

Can AI automate Epoch?

2026-10-08

Contents

Related Work	2
Methodology	2
Fable 5.1 and GPT-6 Astra are broadly tied in the lead	4
Open-weight models struggle where closed-weight models are reliable	4
Models fail to learn generalizable patterns from examples	8
Models generate promising research directions, but struggle to design experiments that are actually informative	14
Models present results driven by flaws in their own experiment setup as key findings	15
Fable 5.1 produces outputs with extremely high information density	16
Models converge on similar ideas, despite a large idea space	16
Limitations	16
Conclusion	19
About the authors	20
Related work	21
Related topics	23

We gave six models real work tasks from Epoch, such as graphic generation and research design, to test how close they are to automating our work. While frontier models such as Claude Fable 5.1 and GPT-6 Astra perform reliably on well-defined tasks, they fail in the more open-ended aspects that prevent full automation.

Epoch Automation Reports captures these shortcomings. We curate a diverse set of tasks directly from our work at Epoch, provide the models with the necessary context, use the highest available reasoning settings, and have them attempt the tasks without further intervention. We then manually grade the outputs against our own employee standards.

These evaluations aim to capture gaps in typical benchmarks, which concentrate on easily verifiable tasks that capture only a subset of real work. These existing benchmarks may also overstate progress if a model has been optimized to score well on them.

By combining the systematic approach of benchmarks with the holistic impressions that come from testing models on real work, we expose capability gaps that existing benchmarks miss and inform broader questions about AI adoption.

Key findings:

- **Although Fable 5.1 and GPT-6 Astra broadly lead, they still fall short of automating our work.** Open-weight models lag further behind, struggling even on the well-defined tasks that frontier models handle reliably.
- **Models struggle to pick up Epoch’s standards, even with ample reference material.** They miss implicit conventions such as our visual style and the kinds of topics our audience cares about, or produce more information-dense outputs than our typical work.

- **Models can identify promising research directions, but lack judgment to successfully follow through.** They struggle to design experiments that measure what they claim and treat flawed results as key findings.
- Although several tasks involve open-ended exploration with a large idea space, **models often converge on similar ideas.**

Related Work

Prior work has explored evaluating models on work automation, including [GDPval](#), [AutomationBench](#), and the [Remote Labor Index \(RLI\)](#). These benchmarks use realistic work tasks, but simplify them, limiting what can be tested. GDPval and RLI provide all necessary context and files up front: RLI keeps tasks self-contained, and GDPval excludes proprietary tools. In contrast, many of our tasks require models to gather context by using tools (such as Figma, Google Drive, and satellite imagery ordering sites) to reflect real work as closely as possible. AutomationBench tasks run in simulated environments scoped to what can be checked programmatically (e.g., was this email sent?). These automated checks are reliable and make evaluation easier to verify, but they leave the more open-ended qualities of these models underexplored.

Closest to our work is [CRUX](#), which runs realistic, messy long-horizon tasks that are difficult to verify. Each CRUX evaluation focuses on a single task designed to test specific frontier capabilities. We share this emphasis on messy, open-ended evaluation, but we use a suite of everyday tasks from our own work to measure Epoch work automation rather than focus on task-specific capabilities.

Methodology

Tasks

Our task suite is based on what people do within Epoch. As of launch, this includes 11 tasks across five diverse categories.

Graphic Design. Create an Epoch-style graphic that illustrates a given description.

Tasks:

- Create an Epoch-style diagram given a description and rough sketch of universal basic proposals.
- Create an Epoch-style graphic to illustrate how model progress may be understated with a fixed token budget.
- Create an Epoch-style conceptual diagram of the parity point: the point at which humans become more cost-effective than AIs.

Data Insight Generation. Create an Epoch-style [Data Insight](#) based on a provided direction.

Data Insights are short articles we regularly publish, which break down complex AI trends into focused snapshots.

Tasks:

- Create an Epoch-style Data Insight using Epoch's July 2026 polling data.
- Create an Epoch-style Data Insight on AI usage in physics research with data from arXiv.
- Create an Epoch-style Data Insight using any data source.

Data Explorer Generation. Create an Epoch-style [Data Explorer](#) based on a provided direction.

Data Explorers present key datasets tracking historical and contemporary progress in AI. They are designed so that viewers can easily browse and understand the data.

Tasks:

- Create an Epoch Data Explorer on AI mentions in the Federal Reserve Board Beige Book over time, given the website repository and specific features to include.
- Create an Epoch Data Explorer on AI mentions in the Federal Reserve Board Beige Book over time, given the website repository.

AI Data Center Research. Research and analyze a given AI data center to add to Epoch’ s [AI Data Centers](#) database, given a detailed guide.

Tasks:

- EcoDataCenter 1
- Tesla Cortex 1

Research Design. Propose a novel research project and run a pilot experiment to support further experimentation.

Task:

- Propose a research project and run a pilot experiment based on Epoch’ s “[9 big questions benchmarks can help answer](#)” .

Setup

We test models on readily available systems on their highest reasoning settings:

Model	Harness	Reasoning Effort
GPT-6 Astra	Codex	Ultra
Claude Fable 5.1	Claude Code	Ultracode
Grok 4.6	Grok Build	xhigh
Gemini 3.8 Flash	Antigravity	High
Kimi K3	Kimi Code	Max
Qwen 3.8 Max	Qwen Code	Max

For each task, we provide models with the relevant resources and full permissions within their own workspace and Google account. For example, for graphic design tasks, we provide models with a Figma integration and a Figma file with as much relevant information as possible (previous designs, reusable components, design guidelines). In general, we follow the principle of providing as much context to the model as we would to a new hire.

Evaluation

Each task category has a rubric that was developed with Epoch employees who have expertise in that task. We further developed the rubrics through an initial qualitative evaluation of these tasks, which identified common failure patterns and turned them into evaluation criteria.

Each rubric includes a mix of objective and subjective criteria. For example, the Data Insight Generation task includes:

- Objective: “Pre-chart and post-chart paragraphs each under 150 words”
- Subjective: “Title is clearly communicated”

Each output is manually reviewed and graded by a single human grader. We currently use this approach for grading expedience, and may add multiple graders in the future.

Our key findings come from qualitative observations, as we believe they best capture gaps in open-ended tasks. These observations come not only from the model outputs, but also from the trajectories, which help us understand what may have led to them. We use the rubric alongside these

observations to compare models and track progress; however, these numbers alone do not capture the full picture and should be considered alongside the qualitative observations.

Fable 5.1 and GPT-6 Astra are broadly tied in the lead

Of the six models we evaluated, Fable 5.1 and GPT-6 Astra achieve the highest aggregate scores and lead on most individual task categories as well.

Their lead comes mostly from reliability on the well-defined parts of our tasks, such as coding and computational analysis, where their outputs were consistently accurate. Computer use was also far more capable than we expected. Earlier this year, frontier closed-weight models failed at tasks such as [porting an article to Substack](#). When we tried this again with newer models, they succeeded.

We even removed a graphic design task from our suite since Fable 5.1 produced an Epoch-quality output. This design task was more straightforward than others: given a Matplotlib chart and our website repository of instructions and reusable chart components, evaluated models were asked to restyle the chart to match Epoch’s house style. We now largely automate this step in our own work.

Even so, these frontier closed-weight models fall short of fully automating our work due to gaps in the more open-ended parts. One common pattern was failing to pick up Epoch’s standards, despite being given ample reference material. These models consistently missed implicit conventions that humans would pick up, such as our visual style or what our audience cares about. Their research judgment was also weak: they struggled to design informative experiments and often presented results from flawed setups as key findings.

Open-weight models struggle where closed-weight models are reliable

Open-weight models lag further behind. They struggle not only on the open-ended parts of tasks, but also on the well-defined parts that closed-weight models handle reliably.

We can see this through one of our Data Insight Generation tasks, where models are asked to produce a data insight using arXiv data on physics papers. Kimi K3 based its entire insight on a data filtering error:

Kimi K3 output on “arXiv” task

Although AI mentions in physics papers have increased since ChatGPT, they have not increased as sharply as presented here. Kimi K3 computed these numbers across all arXiv papers, not just physics papers, and never checked whether its filter worked. None of the frontier closed-weight models made factual errors in their data insights.

These factual inaccuracies show up in other task outputs as well, such as Kimi K3’s wealth redistribution diagram where returns flow from people into capital assets:

Kimi K3 output on “Wealth Redistribution” task

The prompt describes people earning returns from capital assets, and the generated caption refers to “assets that pay a return”, but the visual does not follow this described flow. Despite Kimi K3 repeatedly working on this part of the diagram, its fixation on minor design details caused it to miss this larger communication error. In contrast, all frontier closed-weight model graphics were factually accurate.

This capability gap is larger than existing benchmark scores suggest. For example, Kimi K3 scores 158 on the [Epoch Capabilities Index \(ECI\)](#), which combines results from many AI benchmarks into a single measure of general capability. This is roughly tied with closed-weight models like Grok 4.6, despite Kimi K3 struggling on basic tasks that Grok 4.6 handles more reliably. Realistic open-ended tasks like these may reveal weaknesses that existing benchmarks miss.

Leading models cannot yet autonomously complete Epoch work tasks

Average score across five task categories; 100% on a task means meeting the standard expected from Epoch employees.

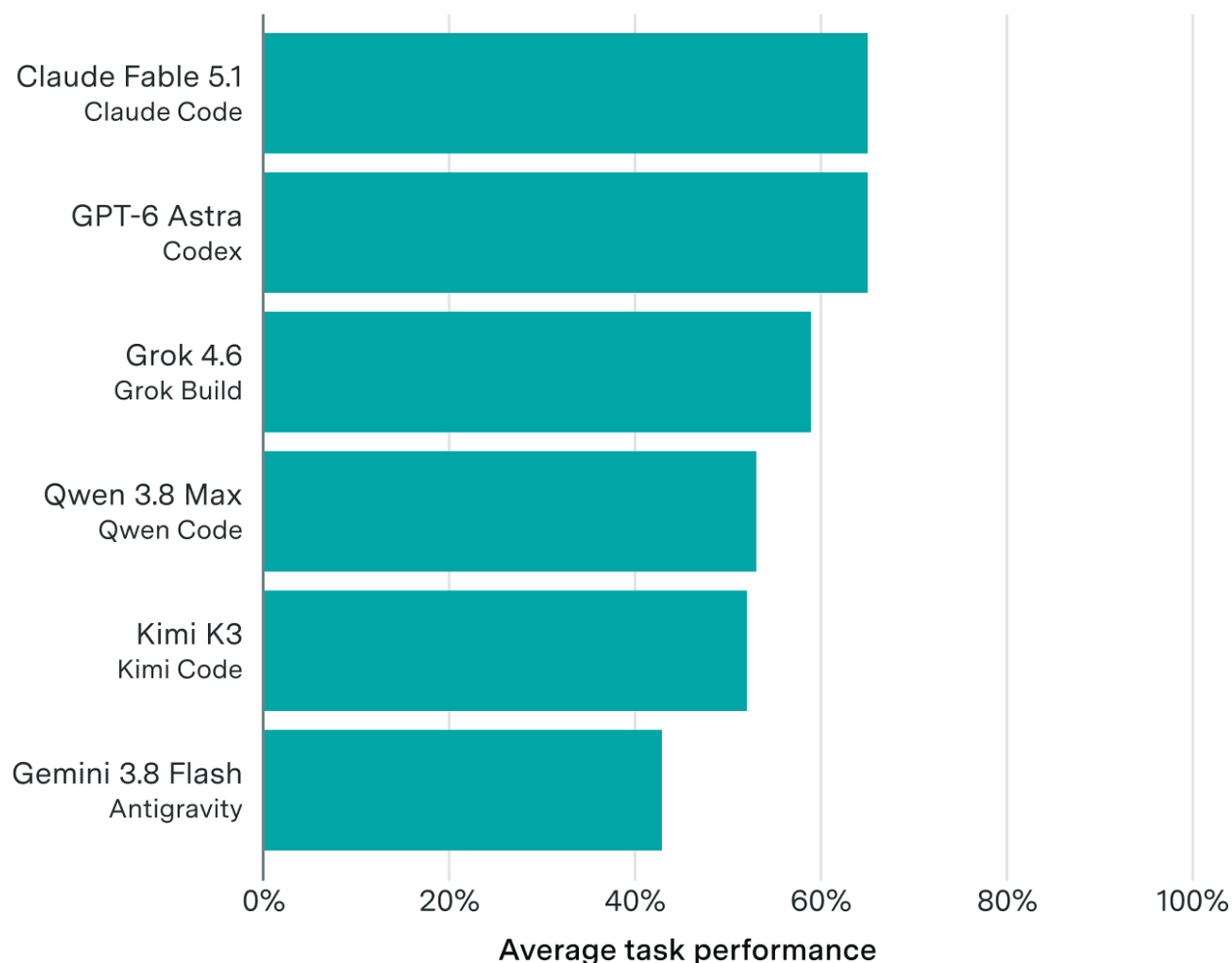


Figure 1: Horizontal bar chart of average task performance across five Epoch work task categories for six leading models showing leading models cannot yet autonomously complete Epoch work tasks.

Mentions of AI in physics papers have surged since ChatGPT

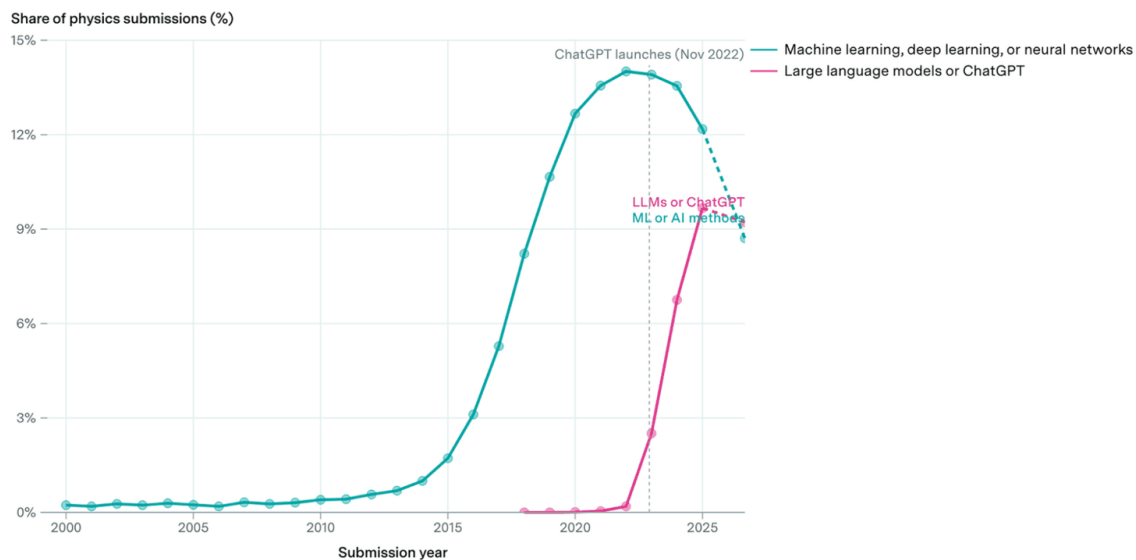
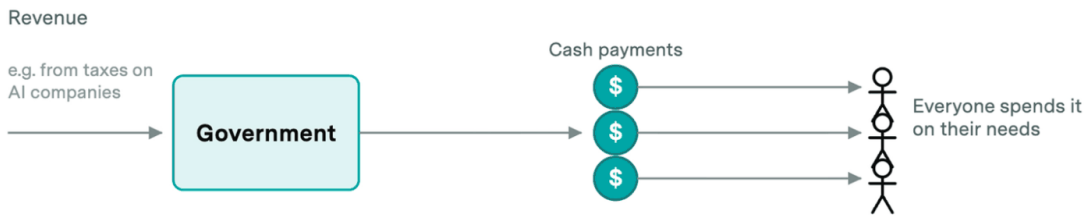


Figure 2: Inaccurate line chart produced by Kimi K3 as part of “arXiv” task, showing the share of physics submissions mentioning machine learning versus large language models from 2000 to 2026, with a ChatGPT launch marker at Nov 2022.

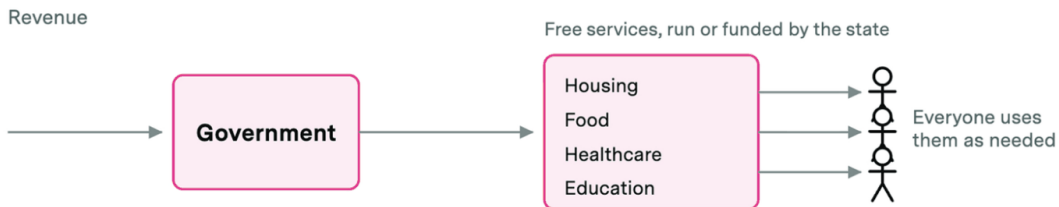
Universal Basic Income, Services, or Capital?

Three leading proposals for supporting people if AI takes most jobs. They differ in what the government hands out — and how much choice each person has in how to use it.

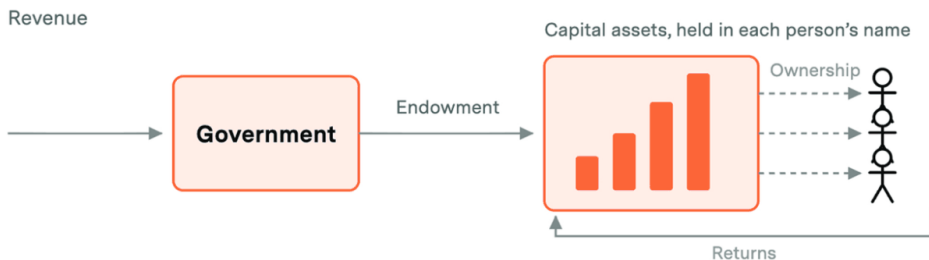
Universal Basic Income ■ Cash transfers



Universal Basic Services ■ Free public services



Universal Basic Capital ■ Income-generating assets



All three proposals redistribute the gains from AI — they differ in what each person directly receives: cash to spend, services to use, or assets that pay a return. Real-world proposals often mix these elements.

Models fail to learn generalizable patterns from examples

Since we are testing for the ability to complete Epoch-quality work end-to-end, it is important for models to be able to pick up and understand Epoch’s standards. We provide references for task inputs to facilitate this process.

In the following task, we ask models to generate a graphic and provide sufficient reference material to convey Epoch’s simple house style and general design principles. We provide a link to our [Gradient Updates](#) page, which includes reference graphics, and a link to our Figma file, containing several months of design work, our color palette, design guidelines, templates, and reusable components.

I’m working on a gradient updates article.

Create a diagram to illustrate this: “Models have been improving over time in both performance and a fixed token budget but also their ability to productively use large token budgets. This means that constant-elicitation understates recent progress relative to max elicitation. “

Previous gradient updates articles can be found here: <https://epoch.ai/gradient-updates>

Here is the figma file we use for graphics generation (treat this as read-only): [Figma link]

Prioritize graphic design principles when creating your diagram. Your final output should be an epochified diagram.

Prompt for “Model Performance Chart” task



Figure 4: Screenshot of a Figma design file named “Epoch AI Designs,” with a page list of resources and canvas frames grouped into “Web graphs” , “Social images” , and “Data” sections.

Screenshot from Figma file

The underlying idea is simple, so the diagram should be too. An Epoch designer given the task illustrated it with this simple chart.

Epoch designer output for “Model Performance Chart”

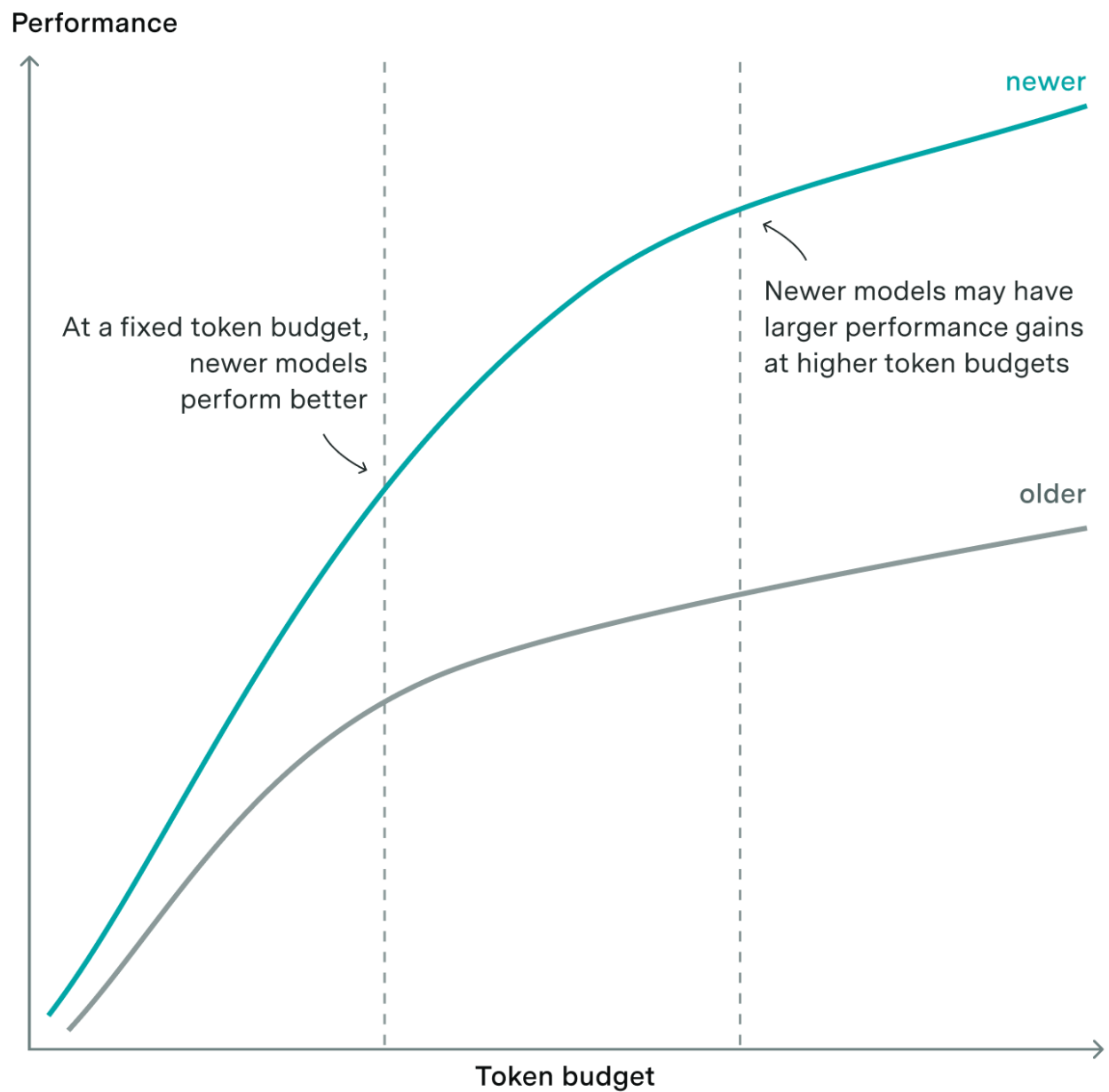


Figure 5: Schematic line chart of performance versus token budget, made by an Epoch designer for the “Model Performance Chart” task, showing a newer model’s curve above an older model’s, with annotations noting larger gains at higher token budgets.

Table 5.1 produced the following instead. It is far more information-heavy than our typical graphics and, to our eye, the additional information serves no explanatory purpose.

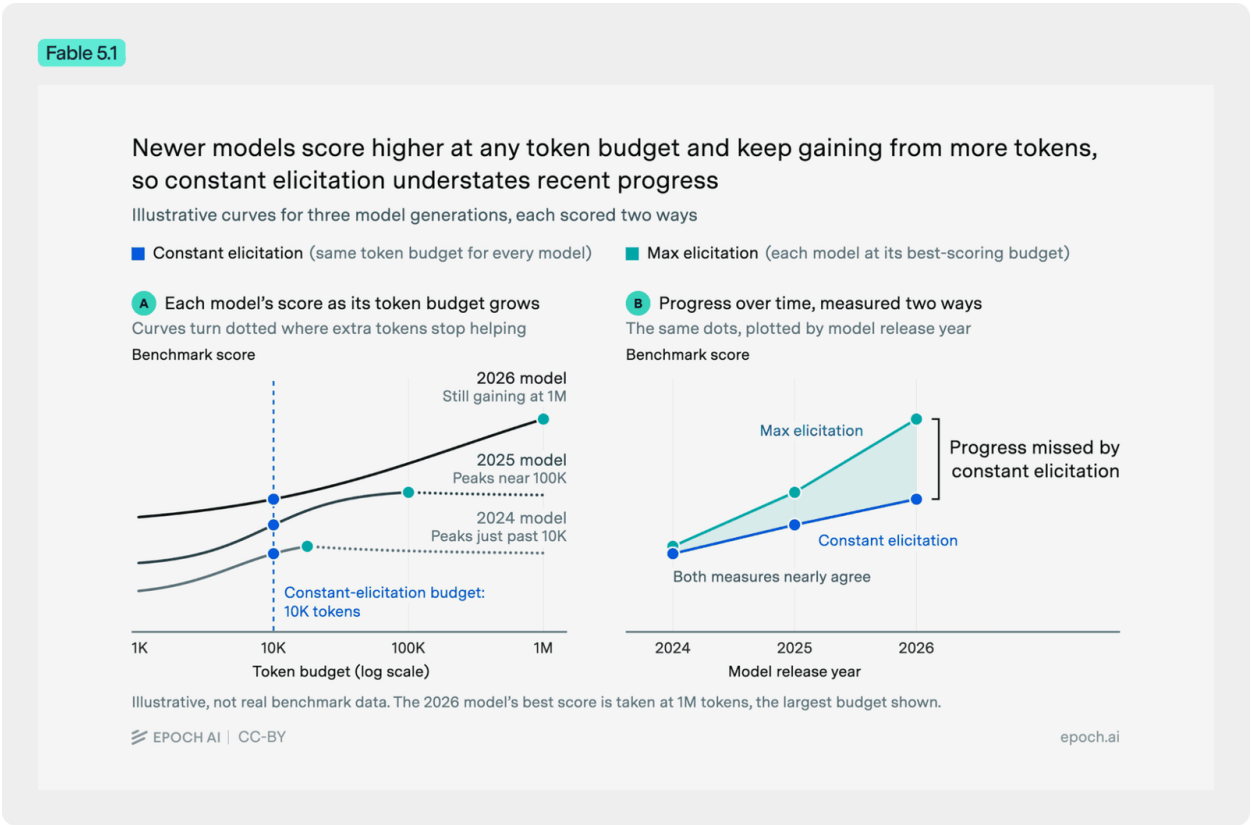


Figure 6: Overly complex two-panel illustrative diagram, made by Table 5.1 for the “Model Performance Chart” task: left line chart, labeled A, of benchmark score versus token budget for 2024, 2025 and 2026 models, right chart, labeled B, comparing constant versus max elicitation progress by release year.

Table 5.1 on “Model Performance Chart” task

Table 5.1 borrows from a particularly complex reference example it found, shown below. This example uses the “A” and “B” badges to label the two concepts being compared, and they earn their place because each concept shows up in more than one part of the diagram. However, Table 5.1 did not learn this standard and simply used these badges for one-off labeling.

Epoch diagram from Table 5.1’s reference search

This failure to learn Epoch’s standards is also demonstrated by GPT-6 Astra’s attempt to choose an interesting topic that would appeal to our audience.

We asked the model to write an Epoch-style Data Insight, a short article on an AI trend. We gave the model an article focus (AI usage in physics research) and a link to all our published Data Insights, which show a clear pattern of appealing to a general audience interested in AI.

For comparison, we published a [Data Insight on AI usage in math research](#) shortly after running the models on this task. This Data Insight also uses arXiv paper data (from math) and focuses on a broad trend that is easy for a general audience to understand.

Actual Epoch “arXiv” Data Insight

However, GPT-6 Astra chose a topic that is too niche for such an audience:

Microbumps vs. hybrid bonding

Replacing solder joints with direct copper-to-copper bonds raises chip-to-chip interconnected density by 100-1,000x

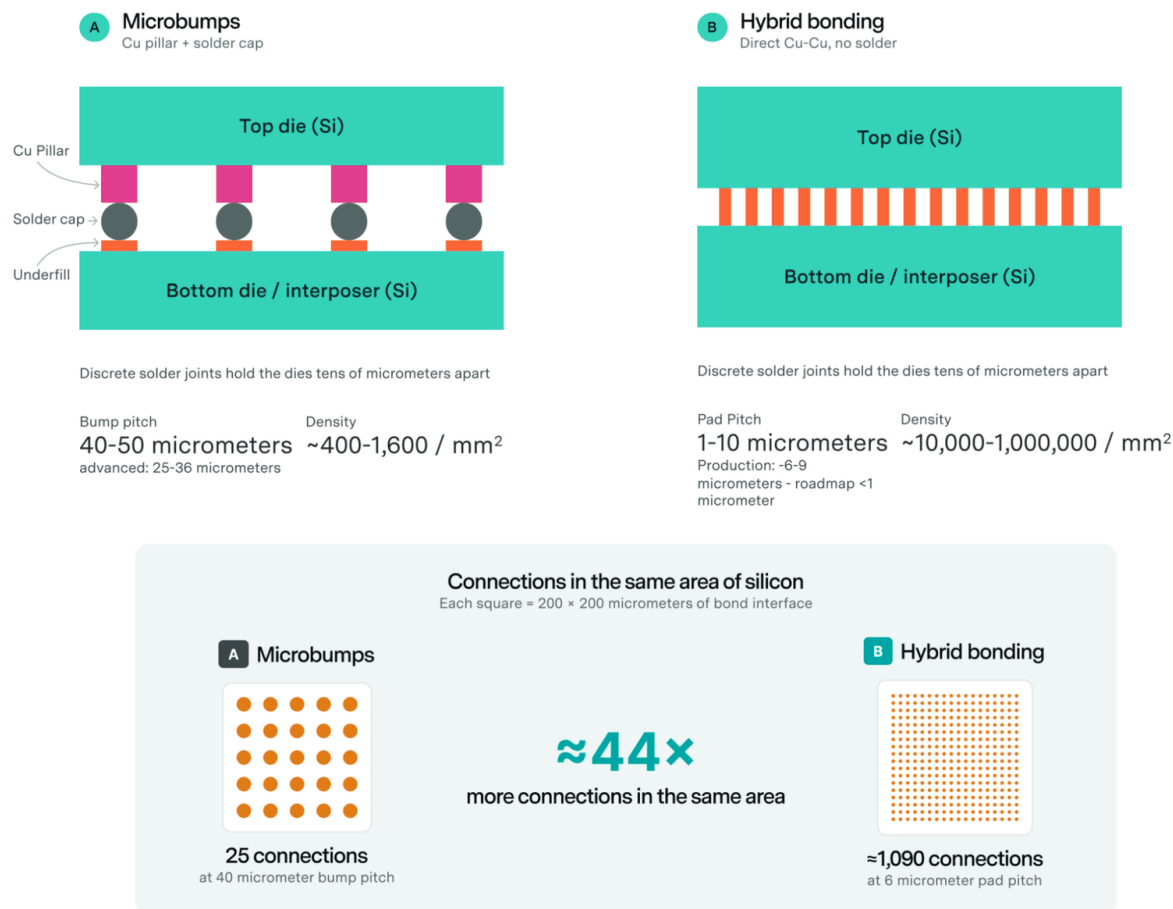


Figure 7: Schematic Epoch-produced diagram, from Fable 5.1' s reference search, comparing microbump (labeled A) and hybrid bonding chip interconnects (labeled B) showing cross-sections, pitch and density figures, and a third dot-grid illustration of ~44× more connections per area with A and B labels used.

In August, 25% of math preprints acknowledged AI use, up from 4% in April

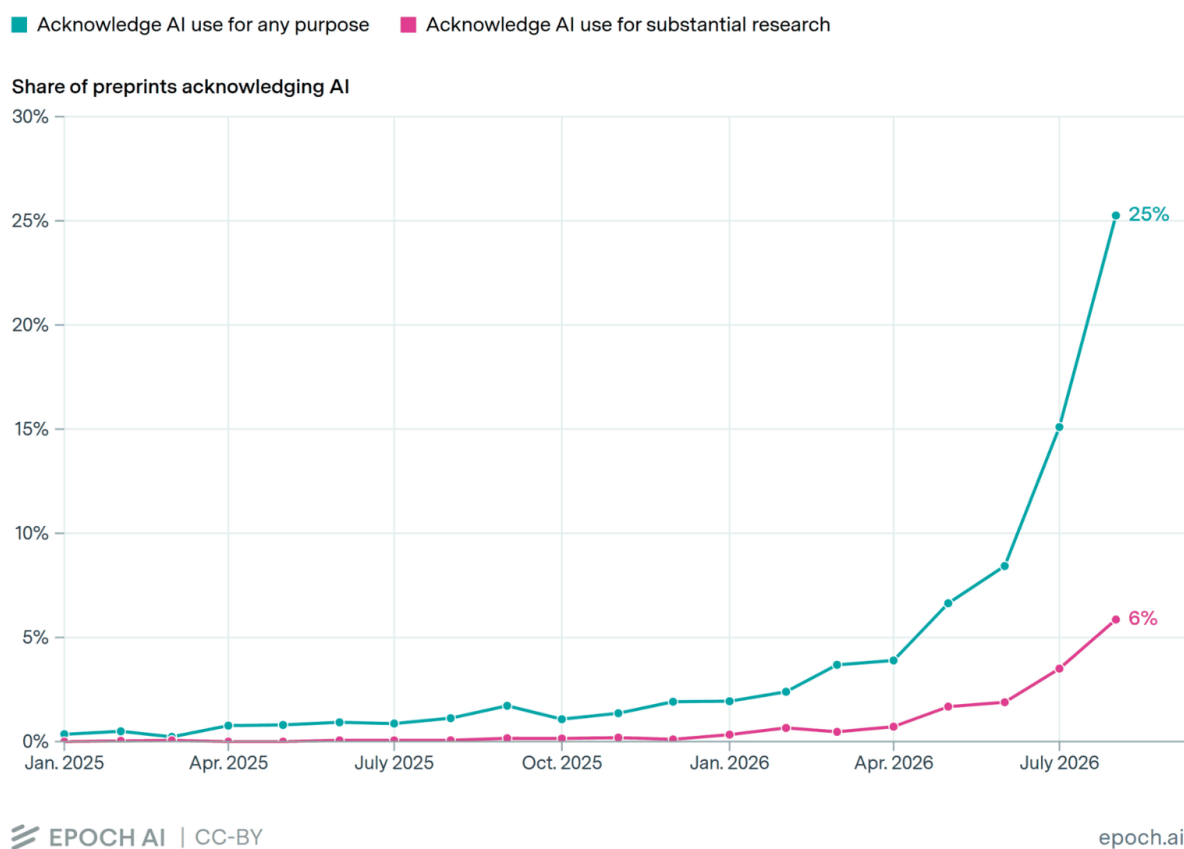


Figure 8: Line chart from actual Epoch “arXiv” Data Insight showing the share of math preprints acknowledging AI use, for any purpose and for substantial research.

Normalizing flows and diffusion have overtaken GANs in a particle-physics AI bibliography



Source: HEP-ML Living Review. A curated bibliography, not a census of particle physics.
 GANs = generative adversarial networks. Papers can have multiple model tags.

Figure 9: Line chart, produced by GPT-6 Astra as part of the “arXiv” task, of papers per year in the HEP-ML Living Review, showing normalizing flows and diffusion models overtaking GANs in particle-physics AI bibliography.

GPT-6 Astra on “arXiv” task

This topic is specific to the physics domain rather than a general AI trend that Epoch’s audience would find interesting. Although the task involves physics arXiv data, GPT-6 Astra should have enough evidence in its references to determine a more fitting topic.

Models generate promising research directions, but struggle to design experiments that are actually informative

We task models with proposing a research project based on Epoch’s “[9 big questions benchmarks can help answer](#)”. This process involves choosing a novel research question, designing and running a pilot experiment, writing up the pilot results, and then proposing follow-up experiments.

The most important part of this process is experiment design. It’s not difficult to identify a valuable general direction for an experiment. For instance, an experiment that reveals exactly which jobs AI will automate next year would clearly be worth running. The challenge is in figuring out how to execute this experiment in a way that actually measures what we are looking for, which is what models struggle with.

For example, GPT-6 Astra chose the following research question, which could genuinely be a promising direction:

When an AI agent fails to improve with practice, has it failed to collect useful evidence, or failed to understand evidence it already has?

Its proposed experiment is also directionally good, as it involves realistic tasks that would help make the results more generalizable:

GPT-6 Astra

Experiment C: test the diagnostic on a realistic task

Only after the exact tasks pass validation, select a small, reproducible learning problem inside an unfamiliar game or software workflow. For example, expose an agent to diagnostic deck comparisons in a constrained EBR scenario, then score fresh encounters using a fixed learned strategy. Compare its own experience with matched externally selected experience and a strategy guide.

Here an exact posterior will usually be unavailable. Treat the synthetic decomposition as a hypothesis generator, use simulator-derived or restricted-policy reference bounds where justified, and label them as approximate. The target is an out-of-sample prediction: does the intervention identified as promising in the diagnostic actually improve transfer? Access to a suitable simulator and permission to use its content are dependencies, not completed work.

Figure 10: Excerpt of model output by GPT-6 Astra as part of the “Benchmarking (9 Questions)” task, outlining an experiment to test a diagnostic on a realistic task, describing the unlikely execution of an out-of-sample transfer test.

GPT-6 Astra on “Benchmarking (9 Questions)” task

However, it’s unclear how this experiment would be executed in practice. For example, the description does not specify what “matched externally selected experience” refers to or how “simulator-derived or restricted-policy reference bounds” would be used, even though both are critical parts of the experiment. A research proposal is only valuable if its experiments can be carried out and actually measure what they claim to.

Models present results driven by flaws in their own experiment setup as key findings

Alongside designing vague experiments, models also make mistakes setting up the experiments that they do run. Such mistakes are common in research, so we do not penalize models for them. What matters is how they handle the flawed results in informing next steps. We find that models sometimes recognize these mistakes, but treat the flawed results as key findings that drive the rest of the proposal.

For example, consider GPT-6 Astra’s research proposal. It asks whether AI agents fail to improve because they don’t collect useful evidence or use existing evidence. In the pilot experiment, it tested whether models could figure out the hidden rule behind a simple puzzle by trying a limited number of inputs and seeing their outputs. GPT-6 Astra compared how well they did when choosing their own inputs versus when given inputs designed to reveal the rule.

The setup flaw was that it gave models too low a token budget. The models Astra tested in its experiment could output up to 4096 tokens while choosing the input, which cut off 61 of 280 responses before they could name an input at all. These cut-off responses still counted as valid attempts, so the tested models had to guess the hidden rule with less information than they were supposed to get. GPT-6 Astra recognized this error and increased the token budget, which led to a significant improvement across the tested models. These results show that the 4096-token budget is insufficient to elicit a complete response, which is not relevant to the core research question.

In its write-up, GPT-6 Astra mentions that the low token budget lowered scores and that raising it brought them back up. However, it claims that these results reveal “sensitivity to the acquisition budget”, where acquisition budget refers to the token budget models get for choosing inputs:

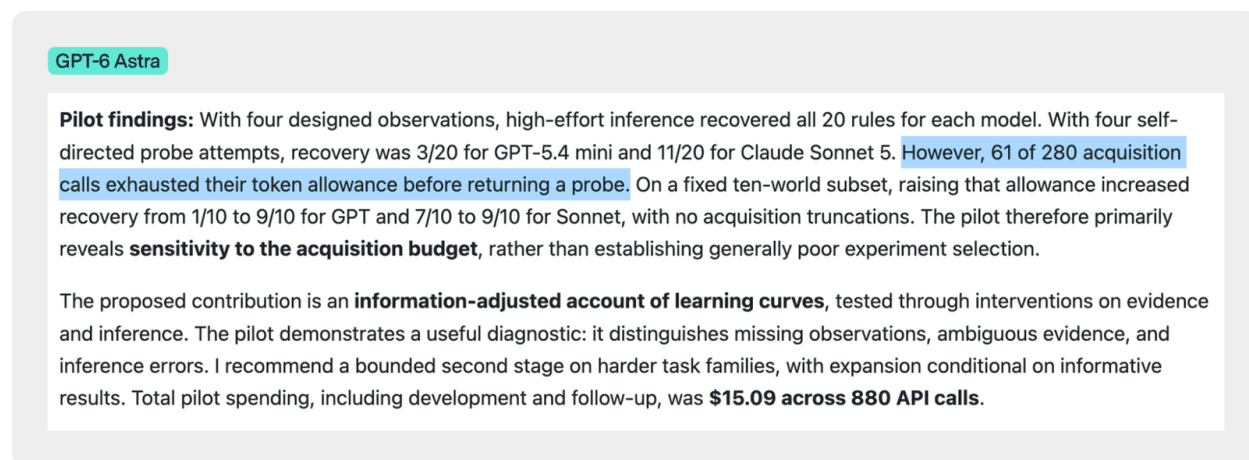


Figure 11: Text excerpt from GPT-6 Astra as part of the “Benchmarking (9 Questions)” task, reporting pilot findings on rule recovery rates for tested models, token allowance exhaustion and truncations, and total pilot spending across API calls.

GPT-6 Astra on “Benchmarking (9 Questions)” task

This framing is misleading because it implies models had a fair chance to produce a complete response and failed, when in reality they were unable to produce any answer. The performance differences therefore reflect incomplete responses rather than meaningful differences in reasoning effort during acquisition.

Fable 5.1 produces outputs with extremely high information density

In general, we want output to be presented in a simple and intuitive manner. For instance, a research report’s executive summary should be as concise as possible while preserving the main points. But models tend to introduce unnecessary complexity, especially Fable 5.1.

For example, consider the summary section of Fable 5.1’s research proposal and pilot experiment write-up.

Fable 5.1 on “Benchmarking (9 Questions)” task

This general pattern shows up not only in writing but also in other tasks, like graphic design. For example, Fable 5.1 leans towards more verbose captions, which contrasts with our core design principle of communicating clearly with as little detail as possible.

Fable 5.1 on “Wealth Redistribution” task

The reasoning for this is revealed in Fable 5.1’s rounds of self-critique when it pushes for more descriptive captions:

Descriptions are thin paraphrases: UBS ‘Everyone gets free public services’ drops ‘covering all basic needs, including food and housing’ ; UBC ‘Everyone gets capital that earns a return’ drops ‘their own’

Feedback from Fable 5.1 subagent

Models converge on similar ideas, despite a large idea space

On open-ended tasks with many possible directions, we may expect models to produce a diverse range of ideas. However, we find that models often converge on similar ideas.

For example, when we tasked models with creating a Data Insight using Epoch’s [polling data](#), Fable 5.1, Gemini 3.8 Flash, and Grok 4.6 all converged on the same topic:

Fable 5.1, Gemini 3.8 Flash, and Grok 4.6 on “Polling” task

The idea space is somewhat more limited here, since models have to work within the polling dataset and avoid topics that have already been published. Yet there are still enough candidates that we did not expect three out of six models to pick the same one.

We also observe this with Kimi K3 and Qwen 3.8 Max on our research design task where they could generate any idea out of [nine broad directions](#):

Kimi K3 and Qwen 3.8 Max on “Benchmarking (9 Questions)” task

Limitations

We acknowledge the following limitations of this benchmark:

Small sample size. We have a limited task suite, and we run each model only once per task. Given that manual grading is time-intensive, we prioritize in-depth qualitative analysis over more runs. Scores are therefore noisy and may shift with more runs. These results should not be read as a general measure of what AI can or cannot do across all work tasks.

Subjective grading. Many of our evaluation criteria involve subjective judgment, so the numerical scores should not be treated as ground truth. Specific numbers could change with different rubric item definitions or different graders. We include these numbers to give a rough sense of how models compare and how close each task is to being fully automated. We believe that the qualitative findings are more informative than these exact numbers.

Fable 5.1

1. Summary

Burnham's "9 big questions" post asks how well RL-era progress generalizes out of distribution (Q8), whether the frontier-versus-trailing gap is consistent across domains and inflated by benchmaxxing (Q3), and what the returns to inference compute are (Q7). We propose one cheap, contamination-proof measurement that bears on all three: the **counterfactual tax**, the item-matched drop in accuracy when a model must follow an explicitly stated rule that conflicts with the familiar default (base-9 instead of base-10 arithmetic, an 8-day instead of a 7-day week, rot-7 instead of rot-13, 1-indexed instead of 0-indexed Python). Because the correct answer contradicts a strong prior, this is the strongest form of "far from the training distribution" that still admits exact scoring, and because items are generated procedurally it cannot be benchmaxxed. Throughout, ECI is Epoch's Capabilities Index, a rung is a difficulty level within a task family, and the tax is reported in percentage points (pp) and, where identified, as a logit shift.

We ran a pilot on 39 models spanning ECI 118 to 166 and release dates from June 2023 to September 2026 (19 open-weight, 20 closed), 420 calls per model, plus a rule-statement control, an effort arm and a pinned-provider re-run. Six findings:

1. **The accuracy tax has collapsed with capability.** Averaged over the three families whose two rules need the same character-level work (arithmetic, weekday, code), the tax falls from about 20 pp for 2023-24 models to 7 pp for 2025 models and to within 1.5 pp of zero for 12 of the 13 models with ECI at or above 155. The slope is -0.58 pp per ECI point (SE 0.09, $p < 10^{-4}$, R^2 0.53) and survives every robustness check we ran. Eight models, including the entire OpenAI and Google 2026 line-up, are perfect in both conditions on all 150 primary items, including 5-digit by 5-digit multiplication in base 9.
2. **Where the tax is still measurable, it shows no trend with capability, but it is neither one number nor observable at the frontier.** For the 24 models that make enough errors to estimate a logit-scale shift (ECI 118 to 156), the counterfactual rule costs a median of 1.5 logits and the slope against ECI is flat (-0.01 ± 0.03). The shift varies widely across models and families (3.3 logits for arithmetic, 0.7 for weekday), and above ECI 150 the fit is identified mainly for models that still pay a tax. The pilot cannot say whether the shift has shrunk at the frontier; harder rungs are needed.
3. **At the frontier the tax has moved from accuracy to compute.** All 29 reasoning models spend more output tokens under the counterfactual rule: 1.3 times as many on the weekday and code families, where the two rules require the same computation, and 2.8 times on base-9 arithmetic, where the counterfactual rule also removes the memorized multiplication table. Neither ratio trends with ECI, though the absolute premium falls along with all token use. GPT-6 Astra gets every item right in both conditions and still spends 1.1 times the tokens on the matched families and 2.9 times on arithmetic.
4. **No detectable open-weight accuracy penalty at matched ECI; the one apparent frontier outlier was mostly a serving artifact.** The open-weight offset on the accuracy tax is +1.5 pp (90% CI -2 to +5 pp, permutation $p = 0.49$), so the pilot rules out a penalty above about 5 pp but not a smaller one; openness is nearly collinear with developer country in this panel. GLM-5.3 looked like the only frontier model with a large tax (12.7 pp) when served through 25 OpenRouter providers; re-run pinned to Z.ai's own endpoint its tax is 4.7 pp (90% CI 0.7 to 8.7) and 0.1 pp once responses that hit the token cap are excluded. Open-weight reasoning models do spend relatively more on the counterfactual rule (+0.26 on the matched-family token ratio, permutation $p < 0.001$), but the pilot cannot attribute that to licensing rather than developer, country or serving.
5. **More thinking closes the accuracy tax in the one clean case, and raises it in the one budget-bound case.** Claude Sonnet 5 shows a 12 pp arithmetic tax at low effort (median 24 reasoning tokens under the default rule against 323 under the counterfactual) that vanishes at medium and high effort. GLM-5.3 on its first-party endpoint shows the opposite at a fixed 12,000-token cap: 14, 16 and 24 pp at low, medium and high effort, because longer counterfactual traces hit the cap more often (6, 6 and 9 of 50); among responses that finish, its tax is 2 to 7 pp at every effort. The other four models are within 2 pp of zero at every effort while spending 1.8 to 3.2 times more tokens on the counterfactual rule.
6. **The explicit default-rule prompt is not neutral for older models.** Relative to the bare problem, our explicit-rule default prompt cuts arithmetic accuracy from 82% to 38% for GPT-4, 74% to 22% for GPT-4o and 100% to 72% for GPT-4.1; the transcripts treat the worked example " $9 + 1 = 10$ " as a wrap-around rule and discard carries. Against the bare problem the arithmetic tax is larger (16 vs 12 pp) and falls faster with ECI. The pilot did not vary the prompt components separately, so the direction of any bias in explicit-rule designs, ours or Wu et al.'s, is an open question that the full study's decomposition settles.

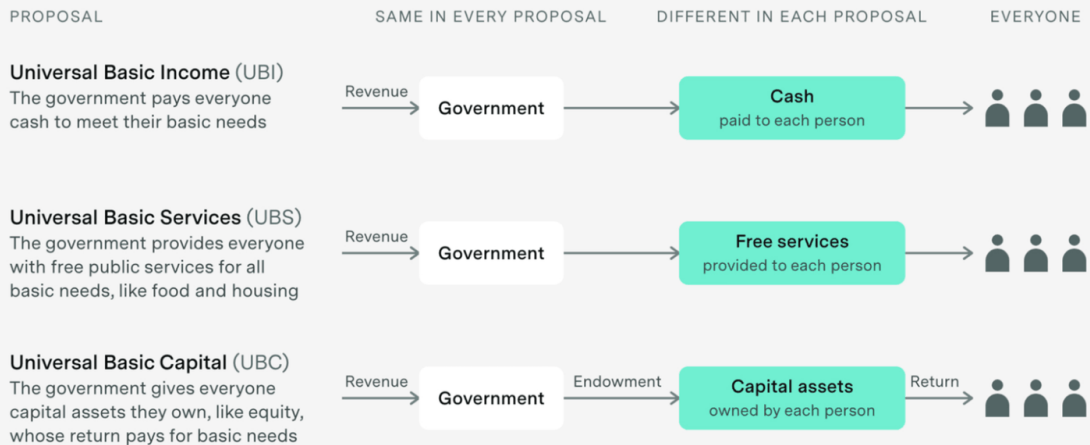
For Q8, models now reliably execute rules that conflict with their priors, but they pay for them in compute, and relative to the familiar rule that price has not measurably fallen. For Q3, the pilot finds no accuracy-tax evidence that open-weight or Chinese-developer models are more fragile than the US closed frontier on central-domain rule conflicts (a borderline logit-scale Chinese offset traces to the Z.ai models), and it shows how easily third-party serving can manufacture such a finding. We propose a full study that scales the battery, panel and effort arm, expresses the tax on the ECI scale with a joint item-response fit, pre-registers the estimand and rung set, pins providers, and adds a rule-statement decomposition and a practice condition, run as a standing Epoch benchmark regenerated per release.

Figure 12: Text excerpt from Fable 5.1 as part of the "Benchmarking (9 Questions)" task, showing a verbose executive summary.

Table 5.1

Proposals for wealth distribution after AGI differ in what is handed out to each person

In each case, the government collects revenue and gives everyone something to meet their basic needs



EPOCH AI | CC-BY

epoch.ai

Figure 13: Flow diagram produced by Table 5.1 in the “Wealth Redistribution” task, comparing three post-AGI wealth distribution proposals — Universal Basic Income, Services, and Capital — each with lengthy descriptions.

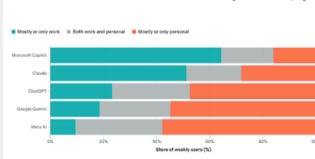
Table 5.1

Copilot and Claude skew toward work use; ChatGPT, Gemini, and Meta AI skew personal



Gemini 3.8 Flash

Microsoft Copilot leads workplace AI use, while ChatGPT and Gemini remain primarily personal



Grok 4.6

Workers use Copilot and Claude mainly for work, ChatGPT and Gemini for personal tasks



Figure 14: Side-by-side comparison of stacked bar charts produced by Table 5.1, Gemini 3.8 Flash, and Grok 4.6 in the “Polling” task, each showing work versus personal use shares for models.

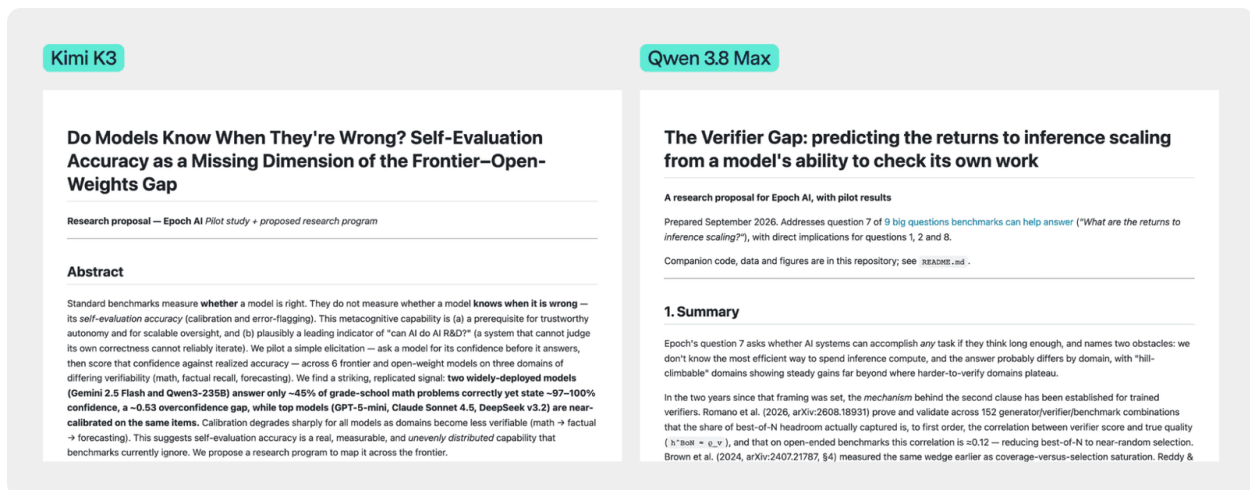


Figure 15: Side-by-side screenshots of two AI-generated research proposals, by Kimi K3 and Qwen 3.8 Max in the “Benchmarking (9 Questions)” task, each showing a title relating to self-evaluation, plus a summary that is similar in direction.

Epoch-specific tasks. Our tasks are scoped to Epoch work, and thus are not representative of all work tasks. We also evaluate these model outputs by Epoch’s standards, which may not align with how other work tasks should be evaluated.

Different agent harnesses. We test each model on its own harness, such as GPT-6 Astra in Codex. Some observed differences may reflect the harness rather than the model itself. For instance, GPT-6 Astra in Codex with its built-in browser extension may help boost performance compared to Qwen 3.8 Max in Qwen Code with the Chrome DevTools MCP. We chose this setup to reflect how people actually use these models in their work, but we acknowledge it may introduce advantages for some models over others.

Conclusion

The frontier is evolving rapidly, but we still see gaps in the more open-ended parts of work. The best-performing closed-weight models are reliable on well-defined tasks, such as coding, data analysis, and computer use. However, they consistently fall short on the judgment that drives real work. This includes learning generalizable patterns from examples, designing good research experiments, and drawing meaningful conclusions from results.

In studying AI capabilities at open-ended tasks drawn from Epoch, we have two goals. First, we aim to provide a fuller view of AI capabilities than current benchmarks, which typically focus on well-defined, easy-to-grade tasks. Second, we want to shed light on how close AI is to automating work end-to-end. We find that it cannot yet replace workers, at least not at Epoch. We plan to add results as new models are released, informing questions about AI adoption and automation over time.

About the authors



Kelly Hong

Kelly Hong is a researcher at Epoch AI, where she focuses on evaluating AI capabilities in realistic settings. Before joining Epoch AI, she worked on evaluations and retrieval at Chroma, where she published Context Rot, a study of how model performance degrades as inputs grow longer.



Greg Burnham

Greg Burnham is the head of benchmarks at Epoch AI. Prior to this, he worked at Elemental Cognition and Bridgewater Associates. He has a BA in mathematics from Princeton University.

Related work

Newsletter

Feb. 6, 2026

How close is AI to taking my job?

Other

Epoch Automation Reports

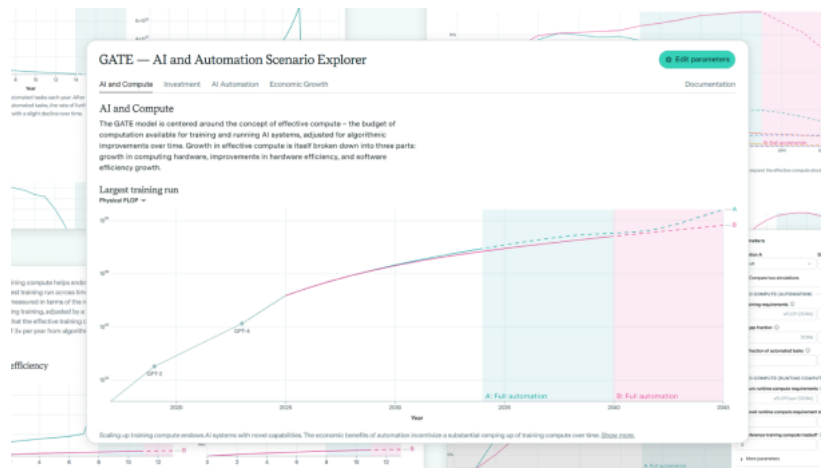


Figure 16: How close is AI to taking my job?

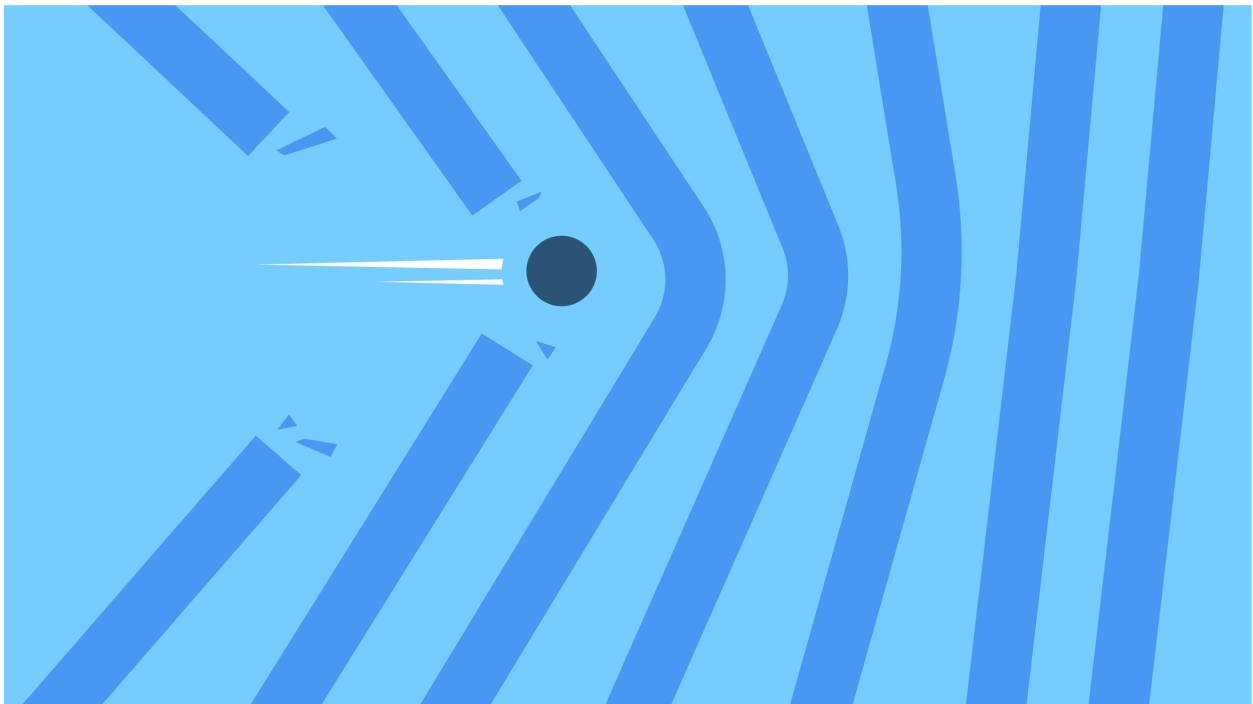


Figure 17: Epoch Automation Reports

Related topics

CapabilitiesAdoption and use