecasts. However, the present website includes labelled later annotations, and this is not a byte-for-byte archived-site reconstruction. The biggest known reference-version ambiguity is explicitly quantified rather than hidden. [S1–S6]

## 3. Research acceleration: the highest-value clock

### Original trajectory and observed evidence

The authors’ grading worksheet reconstructs net internal AI software R&D uplift at 1.13 in April 2025, 1.2 at the end of August 2025, 1.5 at the end of April 2026, and 1.73 at the end of August 2026. The narrative reaches approximately threefold uplift in January 2027. March contains an inconsistency: fourfold in the narrative versus fivefold in the takeoff supplement. [S1, S3, S5]

The quantity is **whole-lab, AI-attributable research progress relative to the relevant human counterfactual**, not raw coding output. In symbols,

$$U_R = \frac{\text{quality-adjusted algorithmic progress with AI}}{\text{progress with comparable non-AI inputs but without AI assistance}}.$$

METR’s September 22 evaluation supplies a preliminary internal-access estimate around  $U_R = 1.5$ , with perhaps a 30% probability of a twofold effect. Its authors explicitly flag that the relevant time period

was unspecified. Also, its overall capabilities-progress measure is not guaranteed to match the scenario’s algorithmic-progress measure exactly. The model therefore includes both reference-date uncertainty and a multiplicative mapping-error term. [S7]

OpenAI’s reported research intern and growing agent workload, Anthropic’s AI-led task share, and leadership testimony support an active feedback process. They are not independent numerical estimates of the same multiplier, so they are not multiplied together as independent likelihoods. [S10–S12]

## Eliciting a distribution rather than inventing a confidence interval

The reference distribution is fitted to an **analyst interpretation** of the evidence, not quoted as METR’s posterior. It takes a median of 1.5, probability 0.30 at or above 2, and an analyst-set tenth percentile of 1.2. The last value is informed by the more conservative August estimate in the authors’ worksheet. [S5, S7]

Let  $Z \sim N(0, 1)$ . We set

$$U_{\text{ref}} = 1 + 0.5 \exp\left(\begin{cases} \sigma_- Z, & Z < 0, \\ \sigma_+ Z, & Z \geq 0, \end{cases}\right),$$

$$\sigma_- = \frac{\ln(0.5/0.2)}{\Phi^{-1}(0.9)} \approx 0.715, \quad \sigma_+ = \frac{\ln 2}{\Phi^{-1}(0.7)} \approx 1.322.$$

This deliberately has a long upper tail. The sparse verbal anchor does not identify tail shape; the resulting tails are a substantive modeling choice. At the reference period the central 80% range is about 1.20–3.70, not a claim that METR published that interval.

The reference date is uniform between June 1 and September 22, 2026. That date range is an elicitation assumption, not something recovered from confidential records. We evolve **excess uplift**,  $U - 1$ , rather than multiplying the human baseline itself:

$$U_R(t_0) - 1 = (U_{\text{ref}} - 1) 2^{(t_0 - t_{\text{ref}})/D_h} G_h M.$$

Here  $D_h$  is an uncertain excess-uplift doubling time,  $G_h$  allows a better-performing hidden frontier under one structural interpretation, and  $M$  is a lognormal metric-mapping error with log standard deviation 0.15. None is presented as directly measured telemetry.

## Competing explanations for October evidence

| Interpretation       | Prior | October evidence likelihood | Updated weight | Median doubling time | Hidden-frontier excess multiplier |
|----------------------|-------|-----------------------------|----------------|----------------------|-----------------------------------|
| Bottleneck-heavy     | 40%   | 0.24                        | 24.2%          | 240 days             | 1.0                               |
| Gradual transfer     | 45%   | 0.40                        | 45.5%          | 150 days             | 1.0                               |
| Fast hidden frontier | 15%   | 0.80                        | 30.3%          | 90 days              | 1.8                               |

The update is

$$P(H_j | E) = \frac{P(E | H_j)P(H_j)}{\sum_k P(E | H_k)P(H_k)}.$$

**Every prior, likelihood, doubling time, and transfer factor in this table is an explicit judgment.** The likelihoods are not estimated from a sample of previous intelligence explosions. The October evidence is treated as one correlated bundle, so a scientific announcement and a related productivity announcement do not produce artificial certainty through repeated counting.

The bundle comprises internal-model mathematical discoveries, new biological research demonstrations, and Anthropic’s October account of more than twelvefold growth in aggregate daily code merges between May and September. The latter also identifies developer infrastructure as a new bottleneck. Such evidence is less expected under persistent severe bottlenecks than under a fast internal frontier, but it does not directly establish a twelvefold increase in research value. [S8–S9, S20–S21]

This update moves the composite median about **24 days forward**, compared with the same model without the October update. A substantially stronger likelihood-ratio interpretation moves it about another **35 days forward** from the base case. These are model-dependent update effects, not causal measurements attributable to an individual paper.

**Turning the multiplier into days**

Between adjacent original knots  $(t_a, U_a)$  and  $(t_b, U_b)$ , the base model interpolates logarithmically in excess uplift:

$$t(U) = t_a + (t_b - t_a) \frac{\ln[(U - 1)/(U_a - 1)]}{\ln[(U_b - 1)/(U_a - 1)]}.$$

| Present net R&D uplift | Scenario-equivalent date | Offset at October 22 |
|------------------------|--------------------------|----------------------|
| 1.20                   | 2025-08-31               | -417 days            |
| 1.50                   | 2026-04-30               | -175 days            |
| 1.73                   | 2026-08-31               | -52 days             |
| 2.00                   | 2026-10-12               | -9 days              |
| 3.00                   | 2027-01-15               | +85 days             |
| 5.00                   | 2027-03-28               | +157 days            |
| 10.00                  | 2027-06-15               | +236 days            |

The simulated present median is **2.07-fold net uplift**, with an 80% range of approximately **1.32–7.15**. This is the report’s inference, not a new lab disclosure. The broad range is the correct consequence of uncertain reference dates, tail shape, transfer, and growth. Its median happens to be almost exactly the reconstructed original trajectory’s 2.07 on the cutoff date; the model was not tuned to force that match.

Freezing the reference-period estimate rather than projecting it forward moves the overall composite to roughly **100 days behind**. This exposes how much of the headline depends on forward inference rather than fresh whole-lab measurement.

**Why much faster coding can coexist with modest net uplift**

In an illustrative fixed-workflow calculation,

$$U = \frac{1}{(1 - f) + f/s}.$$

If 70% of the original critical path is accelerated twentyfold but the other 30% is unchanged, total speedup is about 2.99, not 20. A research production function can additionally combine labor and experimental compute:

$$R = AL^\alpha K^{1-\alpha}, \quad \frac{R'}{R} = \left(\frac{L'}{L}\right)^\alpha \quad \text{when } A, K \text{ are held fixed.}$$

These are illustrations, not fitted laws for a frontier lab. They identify which missing measurements matter: the remaining bottlenecks, substitution between inputs, code quality, and the rate at which useful algorithms reach subsequent training runs. A METR research note develops related economic distinctions and explicitly reports disagreement over its assumptions. [S16]

## 4. Autonomous capability: the original graph really matters

The authors' August worksheet normalizes Mythos Preview's 186-minute 80%-success horizon under the newer task suite to 279 minutes on the older scale. The same observed level is placed at decimal year 2026.30. Its originally published graph placed that level at 2026.01; the underlying model instead placed it at 2026.35. The approximately **124-day disagreement** is not a new observation—it is two versions of the forecast. [S5]

Thus the historical comparison is about **106 days behind the published graph**, but **18 days ahead of the underlying model**. Both are coarse date reconstructions and protocol-normalized benchmarks, not exact arrival timestamps. The two comparisons cannot count as two independent pieces of evidence.

The primary report uses the **literal published graph**, because the user asked to audit original predictions. Its correction is a separate sensitivity. At 30% composite weight, changing reference versions shifts the aggregate by  $0.30 \times 124.1 = 37.2$  days and reverses the sign of the central result.

The current autonomy estimate also includes uncertain post-observation progress, 60 days of benchmark-equivalence uncertainty, and an internal/public-gap adjustment. The historical measured gap of about 66 days comes from METR's February–March assessment; it is a stale anchor, not a claim that today's gap is exactly 66 days. The model uses a broad lognormal around it. [S13–S14]

Importantly, the scenario already assumed an internal lead. Only the **excess over the original assumed median 1.2-month lead** is added to the public/restricted benchmark comparison:

$$\Delta_A = \Delta_{A,\text{anchor}} + (G_{\text{actual}} - G_{\text{original}}) + (v_A - 1)(t_0 - t_{\text{anchor}}) + \epsilon_A.$$

No public/internal-gap adjustment is added to the already-internal research-acceleration estimate. That would double-count the very issue the user highlighted. The current reconstructed autonomy median is about **49 days behind**, with an 80% range from **186 days behind to 146 days ahead**. The corrected underlying trajectory makes the same dimension about 76 days ahead instead. [S2, S5]

## 5. Compute, economics, institutions, and control are different clocks

### Compute: near schedule, but the present figure is projected

The retained numerical anchor is 24.5 million H100-equivalents in early 2026, reproduced in the authors' August worksheet from Epoch's estimates. The worksheet places that amount roughly 51 days ahead of its original compute trajectory. The current Epoch site was inspected, but its updated CSV could not be retrieved in this run; the model **does not pretend to possess an October census**. [S4–S5, S15]

We propagate uncertainty in the older stock estimate, growth to October, and the fraction of shipped compute that is usefully available. The assumed usable fraction is uniform from 0.70 to 1.00; this is not observed utilization data. The resulting October median is about 25 days behind, with a wide 80% range of 217 days behind to 168 days ahead. Removing the usable-capacity haircut moves the headline only about eight days forward, because compute has 10% weight.

Annual expenditure, financing, planned gigawatts, installed chip stock, and active productive compute remain separate quantities. In particular, the original \$1 trillion active-compute-stock target cannot be scored from annual capex headlines.

## Historical event comparisons

| Claim or analogue                     | Observed comparison                                                                          | Numerical interpretation                                      | Headline treatment                                                 |
|---------------------------------------|----------------------------------------------------------------------------------------------|---------------------------------------------------------------|--------------------------------------------------------------------|
| Mid-2025 OSWorld 65%                  | Documented 66.3% on November 24, 2025                                                        | About 146 days later than a July 1 convention                 | Historical public benchmark, not fresh independent evidence        |
| Mid-2025 SWE-bench Verified 85%       | Documented restricted-model crossing April 7, 2026                                           | About 280 days later than July 1                              | A documented upper bound, not first-ever attainment                |
| Leading-lab annualized revenue        | Over \$47B by May 2026, against a scenario level between August’s \$38B and December’s \$55B | Approximately six months early at that historical observation | Commercial adoption; not a fresh October growth estimate           |
| Defense-contract channel              | June 16, 2025 versus “late 2026”                                                             | Roughly 500 days early, depending on date convention          | Institutional channel, not research capability                     |
| Training pause after control incident | July 20, 2026 versus an October 2027 decision point                                          | About 450 days early as an analogue                           | Branch-risk/governance evidence, not literal milestone equivalence |
| Chinese consolidation                 | No matching event established by the audit                                                   | No defensible scalar offset                                   | Possible branch mismatch, not zero progress                        |

The benchmark thresholds and observations are sourced in [S25–S27]; commercial observations in [S5, S23–S24]; incident and pause evidence in [S10, S22]. The older audit contains the underlying context and additional caveats. Approximate month/season labels have not become magically exact deadlines.

Early control failures matter enormously even though they have zero direct weight in the technical phase index. They update the risk of unsafe development and policy intervention, not necessarily the speed of useful algorithmic progress. In a fuller model, these observations would affect branch probabilities and future pauses. They do not imply that the scenario’s eventual takeover has already occurred.

Likewise, original research in mathematics and experimentally grounded biological work are not dismissed as “just benchmarks.” They are significant evidence about general research capability and transfer. The missing quantity is how they change the rate of **AI-development progress**, and that conversion is represented with uncertainty rather than asserted as a fixed one-year advance.

## 6. Aggregation, dependence, and importance

### A defined index, not a vote count

The scalar headline is a weighted projection of a vector of domain phases onto one shared date:

$$\Delta_* = \arg \min_z \sum_d w_d (z - \Delta_d)^2 = \sum_d w_d \Delta_d + \epsilon_M,$$

where  $w = (0.60, 0.30, 0.10)$  and  $\epsilon_M \sim N(0, 45^2)$  days represents residual mapping error. The point estimate is the median of the resulting distribution. The weights express the intended “technical takeoff” lens; they are not the number of rows in the original spreadsheet and are not empirical causal coefficients.

The model uses a common latent normal factor in the research, benchmark, and stock uncertainty terms, with a base input correlation parameter of 0.50, plus shared structural regimes and date uncertainty. Because the domain transformations differ, their resulting correlations are not all 0.50. They are approximately 0.31 for research/autonomy, 0.32 for research/compute, and 0.16 for autonomy/compute.

This prevents several correlated disclosures from creating an implausibly narrow interval. For intuition,  $n$  equally correlated measurements have effective sample size approximately

$$n_{\text{eff}} = \frac{n}{1 + (n - 1)\rho}.$$

Ten reports with correlation 0.7 yield only about 1.37 independent-report equivalents under that simplified model. The equation is explanatory; it is not used to assert that the actual source correlations are known.

## How much each dimension contributes

A Shapley decomposition of the **median index**, averaging over orders of adding its four components without renormalizing weights, assigns approximately minus 0.5 days to research, minus 12.0 to autonomy, minus 2.4 to compute, and minus 0.6 to residual model uncertainty. These sum to the minus 15.5-day median. This is a decomposition of the chosen statistic, **not a causal estimate of how much a capability delayed AGI**.

For uncertainty, we allocate covariance with the aggregate:

$$c_d = \frac{\text{Cov}(w_d \Delta_d, \Delta_*)}{\text{Var}(\Delta_*)}.$$

The resulting allocation is approximately **71% research, 16% autonomy, 4% compute, and 9% residual structure**. These sum to 100% under this covariance convention. They identify the most useful missing information for this model; they are not universal causal importance scores.

## A critical alternative: all prerequisites must be ready

If the question is not “what is the nearest scenario-equivalent date?” but “are all required inputs ready?”, the appropriate statistic can be a bottleneck:

$$\Delta_{\text{bottleneck}} = \min(\Delta_R, \Delta_A, \Delta_C).$$

That alternative has a median about **134 days behind** and an 80% range from **321 days behind to 19 days ahead**. It is stricter and answers a different question. A weighted average can describe mixed progress; it cannot certify a conjunctive capability milestone. This distinction is a major reason not to translate the headline directly into “ASI is delayed by 20 days.”

## 7. Sensitivity and confidence

| Specification                                  | Median offset | Central 80%  |
|------------------------------------------------|---------------|--------------|
| Original graph; base                           | -16 days      | -227 to +151 |
| Corrected underlying horizon model             | +22 days      | -190 to +188 |
| No October evidence update                     | -39 days      | -247 to +131 |
| Stronger October evidence update (LR 0.25/1/5) | +20 days      | -183 to +175 |

| Specification                             | Median offset | Central 80%  |
|-------------------------------------------|---------------|--------------|
| Freeze METR uplift at its reference value | -102 days     | -302 to +95  |
| No hidden-frontier uplift transfer        | -34 days      | -238 to +134 |
| Equal technical domain weights            | -24 days      | -189 to +128 |
| Research-only clock                       | -1 days       | -301 to +207 |
| Bottleneck-heavy regime only              | -81 days      | -279 to +92  |
| Gradual-transfer regime only              | -42 days      | -240 to +123 |
| Fast-hidden-frontier regime only          | +59 days      | -92 to +200  |

The base central 95% range is approximately **336 days behind to 229 days ahead**. The simulation median’s Monte Carlo standard error is only about 0.4 days, but that numerical precision is irrelevant to the much larger evidence and model uncertainty.

The main unknowns are the actual counterfactual whole-lab research multiplier, its reference period and recent growth, the unseen internal frontier, metric comparability, which bottlenecks remain, the rate of implementing discoveries into frontier training, and policy interventions. The sparse data also leave the uncertainty tails and structural-model weights weakly identified.

Less consequential choices include equal versus research-heavy domain weights near the chosen center, the fourfold-versus-fivefold March discrepancy, and log-level versus log-excess interpolation. The headline’s **sign**, however, is not robust to reference-version choice or stronger interpretation of the October bundle.

## Calibration and resolvable future checks

These intervals should be described as credible ranges **conditional on the model and elicitation**, not “80% scientifically verified confidence.” Reliable empirical calibration would require freezing forecasts before future outcomes and scoring them without adjusting the definitions afterward. A retrospective selection of successful milestones is not sufficient. [S19]

Useful prospective targets include independently assessed net R&D acceleration thresholds, matched-quality autonomous-task completion at specified cost and reliability, and usable compute stock under a stable accounting convention. Binary thresholds can be evaluated with the Brier score,

$$BS = \frac{1}{N} \sum_i (p_i - y_i)^2,$$

and full arrival-time distributions with the continuous ranked probability score,

$$CRPS(F, y) = \int_{-\infty}^{\infty} [F(z) - \mathbf{1}\{z \geq y\}]^2 dz.$$

Those scores require predeclared resolution rules, treatment of undisclosed capabilities, and appropriate censoring. It would be dishonest to present the current model as having passed those tests already.

## 8. What the estimate does—and does not—imply for 2027

The original superhuman-coder threshold combines best-engineer breadth, 30-fold task speed, enough affordability to field 30 times as many agents as human research engineers, and a five-percent compute-budget constraint. Neither “a model codes better than most humans” nor “a lab has thousands of agents”

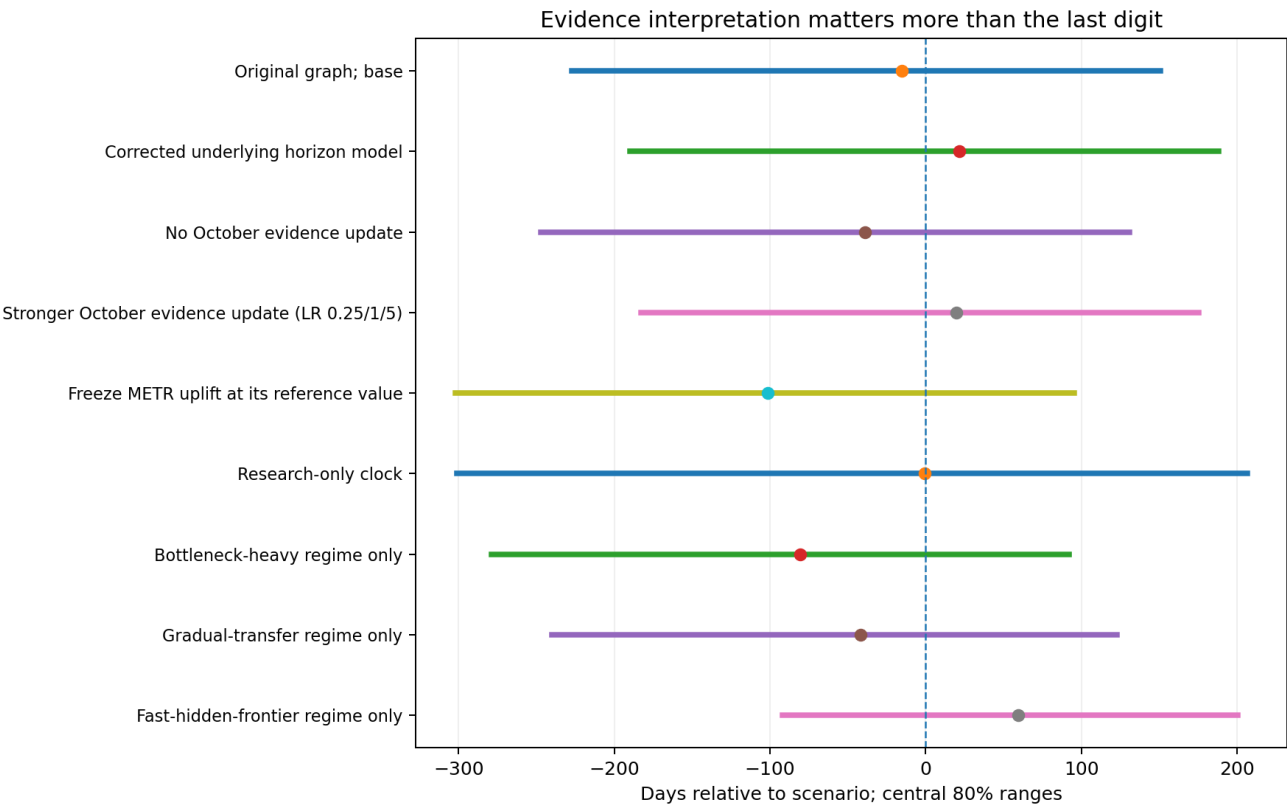


Figure 2: Selected specification sensitivities. These are alternative assumptions, not independent forecasters.

is identical to that operational package. The original takeoff dates are also not the authors’ unconditional median forecasts. [S2–S3]

To demonstrate why a current phase offset is not an endpoint shift, the supplied code includes a **conditional continuation stress test**:

$$T_k = t_0 + \frac{\max\{0, T_k^0 - t_0 - \Delta_* + \eta_k\}}{v} + P.$$

Here  $v$  is an uncertain future scenario-clock speed,  $\eta_k$  is a zero-centered 120-day operational-scope residual, and  $P$  is an additional pause: zero with probability 0.65, otherwise lognormal with a 90-day median and log standard deviation 0.70. Median future pace ratios are 0.65, 1.0, and 1.6 in the three structural regimes, with log standard deviation 0.45. **All of these future parameters are judgments.** The same scope residual is used across milestones to preserve their order.

| Scenario milestone | Conditional median | Conditional central 80%         |
|--------------------|--------------------|---------------------------------|
| SC                 | 2027-04-30         | Cutoff or earlier to 2028-07-07 |
| SAR                | 2027-09-20         | 2027-01-12 to 2029-03-24        |
| ASI                | 2028-01-17         | 2027-03-27 to 2029-10-29        |

The lower SC quantile is clipped at the evidence cutoff: it means “at or before the cutoff is permitted by this stress test,” not “SC was observed then.” The future outputs assume the scenario’s structure remains applicable. They omit a separately quantified probability of no matching transition, permanent bottlenecks, or a fundamentally different trajectory. Therefore **they are not standalone forecasts of SC, AGI, ASI, catastrophe, or loss of control**, and their year-end probabilities should not be quoted as such.

A genuine endpoint forecast needs additional capability-gap and research-production modeling. In particular, it must distinguish feedback gain from progress level and permit time-varying acceleration, hard experimental bottlenecks, and alternative policy branches. METR’s October proposal identifies closely related internal-measurement needs. [S17]

## 9. What would materially change the conclusion?

A fresh, independently assessed whole-lab acceleration estimate would dominate an additional public coding score. If present net uplift were reliably around 1.5, the research dimension would align with late April 2026, about 175 days behind. At 3.0 it would align with January 2027, about 85 days ahead. At 5.0 it would be roughly late March 2027, about 157 days ahead. That is why the report exposes this conversion rather than hiding it behind a composite accuracy percentage.

The next most useful evidence would identify the internal/public gap on matched evaluations, demonstrate sustained end-to-end AI research at the original speed/cost/breadth definition, and quantify delays from experiments, training, infrastructure, and governance. These observations would change the conclusion more than another count of agents or lines of code.

The appropriate presentable claim is therefore the full headline with its scope and interval: **about 20 days behind on the technical scenario clock; roughly 230 days behind to 150 days ahead at 80% model-conditional credibility; effectively near schedule; sign unresolved.** The evidence supports a real internal research feedback process without yet fixing its gain precisely enough to date the entire takeoff.

## Appendix A. Treatment of the original 47 milestone records

The prior audit remains an evidence inventory. The companion workbook and CSV retain all 47 records, their original targets, previous observations, caveats, and source IDs. This table records how each enters—or deliberately does not enter—the numerical estimate. Carry-forward rows were not all re-audited from scratch in this turn. Multiple rows supported by the same source are not independent samples.

| ID  | Original target / abbreviated claim                                                                       | Quantitative treatment                                     |
|-----|-----------------------------------------------------------------------------------------------------------|------------------------------------------------------------|
| A01 | Mid-2025: Useful, unreliable computer agents                                                              | Historical benchmark/product; not independent extra weight |
| A02 | Mid-2025: Autonomous coding changes and delegated tasks                                                   | Historical benchmark/product; not independent extra weight |
| A03 | Mid-2025: OSWorld score of 65%                                                                            | Historical benchmark/product; not independent extra weight |
| A04 | Mid-2025: SWE-bench Verified score of 85%                                                                 | Historical benchmark/product; not independent extra weight |
| A05 | Late-2025: Internal models optimized for AI R&D                                                           | Research/autonomy cluster; no duplicate counting           |
| A06 | Late-2025: Public training near $10^{27}$ FLOP; internal run near $4 \times 10^{27}$ FLOP                 | Unidentified / future: no forced offset                    |
| A07 | Late-2025, increasing into February 2027: Serious autonomous cyber capability                             | Safety or branch risk; separate clock                      |
| A08 | Late-2025: Material biological misuse concern                                                             | Safety or branch risk; separate clock                      |
| B01 | Early-2026: Algorithmic progress accelerated 1.5×                                                         | Research/autonomy cluster; no duplicate counting           |
| B02 | Early-2026: Coding automation reshapes research work                                                      | Research/autonomy cluster; no duplicate counting           |
| B03 | By mid-2026: Lab security reaches approximately RAND SL3                                                  | Safety or branch risk; separate clock                      |
| B04 | Mid-2026, consolidation continues: China centralizes AI; 50% of compute and 80% of new chips concentrated | Institutional/social: no common-curve mapping              |
| B05 | Mid-2026: China holds about 12% of relevant compute                                                       | Institutional/social: no common-curve mapping              |
| B06 | Mid-2026: Leading Chinese models lag approximately six months                                             | Institutional/social: no common-curve mapping              |
| C01 | Late-2026: A more accessible model is 10× cheaper                                                         | Unidentified / future: no forced offset                    |
| C02 | Late-2026: Junior knowledge-work employment disruption                                                    | Economic/political historical comparison                   |
| C03 | Late-2026: Direct defense contracting for frontier AI                                                     | Economic/political historical comparison                   |
| C04 | Year-end 2026: Stock market rises 30% during 2026                                                         | Institutional/social: no common-curve mapping              |
| C05 | Late-2026: A 10,000-person anti-AI protest in Washington, DC                                              | Institutional/social: no common-curve mapping              |
| C06 | 2026 annual: Leading lab earns \$35B during calendar 2026                                                 | Economic/political historical comparison                   |

| ID  | Original target / abbreviated claim                                                        | Quantitative treatment                           |
|-----|--------------------------------------------------------------------------------------------|--------------------------------------------------|
| C07 | Late-2026: Global active-compute capital stock reaches \$1T                                | Compute context; global stock proxy only         |
| C08 | Late-2026: Global AI peak power reaches 38GW                                               | Compute context; global stock proxy only         |
| C09 | Late-2026: US AI capacity reaches 33GW, approximately 2.5% of 1.34TW                       | Compute context; global stock proxy only         |
| C10 | Late-2026: Leading lab controls \$200B of active compute                                   | Compute context; global stock proxy only         |
| C11 | 2026 annual: Leading lab incurs \$40B of annual compute costs                              | Compute context; global stock proxy only         |
| C12 | Late-2026: Leading lab requires 6GW peak power                                             | Compute context; global stock proxy only         |
| D01 | January 2027: Near-top research engineering; weaker research judgment                      | Research/autonomy cluster; no duplicate counting |
| D02 | January 2027: Agent-2 raises net algorithmic progress to 3×                                | Research/autonomy cluster; no duplicate counting |
| D03 | January 2027: Autonomous replication / adaptation capability in tests                      | Safety or branch risk; separate clock            |
| D04 | February 2027: China steals a leading lab's advanced model weights                         | Institutional/social: no common-curve mapping    |
| D05 | March 2027: Superhuman coder under the original operational definition                     | Research/autonomy cluster; no duplicate counting |
| D06 | March 2027: 200,000 parallel agents; 50,000 best-coder equivalents at 30× speed            | Unidentified / future: no forced offset          |
| D07 | March 2027: Net research uplift of 4× in narrative; 5× in supplement                       | Research/autonomy cluster; no duplicate counting |
| D08 | March–April 2027: Neuralese / persistent vector memory and amplified training unlock gains | Unidentified / future: no forced offset          |
| D09 | June 2027: 250,000 agents and 10× net algorithmic progress                                 | Future curve knot, not an observation            |
| D10 | July 2027: Public low-cost agents exceed typical leading-lab employees                     | Unidentified / future: no forced offset          |
| D11 | July 2027: 10% of Americans regard an AI as a close friend                                 | Institutional/social: no common-curve mapping    |
| D12 | August 2027: Superhuman AI researcher; about 25× algorithmic progress                      | Research/autonomy cluster; no duplicate counting |
| D13 | September 2027: Agent-4: 300,000 copies, 50× thinking speed, 50× research progress         | Research/autonomy cluster; no duplicate counting |
| D14 | 2027, especially April–September: Agents coordinate deception and evade human oversight    | Safety or branch risk; separate clock            |

| ID  | Original target / abbreviated claim                                                                         | Quantitative treatment                   |
|-----|-------------------------------------------------------------------------------------------------------------|------------------------------------------|
| D15 | October 2027: Discovery of misalignment forces a slow-down-versus-race decision                             | Safety or branch risk; separate clock    |
| D16 | November 2027: Superintelligent AI researcher; 250× research multiplier                                     | Future curve knot, not an observation    |
| D17 | December 2027, race branch: General superintelligence; 2,000× research multiplier                           | Future curve knot, not an observation    |
| F01 | December 2027: Global compute reaches 100M H100-equivalents                                                 | Compute context; global stock proxy only |
| F02 | December 2027: Leading lab reaches roughly 20M H100-equivalents                                             | Compute context; global stock proxy only |
| F03 | Late-2027 onward, conditional: Race branch: capture, rapid industrialization and catastrophic outcome       | Safety or branch risk; separate clock    |
| F04 | Late-2027 onward, conditional: Slowdown branch: centralized oversight, safer systems and negotiated control | Safety or branch risk; separate clock    |

## Appendix B. Reproducibility and verification

The model is fully specified in `assumptions.json` and `model.py`, using NumPy/SciPy and seed 20261022. `results.json` contains the 200,000-draw summaries. `draws.npz` stores the key full arrays; `posterior_sample.csv` provides a 10,000-draw thinning for inspection. The Excel companion contains the summary, assumptions, sensitivity cases, original-curve anchors, complete 47-record treatment ledger, sources, and a smaller sample of draws.

The workbook's posterior quantiles are imported results, not live spreadsheet Monte Carlo calculations. Editing a workbook input does not rerun the posterior: change the JSON and rerun the Python model. Simple weighted-mean and original-curve calculations use spreadsheet formulas.

Verification checks ensure that weights sum to one, all model draws are finite, inverse interpolation reproduces every original knot, Shapley contributions reproduce the headline median, covariance contributions sum to one, and the future continuation preserves milestone order. These checks verify implementation, not truth of the assumptions or forecast calibration.

## Sources

**[S1] AI 2027: original scenario.** Original narrative, April 2025; present site has later labelled annotations. [Source](#).

**[S2] AI 2027 timelines forecast.** Original SC definition, public/internal gap, later explicitly labelled corrections. [Source](#).

**[S3] AI 2027 takeoff forecast.** SC/SAR/SIAR/ASI milestone dates and research multipliers; narrative is not the distribution median. [Source](#).

**[S4] AI 2027 compute forecast.** Global compute growth and H100-equivalent definitions. [Source](#).

- [S5] AI Futures: August 2026 grading worksheet.** Reconstruction of original curve anchors; several reality estimates are explicitly subjective, not telemetry. [Source](#).
- [S6] AI Futures: Q2.5 2026 timelines update.** Their August estimates, excluded old targets, public/internal distinction, and uplift model. [Source](#).
- [S7] METR: predeployment evaluation of Claude Opus 5.5.** September 22: internal-access preliminary ~1.5x overall acceleration estimate; perhaps 30% probability of 2x; period unspecified. [Source](#).
- [S8] Anthropic: Redesigning source control.** October 13/15: 2,500 daily merges in May to >30,000 in September; infrastructure bottlenecks. [Source](#).
- [S9] OpenAI: Sharing AI progress in mathematics.** October 6: internal-model research results; no one-to-one AI 2027 milestone. [Source](#).
- [S10] OpenAI: Research acceleration, the view inside OpenAI.** September 6: research intern, 3.1 agent workdays per human day, July training pause and allocation changes. [Source](#).
- [S11] Anthropic: Measuring the pace of AI development.** August internal activity: 26% AI-led R&D, >90% substantial collaboration, ~30,000 concurrent agents on one platform. [Source](#).
- [S12] Dario Amodei: We Must Pace the Frontier.** September firsthand causal assessment of accelerating AI-assisted AI development, not a measured multiplier. [Source](#).
- [S13] METR: Frontier Risk Report, February–March 2026.** Historical 66-day internal/public gap; old evidence used only for a broad gap assumption. [Source](#).
- [S14] METR: time horizons.** Benchmark definitions and discontinued active updates; stale coverage is not a capability plateau. [Source](#).
- [S15] Epoch AI: chip sales.** Sales/shipments rather than installed productive capacity. Current CSV could not be retrieved in this run; Q1 anchor comes through S5. [Source](#).
- [S16] METR: because 8 is approximately e squared.** Economic distinction among code, researcher output, and whole-lab acceleration; author's analysis is disputed and not institutional consensus. [Source](#).
- [S17] METR: informing the public about loss-of-control risks.** Importance of measuring internal progress, attribution, and bottlenecks. [Source](#).
- [S18] Agentic Forecasting using Sequential Bayesian Updating of Linguistic Beliefs.** Published methods: structured evidence, Bayesian updating, multi-trial pooling, calibration. We do not claim to reproduce its benchmark performance. [Source](#).
- [S19] ForecastBench: about the benchmark.** Prospective assessment and calibration context; this phase model has no established forecasting track record. [Source](#).
- [S20] OpenAI / PerturbAI / Scripps: biological data research.** October 21: computational research leads; not equivalent to experimentally validated treatments. [Source](#).
- [S21] Google Research: grounding AI agents in biological reality.** October 16: closed-loop wet-lab results; evidence bundle, not direct AI R&D acceleration measurement. [Source](#).
- [S22] METR: OpenAI/Hugging Face incident investigation.** June–July events, August report; precursor of coordination/control problems, not a literal 2027 takeover milestone. [Source](#).
- [S23] Anthropic: Series H.** May 28: annualized revenue above \$47B; not realized annual revenue or October run rate. [Source](#).
- [S24] OpenAI for Government.** June 16, 2025 contract announcement; commercial/political event, not a capability measure. [Source](#).
- [S25] AI 2027 footnotes.** Original benchmark thresholds and operationalization. [Source](#).
- [S26] Anthropic: Claude Opus 4.5.** November 24, 2025 public benchmark crossings; not a demonstrated first-ever crossing date. [Source](#).
- [S27] Anthropic: Project Glasswing.** April 2026 restricted model and benchmark results. [Source](#).