AI in 2030

Extrapolating current trends

This Epoch AI report was commissioned by Google DeepMind. All points of views and conclusions expressed are those of the authors and do not necessarily reflect the position or endorsement of Google DeepMind.

Table of Contents

Executive summary

How will advanced AI be developed, and what will its effects be in the world at large? What will happen if current trends in scaling up AI development persist all the way to 2030? This report examines what this scale-up would involve in terms of compute, investment, data, hardware, and energy. We explore the role of compute across inference and training, the promise of economic value that would be necessary to justify such investment, and potential challenges in data availability and energy.

Based on these predictions for how AI will be developed, we turn to predict future AI capabilities, and the impacts they will have in scientific R&D. AI for science is the explicit goal of several leading AI developers, and is likely to be among the top priorities for AI deployment. Scientific R&D provides a valuable lens for understanding what advanced AI will achieve.

Compute scaling has played a key role in AI development, and will likely continue to do so. Compute for training and inference drives improvements in AI capabilities, and much progress in AI research has come from developing general-purpose methods to enable the use of more compute.

The trajectory of AI development can be forecasted based on continued compute scaling. Scaling has significant implications across many areas of AI development: training and inference compute, investment, data, hardware, and energy. When we predict that compute scaling will continue, we can then examine the consequences within each of these — and how they need to scale accordingly to allow compute scaling trends to continue.

Exponential growth will likely continue to 2030 across all key trends. Across training and inference compute, investment, data, hardware, and energy, we argue that a continuation of existing trends is feasible. We explore each factor in detail, showing how growth could continue to 2030, and discussing the most credible reasons for slowdown or acceleration before then. We argue the most credible reasons for a deviation from trend are changes in societal coordination of AI development (e.g. investor sentiment or tight regulation), supply bottlenecks for AI clusters (e.g. chips

or energy), or paradigmatic shifts in AI production (e.g. substantial R&D automation).

On current trends, the largest AI models of 2030 will require investments of hundreds of billions of dollars, and 1,000x the compute of today's largest models. Investment of this scale is potentially justified if AI can automate significant tasks in the economy. The present trend of 3x annual AI lab revenue growth would lead to revenues exceeding hundreds of billions of dollars before 2030. Finding data for such training runs may be challenging, but between synthetic data and multimodal data, this should be surmountable. Training runs of this scale will require gigawatts of electrical power, approaching the average demand of entire large cities.

Continued scaling will lead to continued progress in capabilities. Once a task begins to show substantive progress with scaling, performance tends to predictably improve with further scaling. Existing AI benchmarks, despite their limitations, cover many capabilities that would be genuinely useful if automated in the real world. Thus, existing benchmarks can inform our predictions on AI's future capabilities. This will be an imperfect view, shaped by the representativeness of existing benchmarks, and limited to where we can already measure progress. We discuss these challenges further in Interlude: from scale to capabilities. Nevertheless, this provides us with a compelling baseline prediction for what AI will be able to do.

At a minimum, AI will act as a valuable tool for scientific R&D. AI systems already excel at helping users find relevant information, implement code, and perform well-defined prediction tasks based on copious domain-specific data. All of these capabilities are set to continue improving.

For example, AI will be able to implement complex scientific software from natural language, assist mathematicians formalising proof sketches, and answer open-ended questions about biology protocols. All of these examples are taken from existing AI benchmarks showing progress, where simple extrapolation suggests they will be solved by 2030. Moreover, AI tools for domain-specific applications will continue to improve. For example, AI tools already offer state-of-the-art predictions for biomolecule

structure/interactions and weather forecasting, and in both areas, progress is set to continue.

Advanced AI will likely lead to a flourishing of desk-based research, which will likely benefit from all of the above advances. In 2030, there will be more software, more mathematical results, more early-stage molecular biology research, more methodological advances in fields such as weather prediction. Areas such as software and mathematics have fewer experimental bottlenecks, and are particularly likely to benefit from AI progress.

For experimental fields, deployment timelines are contingent on hard-to-predict sociotechnical choices. Based on current drug approval pipelines, the drugs that will be approved via clinical trials by 2030 are already in the R&D pipeline today. AI might be contributing to the drug development pipeline by 2030, but within the current regulatory framework, it is unlikely that contributions from AI will lead to approved products available in the market.

The result is a world with increasingly abundant AI-mediated digital services, knowledge, and analysis. By 2030, it is likely that anything physical in scientific R&D will have advanced proportionally less than anything digital. However, if these predictions come to pass, there will be correspondingly strong incentives (and additional resources) to accelerate through these bottlenecks. These efforts may also benefit from AI, but are outside the scope of this report.

Introduction

Compute scaling is the key to AI progress. Using more compute for training and inference is fundamentally what allows AI capabilities to advance. Other crucial factors such as algorithmic innovations and data are important primarily in relation to enabling compute scaling. We will argue this more thoroughly later, but for now, consider the implications if this is true.

What can compute scaling predict?

Assuming that compute scaling drives AI progress, we can predict the near future of AI development by extrapolating recent trends in compute scaling, and the necessary inputs such as investment, data, electrical power, etc. We argue that the baseline for forecasting these things should be trend extrapolation: examine how they have grown recently, investigate the causes, and assume that recent growth will continue unless there is some obvious reason to prevent this. This approach is a common baseline in forecasting (Armstrong 2001), and has been applied in several areas of AI forecasting (Amodei and Hernandez 2018; Sevilla et al. 2024).

As long as investment keeps growing, compute can keep scaling on its current exponential trend until 2030.[1] Then, because AI progress is fairly predictable from scaling, we can predict AI capabilities. Prediction requires existing progress on a relevant benchmark. Fortunately, many relevant benchmarks already provide evidence across economically valuable domains, scientific R&D and otherwise. And these predictions of improved capabilities suggest that investment in compute is likely to continue growing, because such AI capabilities would have large economic value.

This allows us to predict the inputs to AI development. In a world where we "just keep scaling", how much compute is used in 2030? How much is invested in AI clusters to achieve that compute? How much electrical power

[1] Why should the trend be exponential growth, rather than some other form? This property arises when growth is proportional to current value. This pattern is seen across many phenomena in technological and economic progress – for example economic growth, investments, microchip advances, etc.

is needed to supply them? How much data would there need to be for the compute to be productively used? This also allows us to predict the tasks that AI is likely to be able to do at a minimum. What sorts of capabilities will AI have by 2030?

Why compute rather than algorithms or data?

There are two common objections to a scaling-focused view of AI progress: algorithmic innovations and data. We argue that although they complicate the picture, they remain compatible with it.

Algorithmic innovations play a vital role, but they are closely paired with compute scaling. To paraphrase the Bitter Lesson, the most important and effective algorithmic innovations are general-purpose methods that enable compute scaling.[2,3] Moreover, there is some evidence that algorithmic innovations rely on compute scaling for their development. This suggests that we should anticipate algorithmic progress, but enabled by, and focused on, compute scaling. Nevertheless, this is a key uncertainty. Capabilities could improve faster than predicted here, if compute is not a bottleneck.

Data is essential for AI training, and the quality of datasets can significantly influence results. However, there are two reasons to think that compute is more of a rate-limiting input. First, compute is more of a bottleneck in the current paradigm of AI training, at least for general-purpose LLMs. We could scale up for at least a few more years using existing public text data and other modalities (Data). Second, it appears increasingly likely that inference scaling will make training more compute-intensive, effectively using compute to generate data for reasoning training (Data won't run out by 2030, although human-generated text might). Specific data bottlenecks can be important within particular applications, and we discuss these further in Capabilities in scientific R&D. Hence, we must consider data

[2] The Bitter Lesson in brief: "The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin" (Sutton 2019).
[3] Examples of general-purpose compute-leveraging AI innovations from the past few decades include Convolutional Neural Networks, GPU acceleration, the Transformer architecture, and Large Language Model pretraining.

availability when we investigate scaling, but this remains compatible with a scaling-focused view.

What can't compute scaling predict?

What this *doesn't* allow is predicting when we have general intelligence, i.e. AI that can perform any cognitive task at the level of a skilled human. This question suffers from two massive uncertainties: gaps in AI benchmarks and our understanding of them, and gaps in current AI capabilities that might not be filled in the next five years.

Current benchmarks might not adequately represent the most difficult-for-AI tasks that humans do. And not all benchmarks show AI progress yet: there are tasks where AI doesn't yet show much improvement with scaling so far, for example "autonomously prove a new substantive mathematical theorem". It is fundamentally uncertain where AI will have reached expert-level performance by the time it solves existing benchmarks (that show progress). It is also fundamentally uncertain when AI will solve all existing benchmarks, because it only shows progress on some of them. Nevertheless, it is fairly certain that AI will solve many challenging benchmarks by 2030, and these have clear implications for useful tasks that AI will be able to perform.

Meanwhile, gaps in the capabilities of present-day general-purpose AI systems shed some light on the capabilities that AI could fail to achieve by 2030. AI models excel at identifying relevant information from a large training corpus, but frequently veer into illogical hallucinations. They are brilliant at ingesting large amounts of data and identifying underlying patterns, yet fail to reliably apply reasoning steps that would seem natural to a human. Reliability and robustness are problems more broadly, although they have at least shown incremental improvement with scaling. AI excels at solving closed-ended optimisation problems such as games, yet struggles to perform consequential actions in the real world with agency. It can perform shallow processing of long content much faster than a human, but it struggles to use this long-context information for solving challenging problems. There are enough gaps in current AI capabilities that it is hard to even be certain which of them are overlapping – perhaps long-context

comprehension is related to robustness of reasoning, or perhaps they are entirely separate problems. These limitations are related to the challenges with designing and interpreting AI benchmarks: adequately benchmarking these limitations is also an open problem.

It is uncertain which of these AI limitations will improve by 2030, and by how much. It is uncertain whether these will improve by "just scaling" existing systems with small modifications, but also, it is unclear how much compute they would need. Consider the example of reasoning models: pre-existing systems already used inference scaling, but reinforcement learning (RL) made this far more effective, yielding breakthrough results in several benchmarks. Does this challenge the view that scaling is the driver of progress? Arguing in favour of scaling-driven progress, many researchers predicted ahead of time that better inference scaling would be necessary, and this arrived after models scaled up sufficiently for reasoning RL to work. Furthermore, training scaling holds for RL: using more RL training compute improves the capabilities that reasoning models achieve. On the other hand, this emphasises the challenges of prediction from existing results. The areas where AI struggles today can sometimes see breakthrough algorithmic progress, and this is inherently hard to predict.

Even in a scaling-first view, it is unclear how much more scaling needs to happen to reach AGI. It is also unclear whether this will require significant algorithmic advances. It is uncertain whether such algorithmic advances, if needed, might still happen by 2030. These are the biggest challenges to scaling-focused predictions for AI, and particularly when trying to predict beyond continuing progress on already-progressing benchmarks.

Despite these significant challenges, scaling-focused predictions are still useful. We can predict a minimal baseline: tasks that we expect AI to continue improving at with further scaling. We can then examine how the resulting capabilities would affect real work tasks. We can reflect on further tasks that are not covered by the baseline, and the implications for automation if AI did become capable of these. And then, finally, we can follow through to reason about the broader implications this would have within people's work. This allows us to bridge two competing views of AI: AI as a powerful tool, and AI as a virtual worker.

What does scaling predict about AI development?

We examine several key inputs: compute, investment, data, hardware, and energy and the environment. Any of these could undermine continued progress – for example, what if compute scaling stops being effective?[4] What if we run out of data? Some of these arguments are stronger than others, but we see no single compelling argument to prevent current progress continuing to 2030. We explore the implications of this across each of the key inputs: surveying how far they might scale, and how they might be derailed.

In brief, we predict that, on current trends, leading AI models in 2030 will be trained with 1,000x the compute of today's leading models. The clusters used for training such models would require investment of two hundred billion dollars, close to 1% of present-day United States GDP. Training and deployment will require gigawatts of electrical power for the largest models, and total AI datacentre power could easily grow to 2+% of global electricity demand, similar to the level of demand from electric vehicles (around 2% by 2030 [IEA 2025d]) or the Internet (2-3% in 2025 [Rozite et al. 2023]).

KEY FINDINGS FOR AI DEVELOPMENT TRENDS

Compute: Training compute has increased 4-5x per year since 2010, and is likely to continue growing at a similar pace. By 2030, on current trends, the largest AI models are likely to be trained with 1,000x the compute used in today's leading models. Scaling up inference compute will be another important source of continuing AI improvement. This is unlikely to interfere with scaling of training compute, and for a given model, its lifetime inference compute will probably be comparable to its training compute.

Investment: To enable training at this scale, the necessary AI hardware would cost hundreds of billions of dollars on current trends. The amortised cost of developing individual models would be billions of dollars. These

[4] We discuss further in Scaling is not "hitting a wall", although it is getting harder that there are many different senses in which growing investments could become uncorrelated with further AI capabilities improvements.

projections align with current AI investments and valuations, as well as capital expenditure plans from AI cluster developers. Frontier AI labs already earn billions from chatbots, with revenues growing 3x per year in the last couple of years. If AI can significantly raise net productivity across the economy, it will be worth trillions of dollars. This would justify substantial investments in its development. We discuss later how AI could achieve such net productivity gains – this depends on both the capabilities achieved, and being able to deploy them cost-effectively.

Data: Datasets for general-purpose AI training recently grew at 2.7x per year, but further dataset growth could change significantly. A shift towards multimodal and synthetic data may be necessary as high-quality human-generated text data becomes scarce. Recent trends in reasoning training suggest that growing human-provided data at a much slower rate could nevertheless enable compute scaling to continue via synthetic data for reasoning training. If AI capabilities continue to improve, then particular sources of specialist data will become increasingly valuable: essentially, data to enable training on high-value problems.

Hardware: Total installed capacity for leading AI chips is likely to continue growing 2.3x per year, driven by producing more chips with better performance.[5] Large AI clusters, in line with current trends, are already being planned and developed for the largest AI developers. However, it is likely that large AI workloads will be increasingly distributed across multiple datacentres to ease the demand for electrical power.

Energy and the environment: Power demands for frontier AI (both training and inference) are likely to grow at around 2.1x per year, and AI energy demand generally is on track to grow around 1.6x per year. In this case, AI datacentres would grow to 1.2% of global electricity demand. Depending on the energy mix used to power datacentres, AI electricity use could account for 0.03-0.3% of global emissions by 2030. Although significant, this is much smaller than projected emissions from commercial flights (2.5%, [IEA 2025a]). There is demonstrated potential for AI to reduce emissions in areas

[5] 2.3x per year growth in installed computing power is slower than the projected 4-5x per year growth in frontier training compute. Currently, individual frontier training runs use about 2.5% of installed capacity when operating, so there is room for training runs to grow faster than total capacity.

such as energy production, industrial process optimisation, and transport, but this heavily depends on societal decisions about deployment and prioritisation.

What does scaling predict about AI capabilities and impacts?

What capabilities will the AI systems of 2030 achieve, and what impacts might it have in the world? This is an incredibly broad question, and to make it tractable, we narrow our scope to a critical area: automation of scientific R&D. AI for scientific R&D is explicitly the goal of several leading AI developers (Altman 2023; Amodei 2024; Google Deepmind, n.d.), and occupies an important position in the economy due to its ability to improve productivity more broadly.[6] We explore AI's potential for scientific R&D across several different areas: software engineering, mathematics, molecular biology, and weather prediction.

As previously discussed, our predictions are anchored on extrapolating trends in present-day AI capabilities. There are two main reasons this approach could be overly aggressive. The first reason is that if benchmarks are not representative of the capabilities they are intended to measure. We examine this further within each individual section of Capabilities in scientific R&D. In several domains, such as software engineering and biology, there is already some empirical evidence suggesting that benchmark progress is correlated with real-world progress. The second reason is that benchmark progress could be deceptive due to overfitting. Although this is a real challenge for comparing models at a point in time, we believe it is less of a concern for broadly predicting progress across coming years. Benchmarks in the past were also subject to overfitting, but nevertheless, solving them went hand-in-hand with related AI capabilities progress. If current benchmarks overstate progress due to overfitting, then

[6] This is not to say that R&D will necessarily be the first or most economically significant set of activities to see AI automation. There are credible arguments that AI developers face larger incentives, and easier challenges, broadly automating many tasks across the economy. Nevertheless, AI developers' explicit focus on R&D motivates us to give it attention in this work.

our extrapolations will be aggressive, but as long as there is some real underlying progress, they will nevertheless be informative.

Capabilities trends suggest there will be tremendous progress in AI for scientific R&D, particularly in areas such as software engineering and mathematics, where realistic tasks can be trained on entirely in silico. To offer concrete examples: by 2030, existing benchmark progress suggests AI will be able to implement complex scientific software from natural language, assist mathematicians formalising proof sketches, and answer complex questions about biology protocols. We describe this further below.

KEY FINDINGS FOR AI CAPABILITIES IN 2030

Software engineering: Many of today's day-to-day tasks are likely to become automatable by AI agents. Existing benchmarks based on well-defined software issues, such as SWE-bench, are on track to be solved in 2026. Current progress on solving defined hours-long scientific coding and research engineering problems (RE-Bench) is slower, but on its current trajectory would be solved in 2027. A key uncertainty is whether human supervision will be a bottleneck for more open-ended problems.

Mathematics: Challenging mathematics reasoning benchmarks, such as FrontierMath, could be solved as early as 2027 on current trends. Mathematicians predict AI capable of solving such benchmarks might help them by developing sketch arguments, identifying relevant knowledge, and formalising proofs. This would allow AI to fulfil a similar role in mathematics to coding assistants in software engineering today. Even more than for software engineering, a key uncertainty is whether existing mathematics benchmarks are valid for predicting such capabilities. The most challenging mathematics benchmarks today are further from mathematicians' day-to-day work than software benchmarks are from that of software engineers. It is unclear when AI can rise to the level of autonomously proving substantive results, but it is plausible that this will happen before 2030.

Molecular biology: Public benchmarks for protein-ligand interaction, such as PoseBusters, are on track to be solved in the next few years, although the timeline is longer (and uncertain) for high-specificity prediction of arbitrary protein-protein interactions, especially further from training data. Meanwhile, AI desk research assistants are set to help in biology R&D in coming years. Open-ended biology question answering benchmarks are on course to be solved by 2030, albeit with large uncertainty. Importantly, advances in basic biology R&D are likely to take several years to lead to downstream changes in e.g. pharmaceutical development, due to bottlenecks in both wet lab experiments and clinical trials.

Weather prediction: AI weather prediction can already improve on traditional methods across timescales from hours to weeks. Moreover, AI methods are cost-effective to run, and are likely to improve further with additional data. The next big methodology challenges lie in improving prediction calibration at current horizons, rather than extending them further.[7] There are outstanding improvements to be made in two areas in particular: forecasting rare events, and integrating additional data sources. Using more historical data and more finegrained historical data for training can improve predictions, and more real-time sensor inputs could be integrated for better performance in deployment. There are important challenges in development and deployment: funding the research, getting access to data (particularly at low latencies in deployment), and in some cases even permissions to install data recording equipment. Nevertheless, improved weather prediction methods could achieve significant benefits in the wider world, helping in areas such as power infrastructure, agriculture, transport, emergency response, and everyday planning.

Considering the prospect of general-purpose AI assistants, there is a clear vision of AI automating tasks within researchers' existing work. We describe this further in "Claims about AI assistants for scientific R&D". Meanwhile, for areas such as molecular biology and weather prediction, the path ahead is less clear: much progress to date has come from narrower AI tools, and much human labour (or deployment) could be bottlenecked by interaction with the physical world. For such disciplines, it seems likely that desk-based research will flourish, enabled by AI, but with experimentation and broader impact lagging behind. For example, there may be an increase in the quantity and quality of promising candidate molecules for drug development, but due to multi-year timelines for clinical trials and drug approvals, it is unlikely that much of today's AI research will be relevant to the drugs released in 2030.

[7] Until recently, it was widely accepted that deterministically predicting weather beyond a horizon of about three weeks is not possible for simulation-based methods, due to chaos effects. Recent work raises the question of whether this was unduly pessimistic (Shen et al. 2024; Chen et al. 2024). There is also the possibility that integrating more data allows for improvement beyond the limits of pure numerical simulations, and existing long-range forecasts make use of data, as well as ensembling. Note that weather prediction is distinct from climate prediction, which is much longer term, with much lower temporal resolution.

CLAIMS ABOUT AI ASSISTANTS FOR SCIENTIFIC R&D IN 2030

From most to least certain

1. **At minimum, scientific R&D will get AI assistants comparable to coding assistants for software engineers today**. This is almost certain – as we later examine, there are existing benchmarks showing AI progress for the relevant capabilities, and existing AI systems being used for literature review, protein design, etc. These functionalities have differences compared to software engineering, for example more of a focus on reviewing and synthesising large and heterogeneous literature, whereas existing AI coding tools are primarily limited to the context of a single project. Nevertheless, there are important similarities: offering suggestions in response to context, finding relevant information, completing smaller closed-ended tasks in their entirety.

2. **At minimum, AI assistants are likely to improve day-to-day productivity by 10-20%, at least within non-experimental work tasks.** While less certain, this is the starting point from randomised trials on software engineer productivity (see Software engineering for discussion of current evidence, including negative results). Even if the work tasks of a mathematician or a theoretical biologist are less amenable to automation than a software engineer, we already have evidence from relevant benchmarks improving, and anticipate many more years of progress still to come.

3. **The effects could be larger than this.** The 10-20% figure was measured for software engineers using Copilot beginning in late 2023 (Cui et al. 2025). AI systems since then have improved substantially, and early evidence documents the improving capabilities of autonomous software engineering agents.

AI will be at least as important as the Internet by 2030

In short, we describe a world in which scaling AI further leads to further capabilities, consistent with what we have seen so far. Such capabilities can automate meaningful tasks across the economy, scientific R&D among them. Scientific R&D is highly valuable and rapidly evolving work where AI will be adopted fairly quickly,[8] but the same AI advances will be invaluable across many sectors. Deployment takes time, and many bottlenecks must first be confronted. However, if current trends continue to 2030, a radically transformed world will at least be within sight. It may seem extreme to predict that entire power plants' output might be dedicated to AI, but this will be justified in such a world, where AI is becoming comparably important to the Internet.

Inevitably, these claims must be caveated with significant uncertainties. Perhaps AI capabilities will stall near current levels, because today's algorithms are insufficiently general-purpose. We discuss this in Charting the trajectory of future AI capabilities; we argue that simply forecasting from existing AI benchmark progress yields these predictions as a fairly conservative baseline. Perhaps deployment will be slow, particularly in challenging R&D tasks, or in other key parts of the broader economy. In Capabilities in scientific R&D, we argue that although deployment is challenging, AI technologies have seen the fastest adoption curves in history. Current adoption trends are consistent with reaching hundreds of billions of dollars in revenue by 2030. Another common objection is that mass adoption will be bottlenecked by lack of compute. In Interlude: from scale to capabilities, we show that current trends in installed compute argue against this.

These are the predictions that align with current trends, particularly when it comes to the next five years of AI development. These predictions may have substantial uncertainty, but we argue they should be the baseline forecast. By default, the world in 2030 will be filled with highly capable AI

[8] There is survey evidence that academics have rapidly adopted existing AI tools (Oxford Academic 2024), and we discuss deployment prospects separately in several scientific R&D domains, finding that many potential bottlenecks (inference cost, specialist data) do not appear prohibitive.

systems deployed at scale, both as scientific tools (e.g. weather prediction systems and protein structure modelling) and, at least to some extent, as autonomous agents pursuing substantive real-world goals (e.g. in software engineering). We must prepare for that world now.

Scaling and capabilities

There are two key parts to our predictions, corresponding to the two main sections: Scaling and Capabilities.

Scaling has arguably been the most fundamental contributor to AI progress. Training AI systems with more compute, on more data, has led to stronger capabilities. This is not to ignore the role of algorithmic progress – generations of researchers and engineers have spent entire careers developing innovations to improve AI development. However, the history of AI development suggests that these innovations work alongside scaling, either enabling scaling or making it more efficient. We discuss below how scaling compute improves performance, both during training and inference. We also discuss how compute scaling seems likely to continue over the next five years.

The capabilities that result from scaling are fundamental to how AI will be used in the world. We discuss how, if scaling is the rate-limiting factor for AI development, then we can use scaling predictions to chart the trajectory of AI capabilities. This allows us to make predictions about what AI might be doing in the world by 2030: we examine existing trends in progress, and extrapolate forwards.

Scaling compute improves performance

GPQA: a benchmark of multiple choice PhD-level science questions. A similar pattern of performance improving with compute has been seen across many AI benchmarks (Denain 2024).

The reason to scale compute is straightforward: it improves performance. This applies both to training compute and inference compute. Predicting performance from scale is not always straightforward, and it is particularly difficult to anticipate when an entirely new capability will emerge.[9] However, once there are initial signs of AI achieving some capability, subsequent scaling predictably improves performance in most cases (Owen 2024a; Schaeffer et al. 2023).

[9] For example, if AI models perform at random chance on some benchmark, it is hard to predict how much more compute will be required for them to start improving.

Compute scaling has been studied most extensively for training compute. Scaling is most predictable when variables other than compute are held constant, for example when a frontier lab is simply "scaling up" a given architecture and training recipe. However, even when considering many different AI models following different architectures, training approach and data mixes, performance on benchmarks is fairly correlated with compute.[10]

Inference compute scaling, meanwhile, has recently seen significant advances. Previous methods for inference compute scaling tended to be inefficient.[11] New reasoning LLMs allow for scaling inference compute much more cost-effectively.

[10] For example, in popular science questions benchmark GPQA, performance is correlated with training compute with R^2 of 0.45 (Denain 2024).
[11] There are notable exceptions. For example, if we count certain kinds of post-training enhancements as inference scaling, then for some applications inference scaling was very cost-effective. Nevertheless, for many tasks, especially tasks for which people wanted to use language models, inference compute scaling before 2024 was expensive.

Scaling is not “hitting a wall”, although it is getting harder

There is a recent argument against scaling, which has seen a lot of public discussion: that scaling is yielding “diminishing returns”, or is “hitting a wall”. There are several senses in which this might be true, and it is important to distinguish between them.

The most aggressive version of “hitting a wall”: scaling laws, which relate next-token prediction performance to training compute, may break down. There is scarce public evidence to suggest this is true. It may yet happen, but we have no reason in particular to expect it.

A less aggressive version of “hitting a wall”: scaling laws for next-token prediction may remain valid, but improvements on downstream tasks are worse than expectations based on compute scaling or researchers’ intuitions. Journalists have made versions of this claim, attributing it to researchers at leading AI labs. From public information, benchmark performance seems broadly on trend with compute scaling for the first models known to be trained beyond GPT-4 scale such as GPT-4.5 and Grok-3.[12]

Then, there is a much looser version of “hitting a wall”: scaling laws remain accurate, performance improvements are as expected, but further scaling is *harder* than before, e.g. because of the requisite investment, data, power constraints, chip production, and latency. This is compelling. Contemporary AI training datacenters are reaching hundreds of thousands of GPUs. This is approaching the limits of what a single datacenter can power, leading AI developers to run multiple-datacenter training (Gemini Team et al. 2023). High quality public text data may be growing harder to source.[13] We investigate these within their respective sections, and on balance find that none of these would clearly constrain training scaling trends before 2030.

[12] GPT-4.5 scores about 20% higher on GPQA compared to GPT-4o, and 26% higher on Math Level 5. Across many different models on the same benchmarks, performance recently scaled at 14% and 19% per order of magnitude compute scaling. This is consistent with scaling roughly as expected.

[13] Although data overall is unlikely to run out, as we discuss in Data won’t run out by 2030, although human-generated text might.

Finally, there is the claim that training scaling is "hitting a wall", because inference scaling is so much more effective after the development of reasoning models. Reasoning models are a significant advance, and will change the details of training scaling, but we argue that inference compute and training compute are likely to scale similarly. Scaling training leads to more capable models, which can do more with a given inference budget.

Inference scaling is related to an important consideration: what we count as training compute. Recent frontier models have increasingly relied on post-training, and by some reports post-training compute could soon be scaled to the same level as pretraining compute (Amodei 2025). Even if pretraining scaling were thwarted by a lack of data, several notable researchers have predicted that a move to post-training on synthetic data is the next era of AI development.

Charting the trajectory of future AI capabilities

The astounding results from the past 15 years of AI development lead to the question: where will scaling lead us? What will AI be able to do in the future? Will scaling lead to artificial general intelligence (AGI) or beyond - for example, AI capable of performing practically any cognitive task performed by humans (Morris et al. 2024)?

We have little certainty on the required training and inference compute for AI that can do any cognitive task.[14] However, we can chart the trajectory of AI capabilities. Scaling clearly allows predictions for AI evaluation performance. And we already have AI systems that can achieve impressive results in hard evaluations. Hence, we can predict "what sorts of capabilities will scaling yield". We can predict what capabilities AI systems will have in tasks such as implementing complex software, performing biology literature search, and so on. These predictions will be noisy, but they will be grounded in existing progress. This suggests we will be able to chart the trajectory towards advanced AI, even if it falls short of AGI.

Another important challenge to this approach: will compute be the rate-limiting factor? Several leading AI researchers predict that advanced AI will need more algorithmic innovations along the way,[15] but these will be devised faster than the necessary scaling up of computing hardware can happen. Hence, scaling is likely to be the binding bottleneck.[16] There is certainly disagreement on this point – AI researchers have a wide range of beliefs about timelines to advanced AI (Grace et al. 2024). However, in light of recent progress, it is worth taking the possibility seriously. If scaling will play a key role, then "straight line extrapolation on graphs" will be a fruitful way to think about the development of advanced AI.

[14] Several researchers have attempted to anchor the required compute for AGI using equivalent compute used by the neurons in the human brain, or thermodynamic limits from evolution. The results have been highly variable in their predictions.
[15] For example, prior to late 2024, inference-time scaling was relatively inefficient, and reasoning models were an innovation addressing this.
[16] See footnote 1.

Scale

What resources will go into AI development, five years from now? Extrapolating key inputs, such as training compute, allows us to reason about how AI development will proceed. It lets us consider how existing progress might continue, and where significant changes may be required.

Our starting point is to examine existing trends, and interrogate the factors that could cause them to change, going forward. Extrapolating like this is a strong baseline, particularly over shorter time periods. The scaling-up of training compute, for example, has been a relatively constant trend at 4x per year since the Deep Learning era began in 2010. We could have predicted the largest training run in 2024, with reasonable accuracy, simply by extrapolating the trend in 2020 (Sevilla and Roldán 2024).

We first examine scaling of training and inference compute. We argue that training compute trends are likely to continue as long as there is sufficient investment, although training may shift to focus on synthetic data and/or post-training. One reason that scaling of training might cease is if it offered disappointing AI capabilities improvements – we argued above that there is little evidence of this so far, and scaling-driven deep learning has not "hit a wall" in terms of benchmark progress. Meanwhile, recent advances in inference compute scaling indicate a complementary way to improve model capabilities. With significant uncertainty, we expect AI labs will ultimately scale training and inference at similar levels.

Continued AI compute scaling would require a commensurate scale-up of investment. We show how this has happened historically at 2-3x per year, and briefly examine how massive investment in AI development could be justified by AI-driven productivity improvements. Hardware manufacturers' valuations imply the market expects AI to generate over a trillion dollars annually, and supports at least an additional doubling of cluster size scale-up. Investments larger than this could be justified by commensurately more value – which is supported by recent growth in AI revenues.

We then examine trends in training data. We show how general-purpose publicly available text data plausibly could be exhausted before 2027. Nevertheless, AI developers are unlikely to run out of data for large-scale training. This is due to two outstanding sources: synthetic data (particularly for reasoning training), and multi-modal data. We also examine the role of high value specialist data sources, for example problems that can be adapted to generate synthetic data with verifiable solutions. Of particular interest for this report, we discuss the importance of data covering high-value domains in scientific R&D, such as biomolecule structure and interaction data.

In a scaling-driven view of AI development, hardware is vitally important. We show how the largest factor in scaling up training compute was increasing cluster sizes, followed by longer training runs and improving hardware performance. We argue that scaling of cluster sizes is likely to continue, and offer evidence based on the next generation of AI clusters. Meanwhile, we offer tentative evidence that training durations could plateau, as algorithmic progress and hardware progress discourage them from growing too long – and recent reports suggests frontier model training has stabilised around two months.

Finally, we examine the implications of such a scale-up for energy and the environment. We show how power demands for frontier training have doubled annually, and seems likely to continue. In an extrapolation based on power draw from high-end AI chips, AI would make up approximately 1.2% of total electricity demand by 2030. Emissions would vary greatly, depending on the energy mix underpinning its usage. If datacentres exclusively used low carbon intensity power, it could be as low as 0.03% of global annual emissions in 2030. If datacentres used an energy mix comparable to the grid average, similar to natural gas, it could be as high as 0.3% of global annual emissions in 2030. In practice, emissions are likely to be closer to the second figure, unless solar and other renewables expand far beyond current projections.[17] Crucially, AI's overall effect on emissions

[17] Globally, datacentres currently use an energy mix close to the grid average, with US datacentres being significantly greener and Chinese datacentres largely powered by coal. Projections of datacentre electricity supply suggest more of the growth will be from renewable sources, although much of the demand in China may be delivered by coal (IEA 2025c).

would also depend on its uses. Across many applications, AI could reduce global emissions by more than it would increase them. Whether this would happen in practice depends on hard-to-predict societal choices.

Overall, we make a simple forecast about scaling: we predict that the likeliest outcome is that present trends mostly continue. Training continues growing as it has grown since 2010, with inference growing alongside it. To support this, investment must also grow, reaching extreme levels. However, these investments are justified because investors predict that AI will offer commensurate economic value.[18] Consequently, the industry continues deploying more AI chips, consuming correspondingly more electrical power, similar to other key sectors in the economy.

[18] Consultants' projections of economic value from generative AI in existing work tasks run up to trillions of dollars (Chui et al. 2023). In Investment we discuss how current investment trends are consistent with this.

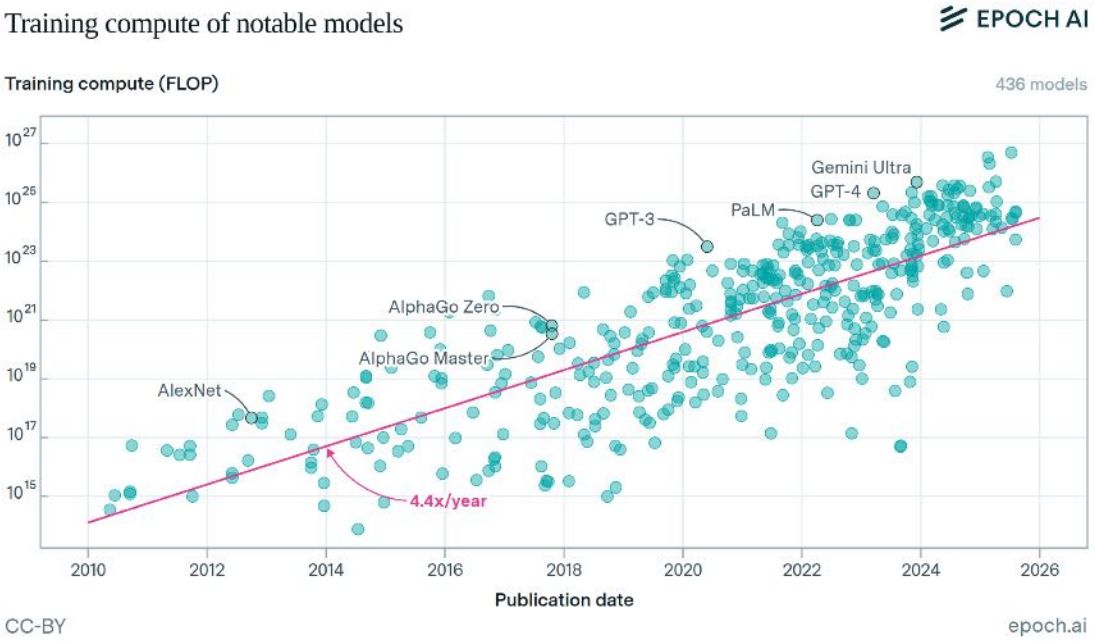
Compute

Compute underpins modern and historical progress in AI. In particular, the shift towards specialised AI hardware and large training clusters drove much of the progress in AI capabilities. Meanwhile, more recently, specialised reasoning models have enabled efficient scaling of inference compute.

We argue that training compute is likely to continue increasing around 4-5x per year until 2030. This trend has been persistent since 2010. Training larger models continues to improve AI capabilities. Although further compute scaling is challenging in terms of data, hardware, and electrical power, all of these technical challenges appear surmountable until 2030. The largest uncertainty is whether investments will continue growing, which we discuss further in Investment.

We further argue that growing demand for inference compute is unlikely to inhibit training compute growth. Inference costs grow proportionally with the number of times a model is used, whereas training compute is an upfront investment in model capabilities. For this reason, as well as evidence from AI deployment so far, we expect that frontier AI labs will scale up both training and inference compute, at around 4-5x per year.

Training compute has increased 4-5x per year, and is likely to continue



Training compute for notable AI models has grown around 4-5x per year since 2010, with a similar pattern for frontier models. Recent frontier models have been general-purpose AI models, with most training compute spent on language training.

Trends in training compute growth have persisted across 14+ years. If they continue, the largest models will be trained using 10^{29} FLOP by 2030 – a quantity of compute that would have required running the largest AI cluster of 2020 continuously for over 3,000 years.[19] Assuming that the necessary algorithms to continue scaling are already in place, or will be discovered along the way, what could change this trend?[20] Some of the most pressing potential bottlenecks that have been suggested are investment, running out

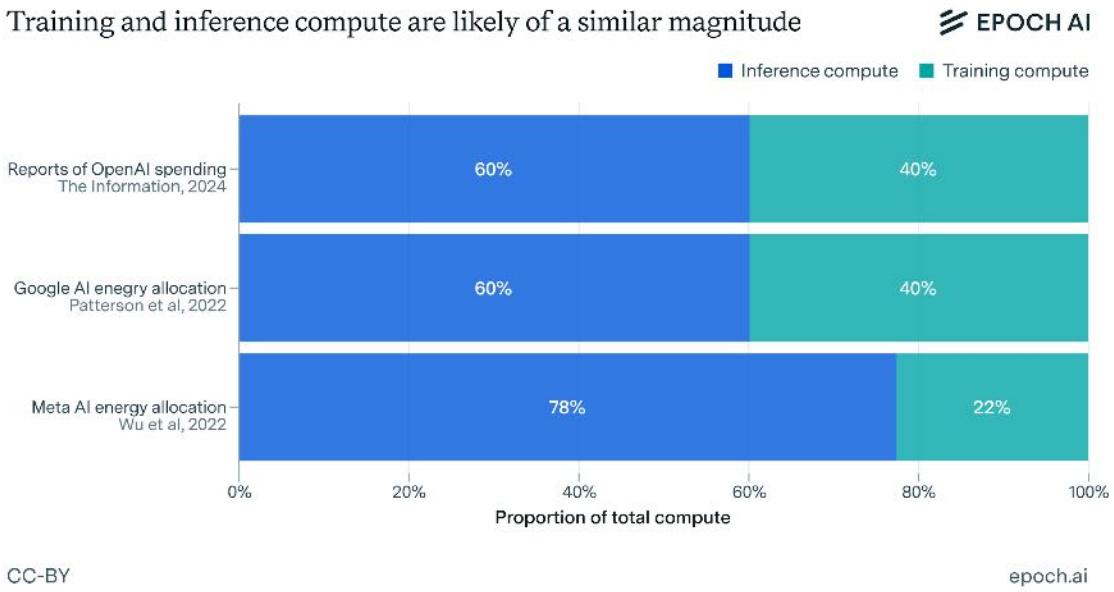
[19] The largest AI clusters of 2020 had peak performance of about 10^{18} FLOP/s (Pilz et al. 2025).

[20] By "the necessary algorithms to continue scaling", we mean both the low-level algorithmic changes necessary to use familiar techniques at a larger scale, and higher-level innovations such as "better approach for long-context memory". As we discussed in Charting the trajectory of future AI capabilities, researchers differ in opinion on whether the latter innovations are on the critical path to advanced AI.

of data, power constraints, chip production, and latency constraints. None of these is clearly an obstacle to continued scaling, as we will discuss within their respective sections.

Investment is a key uncertainty, interacting with all the other potential bottlenecks as well as the broader context of markets, society, and governance. To continue investing in scaling, relevant actors must see sufficient potential for returns. For this to happen, scaling must continue improving capabilities. Whether capabilities will improve sufficiently to justify investment is more uncertain. This depends on the timelines of key AI capabilities (already challenging to forecast), but even these are not sufficient to predict investment, because investments can be made far in advance of their expected returns. We discuss this further in Investment, arguing that current trends and investments support continued scaling, and these trends could continue to 2030 if AI is able to accelerate a meaningful fraction of remote work tasks.

Inference compute scaling won't detract from training compute



Reported allocations of spending and energy allocation suggest that training and inference compute are of fairly similar magnitude for large AI developers. Moreover, total installed AI compute has grown around 2.3x/year, similar to frontier AI training clusters (Pilz et al. 2025), suggesting training has grown similar to inference.

Recent developments in inference compute scaling have been hailed as a paradigm shift. Some have linked this to the idea that training scaling will slow, or even cease, as AI developers focus on using inference compute. However, inference scaling need not imply that training compute scaling will cease, because reasoning models also benefit from training compute.

Evidence to date suggests that compute has been allocated fairly similarly between training and inference (see Figure above). Inference has gotten more compute (60-80%), but the allocations have remained of a similar order of magnitude. More generally, as long as it is possible to trade training and inference compute off against one another, there are reasons to expect that AI labs should continue allocating similar resources to each. Training higher-quality models reduces the amount of inference needed for a given

level of performance, and allocating roughly equally across these allows the most efficient use of a fixed compute budget (Erdil 2024).[21]

Are there any reasons that inference might scale differently to training? The main reason we could foresee is if trade-offs between inference and training are exhausted. This is difficult to forecast based on existing data, where such trade-offs continue to be available.[22] A potential example is power: if large training runs become bottlenecked by the need to concentrate compute in a small number of datacentres, and power for these cannot be supplied, or is more expensive, then the balance would shift towards inference. Still, for inference to actively detract from training scaling, such obstacles need to be extreme. In the particular case of power, we see no impediment to scaling on trend until 2030.

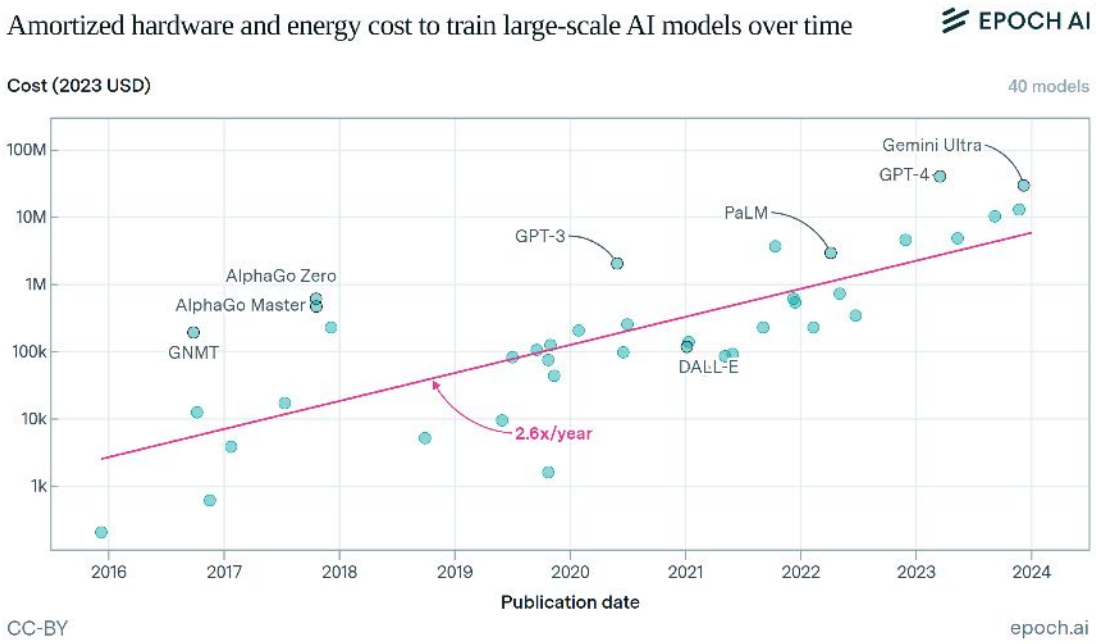
[21] Evidence so far suggests that an order of magnitude in training compute can be traded off for approximately an order of magnitude in inference compute while keeping model capabilities fixed. Consider a lab spending 100 zettaFLOP on inference and 1 zettaFLOP on training, for a total budget of 101 zettaFLOP. This trade-off implies the lab could instead achieve the same quality-adjusted output by adjusting to 10 zettaFLOP on inference and 10 zettaFLOP on training, for a total budget of 20 zettaFLOP. The optimal solution is to allocate them according to the trade-off ratio of log-compute, so even large trade-off ratios (e.g. 5 orders of magnitude to 1) lead to fairly similar compute allocations (20% to training).

[22] One example where the training-inference tradeoff cannot be made for a specific developer is open models, where the developer only pays for training compute, but inference costs are borne by users. However, open models tend to lag behind the frontier, so a different pattern of scaling would have little bearing on the trade-off at the frontier.

Investment

Investment is necessary to provide compute. If compute is to continue scaling on trend, it will require investment of hundreds of billions of dollars by 2030. This is an extreme requirement, but it would be justified if investors believe that AI will provide significant economic benefits. If AI could raise productivity across the economy, it would eventually generate trillions of dollars of economic value, justifying these large investments. This matches present-day trends in AI revenue growth. Moreover, we can already see that present-day investment patterns and spending plans are consistent with scaling until 2028.

Frontier model training costs will likely continue growing 2-3x per year



The costs of training frontier AI models have increased 2.5x per year, and are set to continue.

AI training costs have become steadily more expensive. Currently, frontier AI models require hardware investments of billions of dollars, plus significant energy and labour costs. The amortised cost of compute is reaching hundreds of millions of dollars, with no indication of slowing.

This is fairly likely to continue until 2030. As we discuss elsewhere, the next generation of AI clusters are already priced in for 2028, suggesting relevant actors are currently willing to invest for at least three more years of scaling.

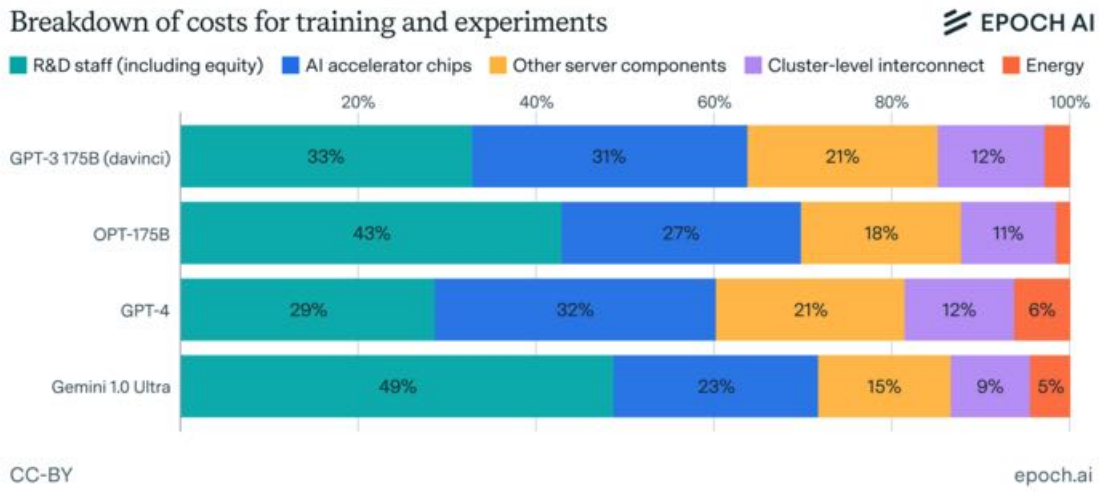
What might disrupt these investment trends? Clearly there are external events that could upset investment trends, ranging from AI regulation to a broader economic downturn or even war. Aside from these, a downturn in investment might look like companies pulling back from compute

investments for training, focusing on lower-cost inference. This might happen in a world in which capabilities advances were relatively stalled. In such a world, investors might not expect to capture much value from large training investments, and returns might accrue mostly from serving AI tools at close-to-existing capabilities levels.

Plausibly, the trend could shift upwards if continued AI scaling leads to increased market confidence in AI's future returns. In a simplified macroeconomic model where a sufficiently large training run can automate all work tasks, optimal AI investments rapidly scale to double-digit percentages of world GDP (Erdil et al. 2025).

Overall, there are fairly strong reasons to think present investment trends will continue. Investment trends could be justified if AI is expected to significantly improve economic productivity. Moreover, current spending plans from AI developers and chipmakers are consistent with current trends, supporting the prediction that scaling will continue to at least 2028.

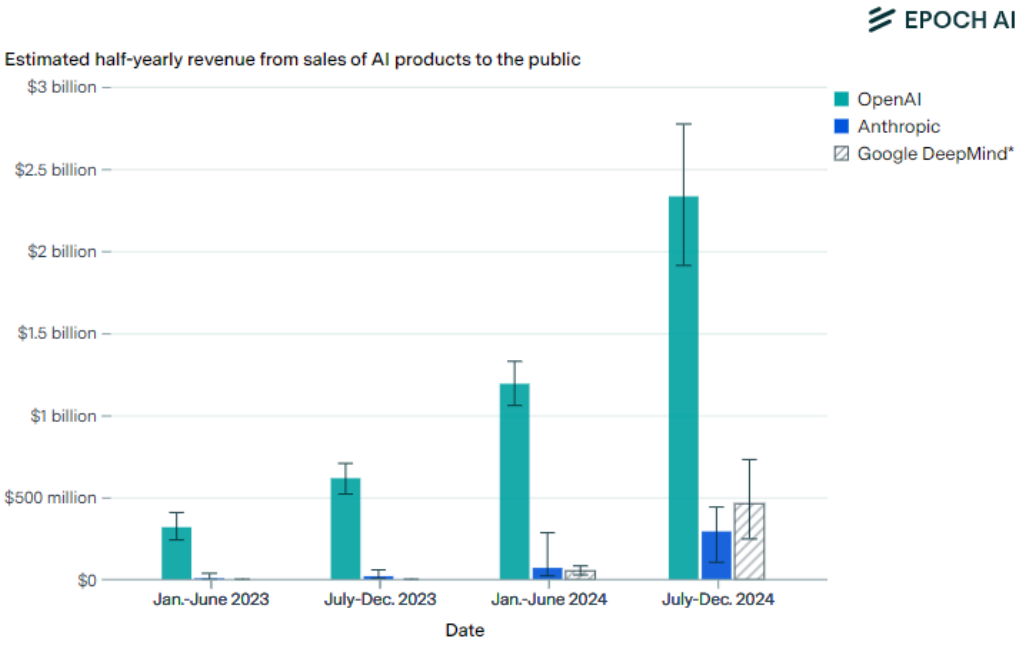
Compute already accounts for half of development costs, and is likely to increase



Cost estimates suggest that compute (AI chips, server components, and interconnect) is the largest cost in developing frontier AI. Cost estimates are based on publicly-reported information (Cottier et al. 2024).

Estimates of model development costs suggest that compute is the largest single cost. This includes compute for experiments as well as large training runs. The largest non-compute contributor is the cost of labour: researcher compensation accounts for a significant portion of spending (Cottier et al. 2024) and recent reporting suggests researcher salaries may increase further (Isaac et al. 2025). If compute scaling continues on trend, it will presumably grow as a fraction of spending compared to R&D staff. There is significant uncertainty here, as there is little public data on the growth of R&D staff costs.

If AI revenues grew on trend, they could match these investments



The largest AI developers are already estimated to earn billions of dollars per year, and these revenues have been growing around 2-3x per year in the past few years. Projecting this trend to 2030 suggests revenues of hundreds of billions of dollars.

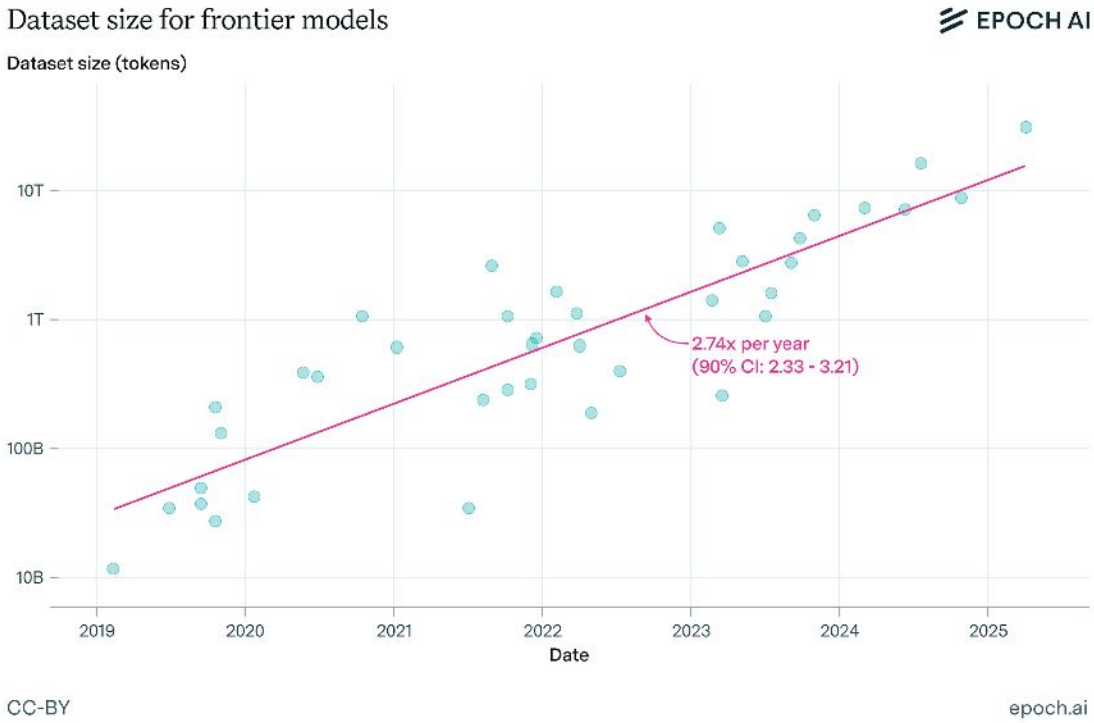
A common objection to a scaling-focused view of AI is the scale of the required investments. Would AI developers really invest hundreds of billions of dollars to create large-scale compute infrastructure? A relevant source of evidence is that existing AI revenue trends match these investments. If this revenue growth continues, AI developers would make these large investments with clear evidence for their value at each stage. This is also consistent with revenue projections for AI hardware: if NVIDIA's revenue grows to match their current price-to-earning ratio, then at current margins its annual revenue would need to grow to about $200 billion. This would suggest even more than this being spent on AI services (Todd 2024).

Data

Data is vitally important to AI development in several ways. First, and arguably most important: large, general-purpose datasets have been vital for scaling pretraining of generative systems in language, images, and other modalities. Increases in training compute have come from increases in both model size and dataset size. Dataset scaling is may be even more important for progress today, as model sizes can only scale efficiently with more data.

Second, there is the necessity of specialist data of various kinds. For general-purpose AI models, specialist data is used to post-train a base model to create a more user-friendly and safe chat model. Specialist post-training data is also important to improve performance on widely-useful skills such as reasoning, coding, and planning. For narrower applications, such as protein structure prediction, there is a straightforward need for models to be trained on corresponding domain-specific data.

Datasets will continue growing, but with a different composition



Training data for language models has increased 2.7x per year.

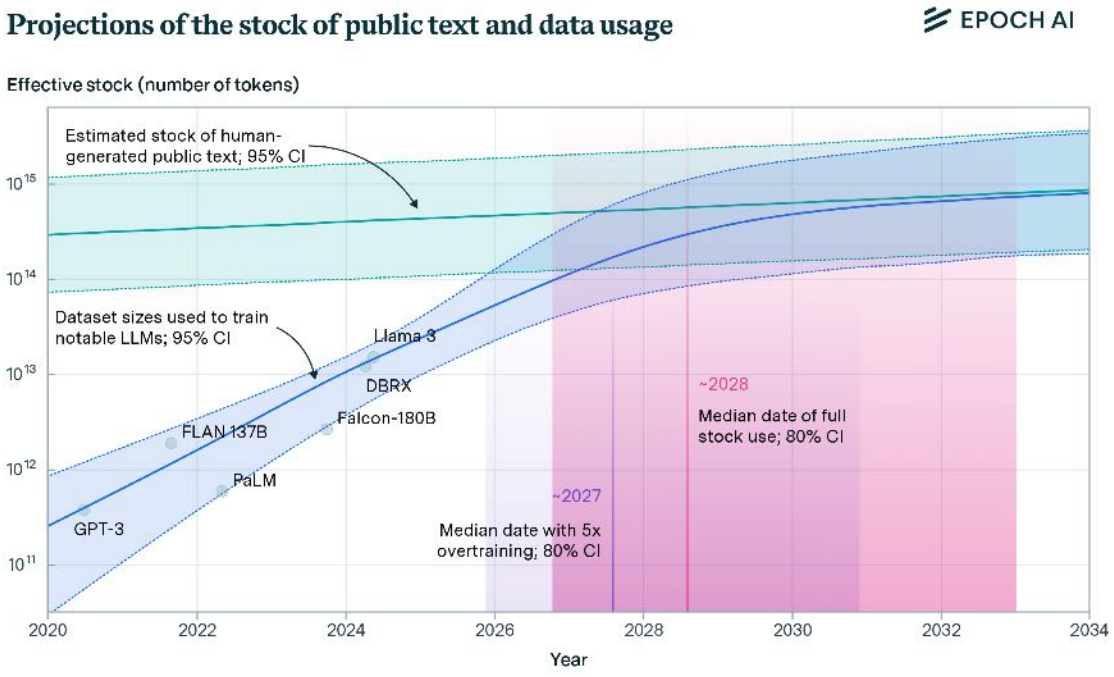
Frontier language models have seen their training datasets grow 2.7x per year. Earlier language models were trained on specific corpora, for purposes such as summarisation or question-answering. The original GPT paper marked a lasting shift towards large-scale general-purpose pretraining. Subsequently, LLMs began to be trained on increasing quantities of text scraped from across the Internet.

In order to continue scaling up training compute, companies are likely to continue growing dataset sizes, although the precise composition of datasets could change significantly, as we discuss below. In particular, training compute may shift towards reasoning training, which uses smaller

quantities of human-generated data.[23] This suggests dataset growth may slow over the coming years, at least when we limit ourselves to counting human-generated data.

[23] For example, DeepSeek R1 reasoning training was estimated to use tens of millions of tokens in human-generated verifiable data, but this led to trillions of tokens of generated data for RL training.

Data won’t run out by 2030, although human-generated text might



The stock of human-generated public text is large, estimated between 10^{14} and 10^{15} tokens, but on current trends frontier training runs could use the entire (human-generated public text) data stock before 2030, especially if models are overtrained.

Data usage trends suggest that the stock of publicly available text data could soon be exhausted. However, there are two important counterarguments to this: multimodal data and synthetic data.

General-purpose models are increasingly trained on multimodal data such as images, videos, and audio. How to measure the equivalence of these data to text is unclear. If we make assumptions based on existing tokenisation schemes, multimodal data might increase the public data stock 10x or more. At historic rates of compute scaling, this would allow pretraining datasets to expand in size until 2030. For this to happen, training on such data would need to provide commensurate improvements in

valuable AI capabilities. Currently, performance in non-text modalities is arguably behind text: visual question-answering benchmarks are below human performance on simpler questions than pure language benchmarks. Hence, we should expect significant scaling of multimodal data in the near term, but it is uncertain whether labs would continue this further.

Synthetic data has recently grown in importance, because it is widely believed that the most recent generation of frontier LLMs are making heavy use of it. In many domains, it is easy to verify solutions even when they are difficult to generate - for example, software engineering problems with tests. In other domains, it may be difficult to verify solutions with high confidence, but existing LLMs may be capable of acting as a judge. It is uncertain how broad and enduring the benefits from synthetic data will be, but current progress suggests it will be an important direction for further scaling.[24]

In early 2024, OpenAI was generating on the order of 100 billion tokens per day – and since then, usage has likely increased. This represents a plausible quantity of synthetic training data that they could generate. It would suggest growing the available data stock by tens of trillions of tokens per year. Moreover, training on synthetic data requires more compute than simply training as-is – it requires multiple inference passes to have a model propose steps, compute to simulate the environment, and potentially inference for a judge model to provide RL signal.

In short: it is likely that traditional pretraining text data sources will soon be exhausted, but this is not anticipated to prevent further compute scaling. As long as either multimodal data or synthetic data proves tractable and worthwhile, there will be enough data to scale to 2030 on current trends. If synthetic data proves particularly generalisable, then general-purpose “data scaling” may never become a bottleneck.

[24] Why might synthetic data be limited? In short, LLMs can only generate data reflecting their learned distribution, and perhaps important capabilities are out-of-distribution. There is even some early evidence characterising this, comparing reasoning post-training with distillation from a larger pretrained model (Yue et al. 2025). This area remains uncertain, but synthetic data of some variety seems likely to play an important role – notably, synthetic data can be generated using more than just a base LLM’s learned distribution.

Specialist data will become increasingly valuable if scaling continues

This raises the question of whether some kinds of data will become more valuable. There is little empirical evidence to draw on, but we can point to some intuitive implications for data if scaling continues to improve AI capabilities.

For general-purpose AI models, this suggests that the following data would be important:

> *Challenging problems with easily verifiable solutions, with value for economically valuable capabilities.* A canonical example here is challenging but easy-to-verify software engineering problems. These can be used to generate synthetic data for reasoning post-training, and hence would be particularly valuable (Rachitsky 2025). There is already evidence here from AI developers' focus on developing challenging benchmarks.
>
> *Data that disproportionately improve "soft" skills of the model, e.g. style and tone, to the extent that these aren't addressable by synthetic data.* This is fairly speculative, but has some evidence so far: AI companies hire researchers to focus on optimising system prompts, and invest in large pipelines to prepare curated example data.
>
> *Data that expand the model's knowledge in valuable areas, particularly if they do this more efficiently than synthetic data alone.* For example, many AI developers are building AI coding assistants. Data that address key limitations, such as worse performance in a programming language with less public code for pretraining, would be valuable.[25] It is unclear whether this would best be achieved through synthetic data or more collection on existing data. An important consideration is whether models will become more proficient at using search tools to augment trained-in knowledge. In

[25] Many AI developers are focusing on coding as a key application of their products. However, performance is unequal across languages, presumably reflecting the training data.

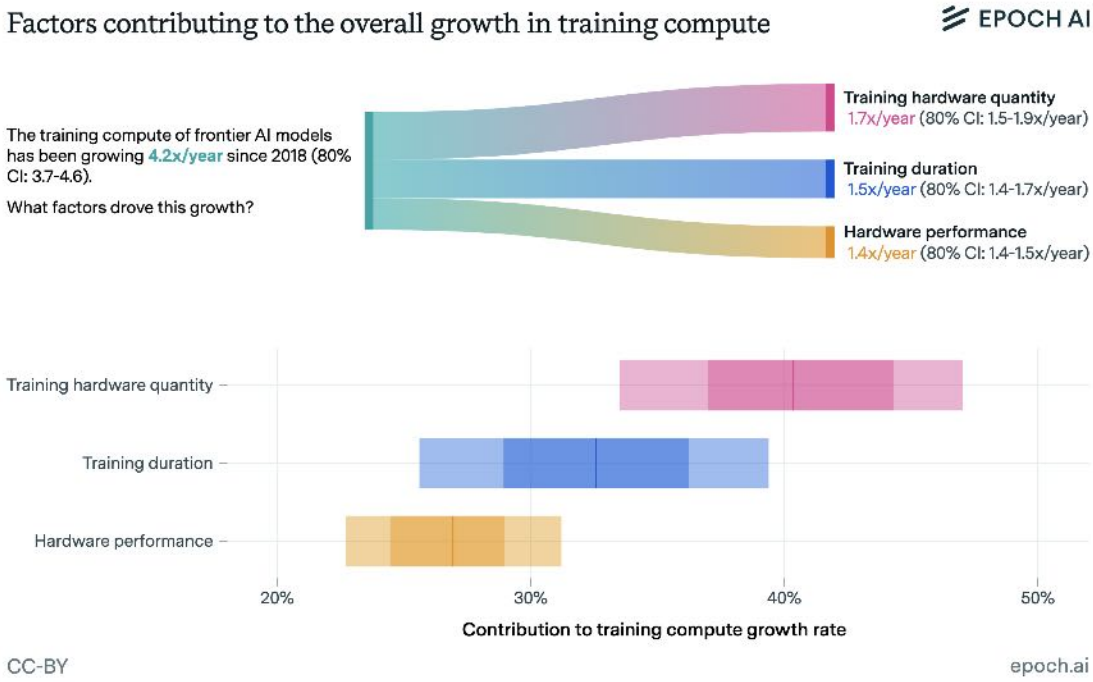
this case, the structure of value changes: data that are available via search become less important for AI developers to collect.

For narrower AI models, the general answer is “data covering the application”. In this report, we focus on scientific R&D. A relevant example is biomolecule structure and interaction data. Pre-existing databases of protein experimental structures were essential for training AlphaFold, and similar databases will need to cover a broader range of molecules and their properties. In general, such data are likely to be crucial when (i) there is no reason to expect transfer learning from other data; (ii) it requires specialist knowledge, skills, or equipment to collect.

Hardware

Specialised hardware has been essential to the development of modern AI, and remains a key driver of progress. We show how most training compute growth has come from scaling up cluster sizes, and argue this is likely to continue in the next generation of clusters. Finally, we discuss how distributed training may make it easier to continue scaling, reducing the need for colocation of compute.

Training compute growth will come from AI chips, probably not from longer training runs



Historically, compute growth has mostly come from increasing the quantity of training hardware. Increasing durations and hardware performance have made respectively smaller contributions.

For frontier AI models since 2018, most compute scaling has come from running more accelerators in parallel, i.e., increasing cluster sizes. Hardware performance improvements have contributed less than increasing cluster sizes or training duration.

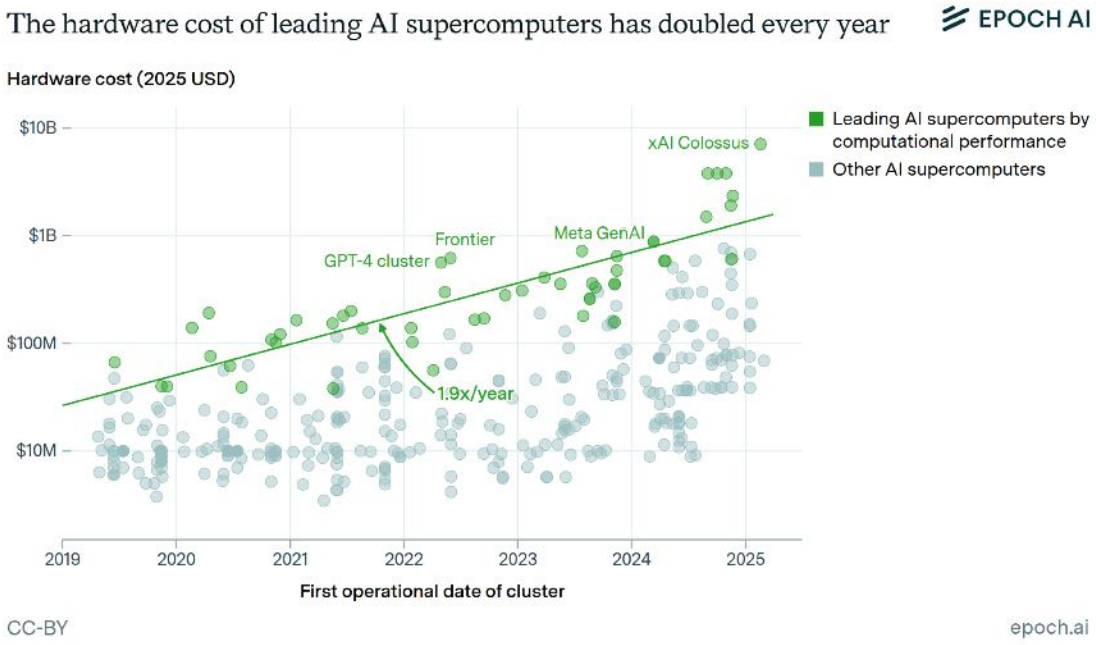
Algorithmic and hardware progress disincentivise long training runs. If a training run is too long, the model risks being overtaken by training that starts later and benefits from these (Sevilla et al. 2022). This suggests that training runs face limits, and may not grow much further than today's typical duration of months.

Meanwhile, AI hardware will likely continue improving. Theoretical analysis suggests there can be at least 50x further improvement, extending far

beyond 2030 on current trends (Ho et al. 2023). Of course, problems could arise sooner in practice. Similar to how Dennard scaling ended for microprocessors, AI chip progress could slow. There is no evidence of this yet, and the forthcoming chip generation suggests progress is likely to continue.[26] Scaling up the number of deployed AI chips is also likely to continue, as we discuss below.

[26] GPU performance per dollar has increased 1.3x per year on average. The NVIDIA B100 has about 1.7x the performance of the H100 for a fairly similar price at launch, about 1.5 years later.

The next generation of AI clusters is already priced in



The upfront investment required to buy hardware for AI clusters has increased 1.9x per year. Large clusters with billions worth of chips are already being constructed.

The next generation of AI clusters provides useful evidence about continued scaling. The largest NVIDIA-based AI clusters are already constructed with over 100,000 H100 GPUs, and larger clusters are in construction over the next year. This strongly suggests that compute scaling will continue for at least one more generation of AI models.

There are four main reasons this trend could change: changes in willingness to invest, breakthroughs in hardware, breakthroughs in training, or clusters being repurposed from training to inference. We have already discussed the significant motivation to continue investing in AI development, if capabilities keep scaling. We have also discussed how repurposing clusters towards inference seems unlikely to happen on a large enough scale to slow training trends.

This leaves the question of breakthroughs in training efficiency, either in algorithms or hardware. Such possibilities certainly exist, but given the

sustained trend of scaling so far, they do not fit into our default predictions. If an algorithmic innovation as large as the Transformer architecture did not disrupt trends, it seems unlikely that such disruption will occur in the next five years.

Training is likely to become distributed across multiple clusters

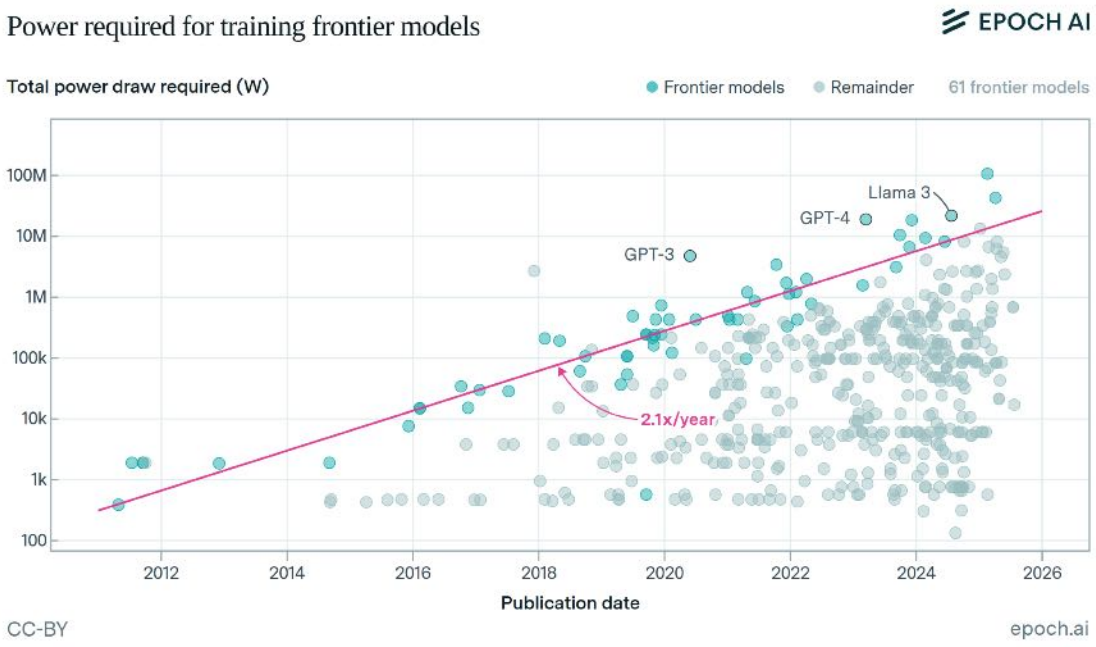
At the same time as individual AI clusters are reaching unprecedented sizes, changes in AI development could make scaling easier by relaxing the need for colocation. Until recently, frontier AI models have been trained using individual clusters. Several developments suggest this is changing – for example, multi-datacentre training officially reported as a contribution in the development of Gemini Ultra. This suggests that the number of chips involved in a training run will continue to increase, though not necessarily in a single datacenter.

A shift towards synthetic data and inference scaling could even further favour multi-cluster training. If synthetic data generation requires large amounts of inference compute, then this might more easily be done across multiple sites. If inference scaling increases the quality of this generated data, then this could favour intensive inference scaling even for the purposes of training.[27]

[27] Other factors might push in favour of centralisation, for example the increasing need to secure model weights. However, it is harder to anticipate the overall impact of such goals; securing several training clusters, and encrypting communication between them, might be sufficient.

Energy and the environment

Power demands are likely to continue growing 2.1x per year



Power for frontier training runs has grown 2.1x per year, and the largest training runs of today are already using over 100 MW. If this trend continues, the largest training runs could require approximately 10 GW by 2030.

Frontier AI training power demand has grown 2.1x/year. Training durations have also increased in this time, so total energy use for frontier training has increased about 3x/year. These numbers reflect the trend in individual models' training – most organisations have trained multiple models in a year, as well as using compute for experiments.

Inference, meanwhile, is less well-documented. AI developers focused solely on large frontier models, such as OpenAI, reportedly spent similar amounts on inference and training (Snodin et al. 2025). Developers who

deploy much smaller models, such as Meta, report dedicating approximately 20% of their total AI compute to large-scale training clusters. However, it is uncertain what percentage of the rest is dedicated to inference, as smaller models also require training (Wu et al. 2022).

The overall amount of power dedicated to AI (training and inference, including non-frontier models) is harder to track, let alone predict. Projections based on AI chip production and hyperscaler capital investment plans suggest somewhere between 1.5x-2x per year (You and Owen 2025). This is nevertheless compatible with individual training runs growing faster: they may simply grow to occupy a larger fraction of total AI energy usage.

The required electrical power is a significant challenge for frontier AI training runs. If scaling continues on trend, the largest training runs will require about 10 GW by 2030. Such power consumption exceeds the generation of all but the largest power plants, and would present enormous organisational challenges. This might lead to a slowdown: perhaps scaling beyond low gigawatt training runs is logistically unachievable by 2030. On the other hand, as previously discussed, frontier AI training runs are already beginning to be geographically distributed across multiple datacentres, which would temper the challenges. Moreover, there are ways to rapidly scale up power delivery, such as solar and batteries, or off-grid gas generation (Datta and Fist 2025). If the demand for AI scale-up continues on its current trend, the largest training runs should at least reach multiple gigawatts, matching planned cluster build-outs (You and Owen 2025).

If compute trends continue, emissions from AI would grow to 0.03-0.3% of the world's projected total

So far, AI appears to have increased net carbon emissions via increased datacentre energy consumption. For example, Google's base carbon emissions before offsetting increased by 48% between 2019-2024. Much of this increase came from AI datacentres (Google Sustainability 2024). Datacentres have two emissions sources: datacentre embodied emissions (construction and hardware manufacture), and operational emissions (datacentre energy supply). In this analysis, we focus on energy supply,

estimated to make up over 70% of total emissions for leading AI datacentres (Google Sustainability 2024).

Depending upon the energy mix that will supply datacentres, current trends in compute and datacentre energy consumption suggest that AI might make up between 0.03% and 0.3% of global emissions. This would be a substantial addition, but nevertheless smaller than the existing emissions from all datacentres, AI and non-AI (180 million tCO2e in 2025) (IEA 2025c). The lower end of this range is an aggressive lower bound, essentially relying on massive solar power provision. The higher end of this range is based on the current average carbon carbon intensity of the grid, which is also close to the carbon intensity of natural gas.

WORKED CALCULATIONS FOR AI EMISSIONS UNDER CURRENT TRENDS

Assumptions

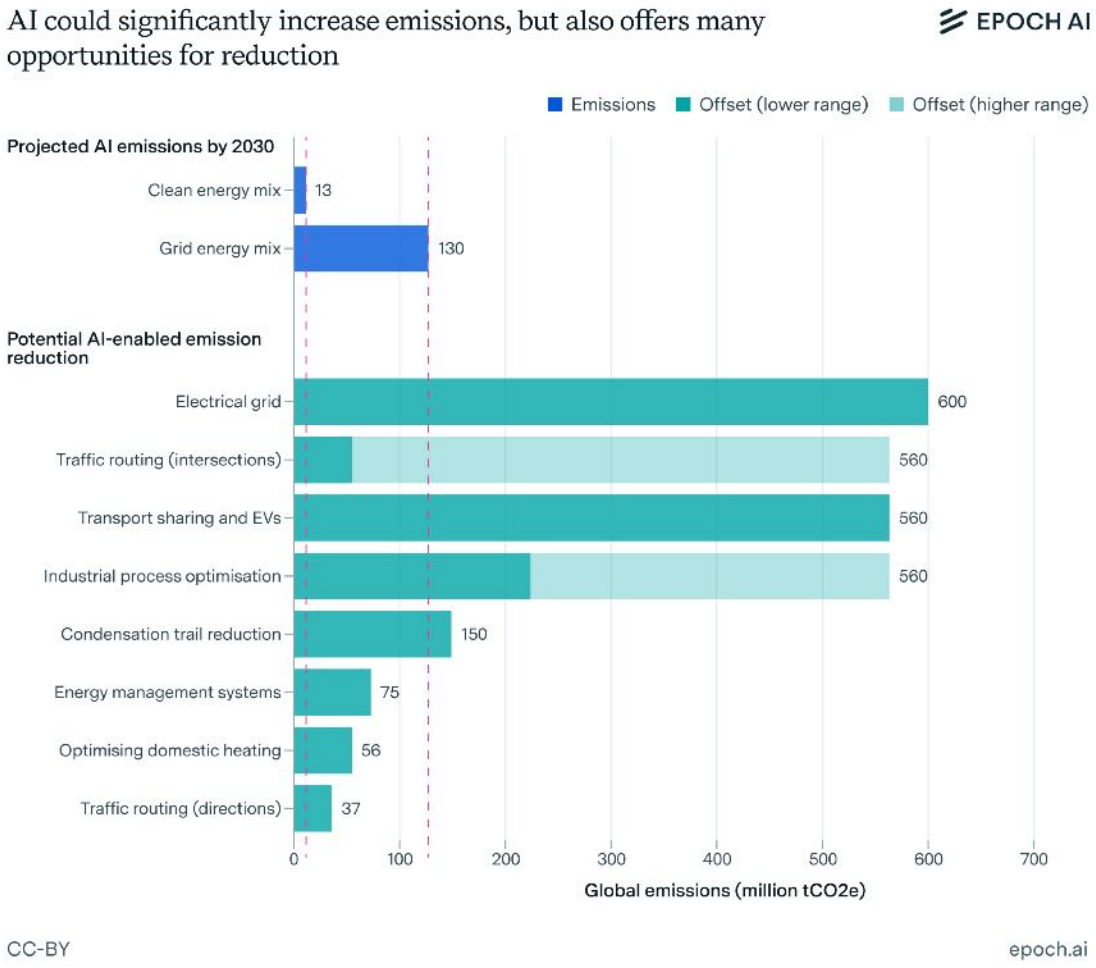
- We focus on operational emissions, i.e. not embodied emissions from manufacturing and construction.
- We assume the starting point is that AI datacentres used approximately 10 TWh in 2023. This was from estimates based on 2023 NVIDIA hardware sales, taking the higher side of the estimated range.
- Alternative check: 3.7M H100-equivalent available by mid-2024. This is about 23 TWh if these were all H100 efficiency (700W) and run nonstop. It makes sense that this is higher than the earlier 2023 estimate, so they're fairly consistent.
- We assume starting total world electricity demand is 23,000 TWh, and other than AI this continues to grow at its recent trend of 2.7% per year to 26,300 TWh by 2030. In this sense, this is a conservative estimate – if non-AI growth is faster than this, as it may be with the roll-out of electric vehicles, AI would occupy a smaller fraction of the total (IEA 2025b).
- Total global CO2e emissions in 2024 were 37.4 billion metric tons. We pessimistically assume that non-AI emissions grow on trend at 7%/decade to 38.7 billion metric tons.

If AI power demands continue growing on trend, how much would AI contribute to global emissions?

- Assume dedicated AI power consumption doubled annually, i.e. following current trends linked to training compute growth of 4-5x per year. By 2030 it would reach 640 TWh. This would be 2.4% of total electricity demand.
- An energy mix at the carbon intensity of the global grid average (400g CO2e/kWh) would suggest emissions of 124 million tCO2e. This is a pessimistic estimate; most datacentres use a lot of renewables. This pessimistic estimate would make up ~0.3% of emissions in 2030.[28]
- If the carbon intensity were instead that reported for solar panels (40g CO2e/kWh), this would be ~0.03% of emissions in 2030.

[28] The most pessimistic imaginable scenario for AI emissions is if the net effect were at the carbon intensity of energy sources such as coal. This could conceivably happen if the AI energy requirements led to existing coal power stations remaining in use in developed countries, and more coal power stations being constructed in countries such as China. Still, given the track record of datacentres so far, and the broader societal shift to clean energy sources, the current grid average seems a more persuasive lower bound in practice.

Energy for AI development will increase, but there is scope to reduce emissions



Projected AI emissions by 2030 could be significant (top), although there are many opportunities for AI applications to reduce emissions (bottom). Most of the projected AI emissions (~95%) are from post-2025 growth. Reduction estimates are approximate calculations of potential reduction, i.e. the full extent of what might be achieved based on evidence from existing programmes and studies. See Appendix: AI's potential to reduce GHG emissions for more detail.

A natural question is whether AI's increased emissions could be meaningfully offset. There are three potential ways this could happen: (i) a

large increase in low-carbon energy, fully offsetting AI datacentres; (ii) AI algorithms and hardware sufficiently improve in efficiency to change the trend in energy consumption; or (iii) AI applications downstream are able to lower carbon emissions elsewhere enough to offset the increase.

AI datacentres already make significant use of renewable energy from solar, wind, and hydroelectric power, although they often use other sources for a reliable base load to complement intermittent solar or wind. However, the question is not as simple as what energy mix datacentres will use. If datacentres relied on renewable energy, but displaced other demand to non-renewable sources, the net effect would be to increase emissions. Hence the question is more involved: could renewables be increased quickly enough to match AI demand growth?

Building enough clean energy capacity to cover demand from AI datacentres would be challenging, but plausible. Renewables are projected to grow from 30% of global electricity generation in 2023 to 46% by 2030 (IEA 2024). Under the IEA's accelerated timeline for renewable energy transition, this could instead reach 60% by 2030. Hence the projected electricity demand from AI (1.2%) is smaller than societal choices on energy transition. The difference between the projected 46% and feasible 60% renewables share of electricity is twelve times larger than the projected demand from AI.[29]

AI algorithms and hardware improving in efficiency seems likely to continue. However, the history of AI so far should give us pause. AI methods have already improved orders of magnitude in efficiency. However, this has happened in parallel with a massive increases in power consumption. As long as there is a strong incentive to scale up hardware for training and/or inference, efficiency improvements seem unlikely to reduce net energy consumption. The historic trend of efficiency improvements is implicitly included in our estimates above.

[29] The analysis of renewable energy's potential to offset AI emissions becomes tighter if we limit its scope to the US and China, assuming that they will host most datacentre activity and must provide their own energy. Still, the US is currently constructing approximately 40GW of renewables per year, after intermittency, which is about 2% of current average generation. This is similar to the projected extra demand from AI, similarly suggesting that there is still scope to offset it.

Downstream applications of AI are the most difficult question to answer decisively. Could AI reduce emissions elsewhere in the economy sufficiently to offset the emissions for which it is responsible? This will depend greatly on the emissions due to AI (how much compute is used, and what energy mix underpins it), and the downstream impacts of AI models (what GHG emission can they avert?)

If AI became responsible for a meaningful fraction of total emissions, it would be challenging to sufficiently offset impacts in other areas. Conversely, for AI to reach such an incredible level of energy consumption, it would need to be extremely valuable, so we would expect its societal impacts to be large. It is difficult to systematically account for all the possible ways that AI could reduce emissions. It seems plausible that AI could be used to reduce emissions more than it causes, averting single-digit percentages of current global emissions.[30] For example, AI can be used to better forecast power supply and demand in the electrical grid, allowing for more usage of renewables; or AI can be used to optimise transport sharing and routing, reducing emissions from cars. This would of course be highly dependent on deployment and prioritisation, and relies on aggressive estimates of what can be achieved.

[30] In many cases, the AI models considered for climate applications would be much smaller than frontier AI models. This points to a potential weakness in trying to assess AI's overall effect: there are many different things that might count as AI. We cover these applications in more detail in Appendix: AI's potential to reduce GHG emissions.

Interlude: From scale to capabilities

Evidence: benchmarks, current AI usage, and domain experts

We have argued that, on current trends, the largest AI models of 2030 will be trained with 1,000x more compute than today. By 2030, we will have seen a jump comparable to the scale-up between GPT-2 and GPT-4. Given such continued scaling for AI development, how can we reason about what AI in 2030 will be able to do? To ground our discussion, we use three sources of evidence: AI evaluations, usage of present AI systems, and predictions from domain experts.

As discussed in Scaling and capabilities, AI evaluation performance tends to improve fairly predictably with scale, once performance begins improving beyond random chance (Owen 2024a). Therefore, we focus on extrapolating from evaluations that show progress so far, and highlighting evaluations that are entirely beyond current AI, and hence are less predictable.

For many domains such as software engineering, highly relevant benchmarks exist, with a fairly clear link to real world problems, and can be used for extrapolation. For other domains, benchmark coverage is less clearly representative of real-world work tasks. Still, we can often see evidence from individual examples of AI capabilities – for example, there may be no systematic benchmark for AI-assisted protein design, but individual results inform us on what AI can currently achieve.

There is an ever-present risk that benchmark results are not reflective of real world performance – through lack of representativeness, through models overfitting on the metric, through test set contamination, etc. Nevertheless, benchmarks play a crucial role in developing AI systems, and

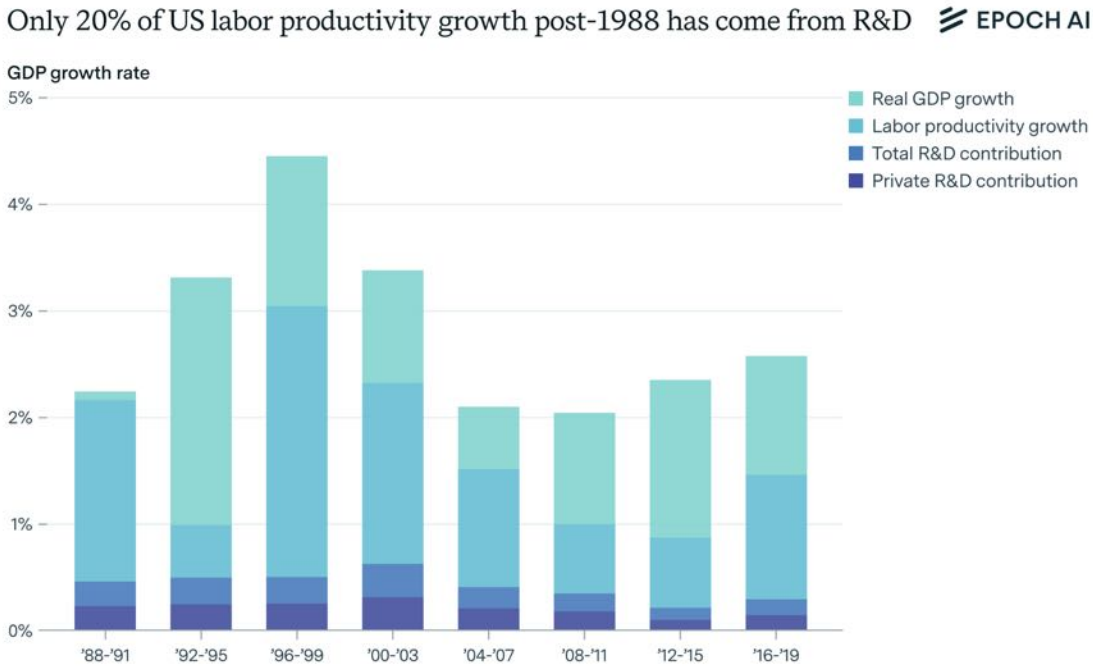
are a valuable signal for tracking progress. Interpreted with care, they should be informative about AI capabilities.

Meanwhile, usage of existing AI systems is strong evidence of real-world usefulness. This is not always future-facing – it requires that existing AI systems are already useful, ideally in a format that matches what we are trying to extrapolate. For example, software engineering AI is already widely used as a code assistant under close human supervision. This is strong evidence that code assistance is useful to programmers, and that improvements are likely to offer further benefits. However, this offers only weak evidence about the timeline on which independent AI coding agents will become practically useful. Nevertheless, where it is present, real world usage suggests AI really is ready to contribute to a field.

Finally, another valuable source of evidence comes from domain experts. Many researchers are actively experimenting with present-day AI systems, and reflecting on how future AI may change their work. Predictions from domain experts provide valuable information on the effects they foresee AI having – and the potential obstacles they see for integrating AI into their work.

By synthesising these sources of evidence into qualitative descriptions of AI capabilities, and how they might operate in an R&D domain, we paint a picture of how AI could accelerate scientific R&D by 2030. But before examining these more specific predictions, we discuss the broader context in which they will be situated: AI-enabled automation of many tasks across the economy.

Broad automation across the economy or a focus on R&D?



Discussion of AI's future impacts often focuses on exciting application areas, scientific R&D being a prime example. That includes this report: in the second half, we will discuss AI's transformative potential across a range of scientific R&D domains. However, there is an argument that in the short and medium term (years or even decades), larger economic effects will come from broad automation across the economy. This informs how we think about automation of R&D tasks.

Explicit R&D accounts for about 20% of US labour productivity growth in recent decades, and in turn this only accounts for about half of real GDP growth. In comparison, capital deepening accounts for about half of labour

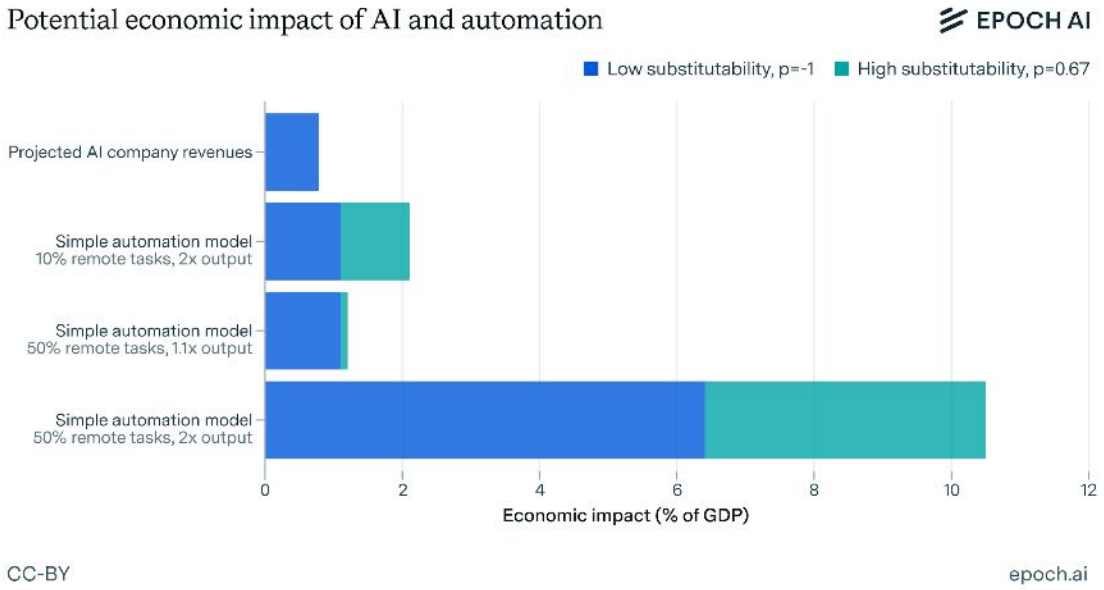
productivity growth.[31] The rest of labour productivity growth is attributable to "better management, learning-by-doing, knowledge diffusion, etc" (Erdil and Barnett 2025). In other words, to maximise near term economic output, it is more effective to scale up the effective labour force (automate many tasks across the economy and run more of them). Since skills required for R&D overlap with these other tasks, it appears likely that broad automation will at least happen in tandem with R&D automation. R&D labour is particularly valuable labour, and hence *a priori* we should expect more effort dedicated to automating it, but not an *exclusive* effort.

Essentially, this is an argument that automation will be a diffuse process across the economy. To give a concrete example, think of software engineering. There are many software engineering tasks in R&D, and some of these are even being automated by AI today. By volume, however, most of the software engineering work that is being automated is not in R&D; it is helping backend and frontend developers develop commercial and hobbyist software. The automation of R&D software engineering may eventually contribute more to new technologies that have a greater impact in the long run, but based on historical examples this would be a longer timescale, and would not be the sole or even primary focus of deployment.

AI R&D may be particularly salient and well-represented for automation, however, and there are arguments that it will be among the first areas to see automation. Potentially, AI R&D is a domain where there are increasing returns to research effort, and automating it leads to a positive feedback loop (Erdil et al. 2024). This is the outcome envisioned in discussions of a software-only singularity, such as AI 2027 (Kokotajlo et al. 2025). In this report, we have focused on a scenario where algorithmic innovations are complementary with compute, and automation of AI R&D overlaps with automation of other tasks in the economy. Nevertheless, it is difficult to rule out rapid automation of AI R&D, and it is a key way in which AI progress could move faster than our projections.

[31] Of course, for capital deepening to occur, R&D must happen in the first place to develop useful forms of capital. For example, hospitals can be made more productive by deploying advanced MRI machines, but those MRI machines were developed through cycles of explicit R&D, technological diffusion, and learning-through-doing. Still, in recent history, more economic growth at any point in time has come from deploying already-developed products, compared to embarking on new explicit R&D projects.

AI automation across the economy could be worth trillions of dollars



Different high-level estimates of economic value AI can generate through automation. Projected AI company revenues (top, discussed in Investment) are broadly consistent with the economic value from a widespread productivity boost, i.e. a ~1% increase in GDP. Meanwhile, a more aggressive model of task automation (discussed below in `Estimating the economic value of AI`) suggests that doubling output for 50% of remote work tasks could increase GDP by 7-10%.

How valuable might AI be, if it were deployed across the economy? We consider several different estimates based on doubling output from non-physical work tasks. The answer is highly contingent on AI capabilities and deployment, but widespread AI tools could plausibly generate trillions of dollars of economic value, simply by improving productivity of non-physical work. Even if such gains were not yet *achieved* by 2030, investments could be made on the basis of beliefs about AI's capabilities and eventual deployment, rather than their deployment by that point in time.

Could we realistically see mass automation rapidly accelerating 50% of remote work tasks by 2030? Based on economic history, there are reasons

for scepticism, even if it were technically feasible. It would require incredibly rapid integration of technology into work. This might be accelerated by AI helping coordinate the reallocation, but nevertheless seems hard to imagine. On the other hand, AI technologies have seen some of the fastest adoption of any technology in history (Hu 2023). Moreover, on a longer timescale, there are strong incentives for deployment, given the substantial economic advantages. If AI revenues continue their current trajectory, then by 2030 it seems likely that either AI will be generating trillions of dollars in economic value, or this will be within sight.

The above projections assume a situation ranging between "AI is a helpful tool for half of remote tasks", and "AI can fully automate half of remote tasks". For a more extreme forecast, consider the implications of a future in which AI can perform any task a remote worker can do today. Is it so far-fetched to imagine that billions of such AI remote workers would achieve large effects in the economy? Technology companies contribute about 10% of present day GDP. Such AIs could presumably create similar organisations, and this already begins to resemble the aggressive projections above.

ESTIMATING THE ECONOMIC VALUE OF AI

We consider a task-based model of automation, similar to existing literature (Barnett 2025; Acemoglu and Restrepo 2018). Labour is allocated to tasks, which taken together produce economic output when combined with capital. Automated tasks have their "effective labour" increased by a multiplier to reflect the effect of AI. We use this model to predict the effect of productivity gains across the economy for widespread AI tools and assistants.

In such a model, tasks are somewhat complementary. For example, imagine a software company that doubles its labour inputs for software engineering tasks (perhaps by hiring more software engineers) but without increasing its labour inputs for sales and advertising tasks. Would such a company double its revenues? Perhaps if the original software labour inputs were far below demand, but generally we should expect that sales would become an increasing bottleneck.

The complementarity between tasks is not obvious. In our modelling, we separately consider both fairly substitutable tasks and fairly strong complementarity, based on the range of values in the literature.[32] This leads to larger and smaller gains from automation respectively.

In the US, approximately 34% of work tasks are estimated to be remote-compatible (Barnett 2025). We use this as a proxy for AI exposure, assuming advanced AI could accelerate some fraction of remote work.[33] We examine different fractions of these being automated, ranging from 10% to 50%.[34,35]

To simplify analysis, we assume the cost of compute for automated tasks is small compared to human wages. There is existing evidence of this in tasks that AI can currently

[32] In practice, for real tasks, complementarities will vary by task and sector, but this broad assumption is common in the literature.
[33] In a less developed economy, a higher fraction of occupations may be incompatible with remote work. However, they also contribute correspondingly less to GDP.
[34] What does it mean when we talk about a percentage of tasks? In a simplified model like this one, we assume there are N equal tasks, and consider a percentage of these. In a more complex model, tasks can be weighted according to features such as their prevalence in an occupation, or the salaries of the occupations in which they are included. Our simple model should suffice for approximate calculations however; in practice, there will be incentives to automate where automation brings most value.
[35] A common concern about this line of reasoning is whether there would even be enough inference compute for so many tasks to be automated. If we assume that a virtual worker requires 10^{14} FLOP/s, similar to some of today's leading models, then requiring 10% of the global population (a third are in the workforce, and they work a third of their time) corresponds to about 10^{23} FLOP/s of compute capacity required. On current trends, by 2030 there will be 2.6×10^{23} FLOP/s installed compute capacity for NVIDIA chips alone.

perform.[36] Moreover, a given level of AI model output rapidly becomes cheaper to generate, based on observations of inference prices for LLMs from the past several years (Cottier et al. 2025). We ignore further possible gains from reallocation of human labour. We model the economy as a fixed set of tasks, with a fixed allocation of human labour, which sees productivity increase in automated tasks.

Doubling output in 10% of remote tasks would give a 1-2% increase in GDP, producing trillions of dollars in economic value. Increasing output by 10% in half of remote tasks would have a similar effect (~1% of GDP). Doubling output in half of remote tasks would lead to a 6-10% increase in GDP.

The key question lies is how quickly this growth would be realised. Our economic model says nothing about the timeline over which such effects occur. This would depend on deployment and adoption. Projected AI revenues are consistent with *spending* on AI roughly in line with 1-2% increase in GDP, suggesting the timeline could be as soon as 2030. However, spending could pre-empt effects in output, arguing for longer timelines. How far in advance could spending happen? A relevant example is investment in the web, where it took about a decade for ecommerce sales to match the IT company investments of 1999.[37] It seems safe to assume that "decades" is a pessimistic upper bound, as long as AI actually does achieve the necessary capabilities.

A common objection is that growth projections from automation are too optimistic because they fail to consider Baumol and Engels effects. Both of these are effects that reduce the value from productivity improvements, because productivity improvements change the relative value or structure of different parts of the economy. We explain each further below.

Baumol effects limit economic gains from increased productivity when stagnant economic sectors demand similar wage increases to sectors that see automation. This can also be understood at the level of tasks: the tasks that are difficult to automate end up becoming more economically important, and the tasks that are easier to automate end up reducing in their marginal value, precisely because they are abundant. Here, Baumol effects are implicitly captured by the inter-task complementarity. As effective labour increases for automated tasks, the marginal value of labour for non-automated tasks increases correspondingly. We do not explicitly model wages, but the economic value of tasks would end up setting their wages (in principle), so these are essentially capturing the same Baumol effects (Acemoglu et al. 2024). There is also empirical evidence on the overall size of Baumol effects, which we discuss below, after covering Engels effects.

[36] In a benchmark of AI research engineering tasks, AI agents could outperform human baselines up to a couple of hours, at prices 10x cheaper than corresponding human wages (Wijk et al. 2025).

[37] Investments in IT reached hundreds of billions of dollars in 1999. Although scholars disagree on when the web first showed productivity improvements on this scale, it appears that ecommerce sales reached similar levels around 2010 (Winters et al. 2011). Notably, this happened despite the infamous Dotcom bubble bursting in late 2000.

Meanwhile, Engels effects limit gains from increased productivity as rising incomes lead to more demand for discretionary goods and services. To the extent that these see lower productivity gains, Engels effects exacerbate Baumol effects. We do not model Engels effects here. Previous empirical estimates suggest that Baumol and Engels effects reduced US GDP growth by about 25% between 1948 and 2014 (Baqaee and Farhi 2019), which would not substantively change conclusions from this model.

As will be discussed in How capabilities are deployed, we believe there would be enough inference compute for a widespread deployment of AI. On current trends, there would be enough AI compute for all existing remote-compatible work to be assigned a H100-equivalent. Of course, there would be immense engineering challenges in provisioning and serving this compute across many requests; however, this extrapolation suggests there would be enough physical compute available.

Capabilities in scientific R&D

We turn now to examine the specific capabilities that AI is likely to achieve, and how they will affect scientific R&D. A recent framework highlights five key opportunities for AI in science: knowledge, data, experiments, models, and solutions (Griffin et al. 2024). Here, we review several different scientific R&D areas, each presenting a different profile in terms of these opportunities (*italicised* throughout).

In software engineering, on current trends, AI coding assistants and agents will likely lead to an abundance of software for well-scoped problems. This could clearly contribute to software development for science. AI for software engineering would act as a general-purpose productivity boost, relevant for practically all of these opportunities. Areas such as *data* analysis and software-based *experiments* and *models* would clearly see benefits from a large increase in available software engineering.

In mathematics, it seems likely that AI assistants will follow the path of software, becoming increasingly useful and independent over time. AI may transform how mathematicians generate and share *knowledge*, if it can reduce barriers to formalisation. AI may also help develop intuitions towards full proofs, as a *solution* tool – directly solving subproblems at a meaningful scale.

In molecular biology, two different visions exist for what will drive AI acceleration. Targeted AI tools such as AlphaFold will continue to improve, leading to unprecedented *data* and *models* concerning key biological processes. Meanwhile, general-purpose AI assistants might revolutionise *knowledge* sharing and accelerate *experiments* through feedback. Both pathways will be pursued in parallel, and basic scientific research in fields with plentiful data should flourish. However, translation to wider societal benefits is likely to happen on a slower timeframe.

In weather prediction, AI can enhance *models* of weather systems and will lead to continuing improvements in forecasts for everyday weather and extreme weather events. Integrating *data* from a vast array of different modalities raises the prospect of further improvements. Existing societal decision-making should benefit from improved forecasts in areas such as agriculture, emergency planning, transport, and power and water infrastructure. With significant improvement, weather forecasts might be used in other areas where currently they are neglected, although this is harder to predict.

How capabilities are deployed

Across all these fields, there are two recurring topics:

1. How does benchmark progress relate to progress in real-world capabilities?

2. When does deployment of those real-world capabilities happen, and what effects do they have?

Benchmark progress is astounding. AI rapidly improves at practically every task we have defined in detail, including challenges that domain experts find difficult. There are significant caveats for interpreting such results, but we argue that even imperfect benchmarks act as an informative signal about real progress in AI capabilities. Tasks created for benchmarking tend to be artificial in some way – they need to be easily verifiable, they are created by researchers aiming to probe current AI's limitations, etc. Nevertheless, benchmark progress clearly reflects some underlying real progress. And future benchmarks are informed by the outstanding gaps discovered by AI solving the benchmarks. Hence, a benchmark may be solved before the underlying capabilities are perfected, but significant progress will be made.

Then, there are key issues in deployment. Particularly common to discuss are *reliability*, *integration* into a *workflow*, and *cost*. Another issue, cutting across both development and deployment, is *specialist data*. We discuss

each of these in turn, before using them to examine potential impact for AI in scientific R&D within each domain.

Reliability. When systems are unreliable, it becomes difficult to deploy them at scale, and difficult to deploy them autonomously. AI systems can be notoriously unreliable, despite showing impressive capabilities in benchmarks and demonstrations. For example, LLMs often show degraded performance even on small perturbations to benchmark examples (Mirzadeh et al. 2024). This is more of a problem in some applications than others: for example, if mathematical results can easily be formalised and checked, reliability issues are relatively minor. Meanwhile, suggest that reliability is also improving over time (Kwa et al. 2025; Vendrow et al. 2025). This suggests that deployment will happen first in the areas where reliability is less crucial, but that it is unlikely to be a long-term obstacle.

Integration into a workflow. Using AI systems in real-world work often requires complex changes across many mutually interacting tasks. This can significantly hamper productivity improvements. For the most part, this is quite specific to individual workflows. We discuss deployment prospects under each heading. Generally, deployment is easier in areas with less serious consequences from mistakes, for example mathematics versus biology research. Deployment is easier where there is less need for slow empirical feedback loops, for example literature research versus wet lab experiments. And deployment is easier where there are fewer data availability issues. A key question that repeatedly arises is the nature of the AI systems being used: whether they automate tasks fully or partially, and the time horizon of the tasks that they automate.

Cost of deployment. There are two costs involved: costs from changing workflows, and inference compute. Changing workflows can change costs as tasks are rearranged; for example, a biochemist might be able to use biomolecule structure prediction to reduce their time and budget spent on lab experiments, but this might also require them to spend more time figuring out which experiments remain necessary. These changes can be hard to predict, although we discuss their prospects in each section. This is closely related to the previous point about reliability.

Meanwhile, for inference compute costs, there is significant cause for optimism. Evidence to date suggests that AI inference costs for a given level of capability rapidly drop over time, 10x per year or faster (Cottier et al. 2025) If this persists, then even when existing state of the art benchmark results use an expensive level of inference compute, they will rapidly become cheaper. Relatedly, when AI is capable of performing a task, so far it usually costs less than human workers (Wijk et al. 2025). This suggests that inference costs would only be a long-term bottleneck if (i) inference unit costs plateau, or (ii) AI automation requires performing more instances of tasks than at present.

A useful way to consider the required capacity is to examine trends in total installed AI compute, and compare with the required amount of inference compute. Total installed AI compute capacity, on current trends, would be 600 million H100-equivalents by 2030 (Hardware). Where will this compute be used? It seems likely that at least half of it would be allocated for inference. What would the inference be used for? Automating tasks, ranging from image generation to coding to myriad other applications. The global workforce is about three billion, working about a third of their time, with about a quarter of their tasks being remote-compatible. If each remote worker needs an H100-equivalent for their AI usage in 2030, this requires about 250 million H100-equivalents – that is, roughly half of the projected compute available. In practice, inference compute will be allocated across different tasks according to both their value and their susceptibility to automation. Nevertheless, this rough calculation suggests that there should be enough compute for AI capabilities to be deployed at scale.

Specialist data. Issues around data availability can affect both deployment and development. Data collection is particularly challenging when it is expensive or logistically difficult. Examples include real-time deployment requiring sensor installations (weather prediction), or specialist wet lab data collection (biomolecule interaction). We consider the details of this separately within each area. Generally, domains such as software and mathematics suffer less from this problem, due to the possibility of easily generating data without physical experiments.

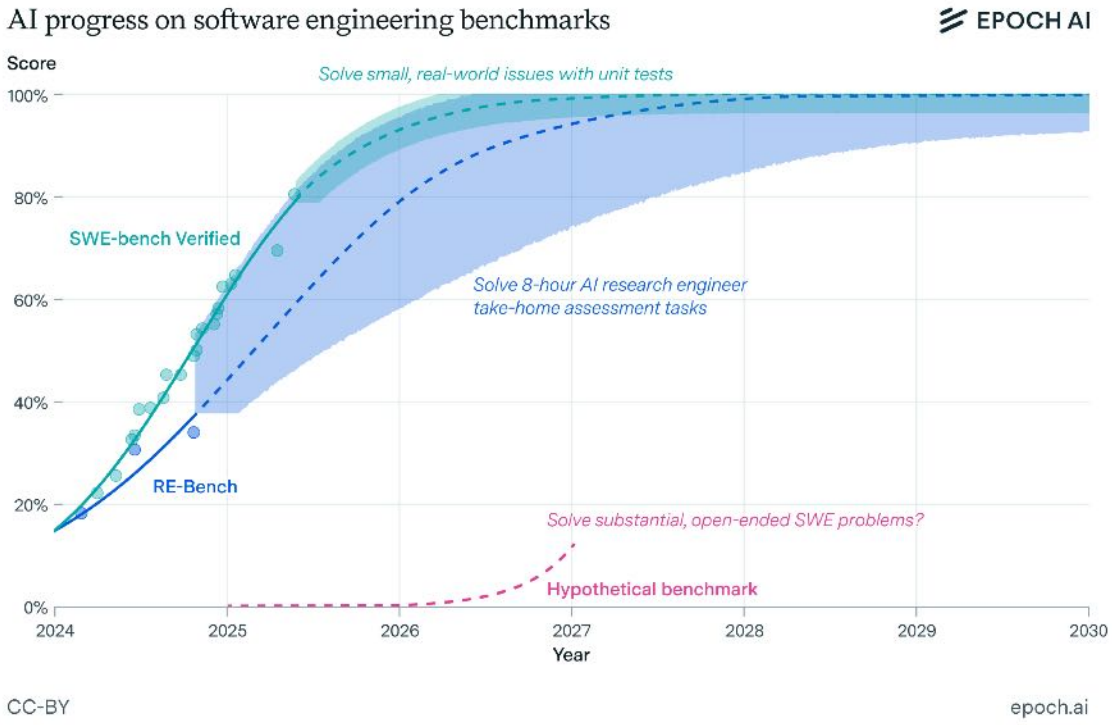
We provide a high-level summary of these challenges by domain in the table below, before discussing each domain in more detail in dedicated sections. Across all the R&D areas considered in this report, we see that AI opportunities are plentiful, and current trends point towards tremendous impacts. This is particularly compelling in areas such as software engineering and protein structure prediction, where existing real-world usage confirms the initial promise of benchmarks.

SUMMARY OF CHALLENGES TO DEPLOYMENT

	Reliability	Integration into a workflow	Cost of deployment	Specialist data
Software engineering	✓✓ Quickly checkable compared to time savings, errors usually not highly expensive.	✓✓ Already being integrated successfully, fast feedback loops, task time horizons showing steady improvement.	✓✓ Already being deployed cost-effectively, likely to become cheaper.	✓✓ Plentiful data exists, and more can be generated fairly easily.
Mathematics	✓ Likely to be checkable compared to time savings, errors not highly expensive.	? May require shift to formalisation, little real-world success so far.	✓✓ Assuming similar to software tools or existing math tools, should be cost-effective.	✓ Potential bottlenecks, synthetic generation might ameliorate these.
Molecular biology: molecule prediction	✗✗ Error rates hard to characterise, not easily checkable, costly.	✓✓ Several real-world success stories, unclear what shape this will ultimately take.	✓✓ Inference costs cheap based on current systems, unclear costs for workflow adaptation.	✗✗ Specialist data is likely a bottleneck, requires experiments to collect.
Molecular biology: AI desk research assistance	✗ Not easily checkable.	✓ Some real-world success stories, ultimate extent unclear.	✓✓ Already being deployed cost-effectively in a limited form, likely to be cheap.	✓ Fairly plentiful literature data, no experiments needed, risks around non-paper data.
Weather prediction	? Not easily checkable, although preexisting numerical methods can sanity check.	✓✓ Successfully deployed at scale.	✓ Cheap to run, potentially large costs from workflow rearrangement.	✓ Specialist data is partly a bottleneck, e.g. extreme events. Existing data supports valuable uses.

Qualitative ratings of different challenges, where ticks indicate a challenge is addressable or non-blocking, and crosses that it could prevent adoption. Double icons indicate more confident conclusions, for example present day adoption or stronger arguments.

Software engineering



SWE-bench Verified: a coding benchmark based on solving real-world GitHub issues with associated unit tests. Results include those reported from model cards, including those with private methodology such as Claude Sonnet 4. The trend would be similar if limited to the public scoreboard.

RE-Bench: a research engineering benchmark based on tasks similar to take-home assessments for job candidates, taking approximately eight hours for humans.[38]

AI is already transforming software engineering through code assistants and question-answering. By 2030, on current trends, AI will be able to autonomously fix issues, implement features, and solve difficult (but well-defined) scientific programming problems.

Software engineering is a particular area of interest for frontier AI developers, with both chat interfaces and tools like Copilot extensively

[38] For RE-Bench, the maximum achievable score is uncertain. For these fits, we set 100% as a normalised score of 1.5, i.e. the low end of the estimated maximum average score. As the benchmark is not yet near saturation, this has little effect on the extrapolation.

adopted (Yepis 2024). Moreover, software engineering is a key part of scientific R&D across many domains. AI R&D is particularly coupled to software engineering, because much of AI research revolves around software engineering to devise new algorithms, develop new AI models, etc. However, software engineering is an important part of scientific work in other fields such as physics, chemistry, biology, etc.

What does existing progress suggest about AI for software engineering in 2030? We examine three sources of evidence: real-world usage of AI for software engineering today, benchmark progress, and current open problems and research as articulated by domain experts. Taken together, these suggest AI will dramatically change software engineering, and is already having significant effects. However, significant uncertainty remains about AI's capability to autonomously perform challenging tasks end-to-end in the real world. Benchmark results suggest rapid progress towards this level, but domain experts remain divided, particularly on reliability and workflow integration.

Task	Relevant progress
Autocomplete snippets of code	Simple coding benchmarks such as HumanEval are solved. Coding assistant products are deployed widely, generate billions of dollars in revenue, and improve developer productivity in trials.
Solve small real-world issues with unit tests	SWE-bench steadily improving.
Solve well-defined scientific R&D coding problems from natural language description	SciCode, RE-Bench and similar benchmarks are steadily improving. Prominent engineers report using recent AI models as a "language to code" assistant (Wikipedia 2025, "Vibe coding").
Solve real-world coding problems priced at hundreds to thousands of dollars on a freelancer marketplace	SWE-Lancer benchmark performing above chance, and showing steady progress.

Complete professional-level "capture-the-flag" cybersecurity challenges	Cybench shows steady progress. An LLM-based system recently found a new vulnerability "in the wild" in the popular sqlite package (Big Sleep Team 2024).
Replicate results in code from a research paper	PaperBench (only the paper provided) shows early signs of progress. CORE-Bench (repository also provided) shows steady progress.
Solve substantial open-ended software engineering problems, e.g. developing a new database given high-level requirements	Arguably no benchmark thoroughly covers this.

AI systems today can already provide implementations from a natural language specification, make suggestions during code editing, and autonomously investigate and resolve bugs (Cui et al. 2025; Jimenez et al. 2024). However, as of today, these AI capabilities are not reliable, and typically apply to problems at the easier end of engineers' work (Miserendino et al. 2025). Consider SWE-bench Verified as an example. These problems are taken from real GitHub issues, but only those with a unit test to provide unambiguous resolution of whether the AI's attempt succeeded. As a result, almost all of these problems touch one or two files, and are primarily resolving small issues. The benchmark-leading score of today is around 70%. This is much better than random chance, but far from reliable. Hence, AI today is mostly used as an assistant, with close supervision. Most field studies have found productivity improvements of 20-70%, varying significantly by developer and area, although one rigorous field study found a surprising 20% slowdown (Cui et al. 2025; Becker et al. 2025).[39]

However, many popular benchmarks frame the problem in terms of autonomous engineering agents, performing significant software tasks

[39] In a recent study of AI's effects on software engineering, literature review identified seven empirical studies. 6/7 found 20-70% speed-ups or increases in output (Becker et al. 2025). The remaining study found a surprising 20% slowdown, although it has a claim to the most thorough methodology. We take 20% as the starting point, then, but we acknowledge there is considerable uncertainty in current evidence.

end-to-end (Jimenez et al. 2024; Miserendino et al. 2025; Wijk et al. 2025). At its extreme, this could change the nature of software engineering, with human engineers overseeing coding agents (Yang et al. 2024). Solving the above benchmarks would not clearly entail reaching this extreme outcome: in both cases, compared to real-world problems, the benchmarks are more crisply-defined, shorter, and conceptually simpler. Nevertheless, solving these would be a clear sign of progress.

Matching this interpretation of overseeing a team of virtual engineers, several AI researchers have predicted that AI will be able to autonomously perform substantial implementation tasks from their work in the next five years, before being able to compete on higher-level planning and creating research ideas (Owen 2024b). Recent evidence suggests that there has been a steady 3.3x per year increase in the time horizon of autonomous software benchmark tasks that AI can perform for a given reliability level (Kwa et al. 2025). In AI research, with computationally-expensive experiments and training, AI agents would need to be extremely reliable if they were themselves allocating significant compute resources, for example for ML experiments. However, for lower stakes tasks, such as implementation and debugging of a webpage, the picture is more optimistic.

What challenges could stand in the way of automation-led software abundance? Some of the most commonly raised concerns are inference costs, reliability (with the resulting need for human supervision), and potential AI deficiencies in open-ended problem solving (Owen 2024b).

So far, inference costs are relatively affordable for software agents: in the more challenging benchmark problems that AI has successfully solved, inference costs are much lower than the corresponding human wage for that problem (Wijk et al. 2025). However, there is the important caveat that solving more difficult problems may require further scaling of inference. Balanced against this, inference costs for leading models have gotten dramatically cheaper, at a rate of 10x per year or more (Cottier et al. 2025). Even if there is inference scaling up comparable to scaling of training compute (4-5x per year), on current trends the costs would *decrease*. This tentatively suggests that inference costs are unlikely to be a bottleneck in the medium term.

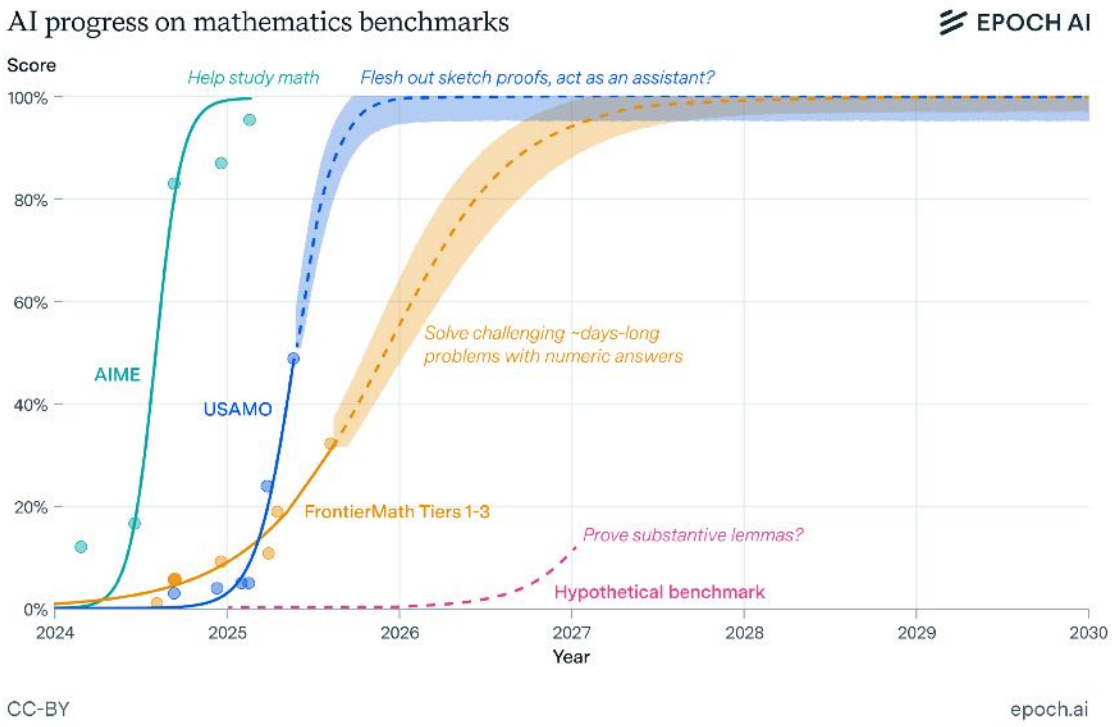
Another key obstacle for "overseeing a team of virtual engineers" arises from reliability. If there is any need for human engineers to intervene and dive deep into the code, this would act as an important bottleneck. A natural comparison is to overseeing junior engineers today - with the associated need for a more experienced engineer to provide occasional detailed input.[40] If AI reliability remains meaningfully below that of humans, this could curb the usefulness of software agents, leaving them closer to extended coding assistants.

Meanwhile, the question of whether AI will be able to autonomously solve more open-ended problems remains difficult to predict. Recent benchmarks such as SWE-Lancer suggest that real world freelancer tasks are already within reach of software agents, despite being more open-ended than SWE-bench tasks.[41] AI agents may soon be able to solve tasks comparable to hours-long freelancer software jobs, with more difficult tasks remaining accessible only to an AI-human combination. On current trends this seems likely to continue improving, leading to a world in which, at a minimum, software engineering agents for prototyping or analysis are cheap and ubiquitous.

[40] If AI agents continue to struggle with reliability, then they might be particularly difficult to deploy for software engineering in R&D, where results often are less easily checkable than, for example, web development (Owen 2024b). For this reason among others, domain experts are divided on how soon they would predict such a transformative shift to hit software engineering for ML experiments, compared to other areas.

[41] SWE-Lancer tasks use end-to-end tests rather than unit tests as in SWE-bench. This should allow for more open-ended tasks, as the tests do not hint to the agent at how to implement code.

Mathematics



Results show general-purpose LLMs only, excluding domain-specific systems like AlphaProof and AlphaGeometry2 (mid-2024).

AIME: a high school mathematics exam used for determining entry to the US Mathematical Olympiad, integer answers.

USAMO: US Mathematical Olympiad, a high school mathematics exam with proof-based answers.

FrontierMath: a mathematics benchmark focused on challenging questions up to expert level, but still offering straightforwardly-verifiable answers (numeric or simple expressions).

AI for mathematics may soon be able to act as a research assistant, trying to flesh out proof sketches or intuitions. Early accounts from mathematicians already document AI being helpful in their work. However, it may be necessary to make significant changes to mathematicians' workflows for AI tools to be widely used. Notable mathematicians differ

greatly in how relevant they think existing mathematical AI benchmarks are for their work, as well as in their predictions for how soon AI will be able to develop mathematical results autonomously, rather than as an assistant.

Benchmarks for mathematics are further from professional mathematicians' work tasks than software engineering benchmarks. Many common mathematics benchmarks focus on exams - school exams, for example, or more challenging invitational competitions including various Mathematical Olympiads. These may be informative about AI progress, but have a less natural interpretation in terms of useful capabilities once the benchmarks are solved. A notable exception is FrontierMath, which attempts to formulate mathematical questions that are similar to those faced by early-career research mathematicians, albeit while remaining easily verifiable.

Several notable mathematicians emphasised the significant difficulty of FrontierMath's hardest problems.[42] Subsequent rapid progress on this benchmark raises the question of whether the problems are as difficult as they seemed. One potential issue is that, in order to make the solutions verifiable, many problems use numerical answers. Numerical problems may be susceptible to brute force, despite the designers' intention to avoid this. Hence there is a risk that the benchmark overestimates progress in challenging mathematical reasoning. Nevertheless, separate to the question of benchmark validity, several prominent mathematicians anticipate rapid progress in AI for mathematics, even predicting a (highly uncertain) ten-year timeline to full automation of mathematical research (Glazer et al. 2024).[43]

What would the implications be for AI in mathematics if ambitious mathematics benchmarks like FrontierMath were solved? Mathematicians have shared their ideas of what utility an AI capable of solving such problems would contribute to their work. They suggested such an AI might

[42] For example, Terence Tao stated of the hardest subset of problems, "These are extremely challenging. I think that in the near term basically the only way to solve them, short of having a real domain expert in the area, is by a combination of a semi-expert like a graduate student in a related field, maybe paired with some combination of a modern AI and lots of other algebra packages..."

[43] For example, Richard Borcherds stated, "When is AI going to overtake humans at research? Well, not in the next year, and almost certainly in the next 100 years. So I'll go for about ten years or so."

"verify calculations, test conjectures, and handle routine technical work while leaving broader research direction and insight generation to humans" (Glazer et al. 2024). Another area, where several mathematicians are interested, is using AI for formalisation and communication.[44]

Task	Relevant progress
Simple arithmetic	GSM8k and similar benchmarks are solved. However, even in fairly recent models, reliability is still an issue.
Solve typical school math exams	School-finishing exams like SAT and GRE are solved, although reliability can be poor.
Solve challenging high-school math competitions	Exams with numerical solutions, such as AIME have been largely solved. Similar exams requiring proofs, such as USAMO, show sharp progress from Gemini-2.5, suggesting they may soon follow. Moreover, specialised systems such as AlphaProof have performed better, but focusing specifically on formal settings where they can be trained via self-play. PutnamBench (formal proofs for Putnam exam questions) shows some signs of early progress, but not enough to be confident in a timeline.
Solve challenging expert-level ~days-long questions with numeric answers	FrontierMath has seen rapid progress, although interpretation is challenging in light of varying difficulty tiers and model types.
Formalise proofs from informal language	This area remains nascent, with most projects testing the formalization of natural language problem statements drawn from undergraduate education and math competitions, and few models systematically benchmarked (Azerbayev et al. 2023).
Prove substantive lemmas or theorems	No systematic benchmark exists yet. Specialist AI systems have helped mathematicians identify

[44] For example, Terence Tao stated, "If I were to write a math paper, I would explain the proof to a proof assistant... and they would help formalize it."

	promising conjectures or prove results, albeit requiring significant mathematician input.

Complicating our analysis, there are several narrower AI systems for mathematics, many of which have achieved some of the most impressive results to date. Formal systems such as AlphaProof achieved high scores on IMO questions before general-purpose systems did, but have not yet been documented as being useful in research. Other AI tools have been used to guide researchers towards promising conjectures or optimise problems under constraints, leading to novel results with significant mathematician input (Davies et al. 2021; Romera-Paredes et al. 2024; Novikov et al. 2025). This has significant overlap with earlier work on experimental mathematics, but can take advantage of deep learning methods to detect patterns that traditional machine learning would not detect. These results came from both narrow AI systems and task-specific systems built on LLMs, used in combination with extensive problem-specific setup work from domain experts.[45] It is plausible that narrower AI tools will become useful at scale before general-purpose systems, or even at the same time. This largely depends on broader uncertainties around near-term AI capabilities.

Unlike software engineering, there are no systematic studies examining productivity improvements for mathematicians from existing AI. However, there are notable claims of mathematicians using AI to assist in their work. In addition to the above results from narrower AI-enabled tools, mathematicians have shared their early impressions of working with general-purpose LLMs. Current accounts suggest they are far from reliably helpful, but sometimes meaningfully help in day-to-day research, for example successfully formulating a derivation (Burnham 2025).[46]

[45] For example, a subsequent review preprint described how “[d]eep mathematical knowledge of the problem was required here at practically every stage: In designing the DL architecture in step 1, in creating the data set in step 2, in choosing the experiments in step 4, in interpreting them in step 5. and of course in proving the theorem in step 6” (Davis 2021).
[46] For example, using o1 was like “trying to advise a mediocre, but not completely incompetent, (static simulation of a) graduate student [...] [i]t may only take one or two further iterations of improved capability (and integration with other tools, such as computer algebra packages and proof assistants) until the level of "(static simulation of a) competent graduate student" is reached, at which point I could see this tool being of significant use in research level tasks” (Tao 2024).

What could hinder real-world deployment of AI for mathematics? Several mathematicians have noted the importance of affordable deployment, lack of specialist data, and the importance of solving open-ended problems (Glazer et al. 2024).

For the reasons discussed in Software Engineering, there is reason to expect that even if inference compute scales up dramatically, cost reductions are likely to compensate for this. Meanwhile, a lack of specialist data could be an important bottleneck: many research domains rely on a small number of relevant papers, and depending on the data efficiency of AI systems, there simply may not be enough data for useful learning. A lack of data may also be related to the concern about open-ended problem solving: mathematicians publish proofs for their most compelling problems, rather than sharing extensive documentation of their reasoning process, errors, and progress (Glazer et al. 2024).

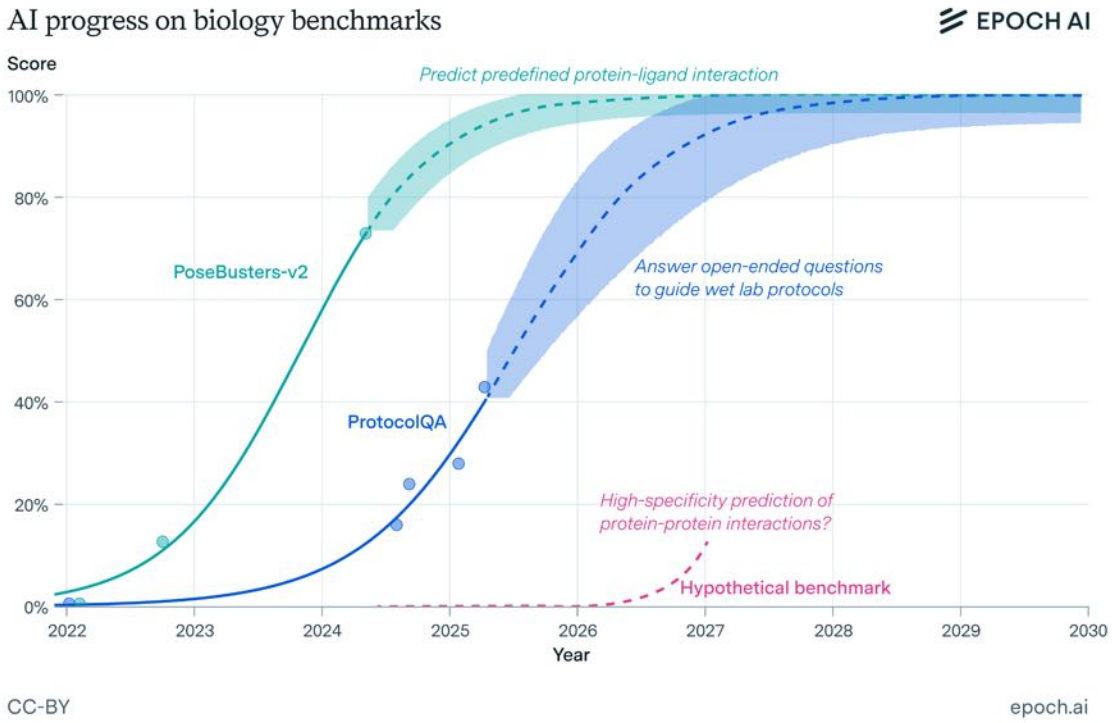
Finally, a flourishing of AI-assisted mathematical results could be bottlenecked by ways of working: to keep up with a large volume of AI-generated results, occasionally prone to hallucination or subtle error, formalisation would need to become more common, which might bring its own challenges (Yang et al. 2024).

Nevertheless, the overall picture from AI progress and expert opinion is fairly optimistic: AI is likely to contribute to mathematics research, at the very least becoming a helpful assistant similar to those used in software engineering today. We expect the pace of research discoveries to increase commensurately with how powerful these assistants become. By default it will take years for new mathematical results to become relevant for applied research, suggesting few visible impacts on broader society to begin with (Frontier Economics 2014). Existing research on the economic benefits of mathematics R&D tends to focus on applied R&D, counting the GDP contribution of occupations involved in computer science, data analysis, etc (Deloitte 2013). In the long run, increased output from basic mathematics R&D should have substantial spillover benefits for applied uses.[47] Gradually,

[47] Social returns to scientific R&D in general have been estimated at about twice those of private benefits, but with relatively scarce data (Frontier Economics 2014). Estimates of the rate of return on marginal R&D funding are high, centered around 50% per year (ibid).

an AI-enabled scale-up of mathematics research will transform the more applied sciences, with results likely to diffuse across cryptography, statistics, computing, physics, and beyond.

Molecular biology



PoseBusters-v2: a benchmark for protein-ligand docking (spatial interaction). We only include blind results, where the protein's binding pocket is not provided.

ProtocolQA: a benchmark for questions about biology wet lab protocols, here evaluated without multiple choice answers.

Protein-protein interactions: there is significant progress predicting protein-protein interactions, but predictions for arbitrary pairs have a high false positive rate. Our illustration of progress is highly uncertain, and would depend on benchmark details.

Scientists envision two different paths for AI in molecular biology: transformative AI tools for tasks like biomolecule modelling, and general-purpose AI assistants to automate parts of the research process. Tools such as AlphaFold are already revolutionising the field, and will expand to predict more properties for more complex structures. Meanwhile, AI assistants for biology research are at an early stage, but offer the promise of accelerating many key R&D steps.

There are many different applications in biology where AI is already showing great promise. AI is already being investigated across many areas: prediction of biomolecular structures and interactions, analysis and editing of genomic data, imaging, lab robotics, and more. Due to the breadth of the field, we focus on two key areas that represent the divide between specialised tools and general-purpose agents: AI for biomolecule prediction and design (particularly proteins), and biology desk research. Other research areas could be vitally important, but are either less straightforwardly within the purview of AI (such as robotics and imaging for wet lab research), or less straightforward to analyse.

AI for prediction and modelling of biomolecules has seen staggering success. AlphaFold2's principal authors recently shared the 2024 Nobel Prize for Chemistry.[48] AI approaches such as AlphaFold have revolutionised protein structure prediction, achieving near-experimental accuracy for many well-characterised protein domains in their equilibrium state.[49] Subsequent work has attempted to carry these successes over to other problems, such as other biomolecules like DNA/RNA, dynamic structure, molecule interaction, and even target identification and protein design.[50]

Task	Relevant progress
Predict protein equilibrium structure	Solved for basic structures (AlphaFold), though remaining challenges around hallucination, intrinsically disordered proteins, etc
Predict structures for other biomolecules / complexes	Progressing on relevant benchmarks such as held-out sets from PDB, CASP competitions, etc.

[48] David Baker shared the same year's prize for advances in protein design. Many of his lab's tools did not use AI approaches originally, although for example RoseTTAFold does.
[49] Within the prediction of equilibrium structure for single proteins, challenges remain for intrinsically disordered regions, protein complexes, and certain novel folds.
[50] There is a divide between relatively specialised models such as AlphaFold and RoseTTAFold, which primarily use spatial data, and more general protein language models (PLMs) such as Evo and ProGen, which are trained on a large corpus of biological text data. We separate our discussion into narrow tools versus general-purpose AI, but the distinction is fuzzy. Potentially, the best general-purpose AI assistants for biology will incorporate usage of tools such as AlphaFold, or be additionally trained on domain-specific biomolecules similarly to Evo.

Design custom binders without high-throughput screening	Achieved in demonstration examples (AlphaProteo), but not necessarily for arbitrary targets.
Predict single nucleotide variation effects	Steady progress on benchmarks such as ClinVar.
Predict arbitrary protein interaction and binding	Docking benchmarks like PoseBusters show progress, but ad hoc trials suggest current methods struggle for arbitrary real-world protein-protein interaction (Dickinson, 2024).
Predict small molecule interaction	Small molecule binding prediction performance struggles to rise above chance on an “in the wild” benchmark (Leash Bio, 2024).
Predict properties such as efficacy or toxicity	There is much active research interest in this topic, and individual results suggesting AI methods perform better than chance. These are usually considered within pre-specified domains, with significant expert setup, rather than “predict properties for an arbitrary biomolecule”.

Meanwhile, AI for desk research tasks has only seen advances more recently, with LLMs recently solving some of the first multiple-choice benchmarks on challenging literature search questions, reasoning about experimental protocols, and interpreting figures (Laurent et al. 2024). Currently, AI for biology research tasks mostly acts as an expressive but error-prone search engine for specialist knowledge. In biology, the most visibly impressive results from AI so far are from “tools” rather than agents, although recent exciting results exist where AI literature research tools have suggested targets for drug repurposing, novel treatments, and other applications (Gottweis et al. 2025; Lu et al. 2024; Huang et al. 2024).

Task	Relevant progress
Answer school-level questions about biology	Biology categories within MMLU and school exam benchmarks are solved. Robustness and reliability are an outstanding challenge, but appear to be steadily improving.
Answer challenging graduate-level exam-style multiple choice biology questions	Benchmarks such as GPQA are making steady progress.
Answer open-ended questions about biology wet lab protocols	Open-ended ProtocolQA and BioLP are making steady progress.
Answer well-defined questions about recent biology research literature	Benchmarks such as LitQA show steady progress.
Perform end-to-end bioinformatics analyses	BixBench does not yet show clear progress, but is recently released. Benchmarks with less challenging bioinformatic tasks, such as BioCoder and ScienceAgentBench, show steady progress.
Answer open-ended biology research questions, propose hypotheses and identify relevant experiments and related work	No end-to-end benchmark fully covers this. Demonstration results from "co-scientists" and similar systems suggest progress.

What might AI-assisted biology R&D look like in 2030? Biomolecule prediction benchmarks and real-world usage offer a hint. Again, we focus this discussion on the specific areas around biochemistry, drug targets, and ultimately drug development. Benchmark progress suggests that other biomolecule prediction tasks (RNA, DNA, protein complexes, small molecules, interaction, etc) will see similar prediction advances to proteins, as long as sufficient data can be found or generated. Structure prediction will improve steadily, enabling better prediction of other properties like receptor binding (Zambaldi et al. 2024).[51] Lab experiments will remain vitally

[51] How far can this prediction improve? To give a sense of current methods, recent results from RFDiffusion and AlphaProteo suggest that researchers can predict binding affinity well enough to design novel proteins to bind to receptors without high-throughput screening.

important, but investigating a given target should require fewer hours in the lab. Meanwhile, exploring literature, debugging experiments, and analysing results are likely to be assisted by AI – which seems likely to saturate existing benchmarks on desk research and protocol debugging. Expert disagree on which will contribute more, but expect both areas to advance.

This should significantly accelerate early R&D: coming up with a new drug target and investigating it should require less researcher time overall, while steering towards drugs with better characteristics (higher binding affinity, lower toxicity, fewer interactions with other targets, etc). However, downstream drug development is likely to see modest end-to-end productivity effects by 2030, particularly given the time requirements and regulatory processes for new biomedical treatments. A new drug typically takes eight years to go through trials and approval (Brown et al. 2021). It is likely that the drugs approved in 2030 are those already in the trial pipeline today, and hence any AI involvement in their early development already occurred in the last few years.[52]

This is not to downplay the longer-term impacts of AI. The long duration of pharmaceutical pipelines also underpins the opportunity for AI to accelerate drug development. Historically, for a given medicine, most time is spent in its early development, with one study identifying a median of twenty-eight years between initiation and the first clinical trials for a novel target (McNamee et al. 2017). Accelerating early research could correspondingly accelerate these timelines.

In the longer term, in silico predictions may lead to treatments that are substantially better than those currently being trialled. AI-designed treatments could be more efficacious, have fewer side effects, and see higher trial success rates, which in turn could improve the economics of drug development. Currently, about half of pharmaceutical R&D spending is

However, an "in the wild" experiment suggests that interaction cannot yet be predicted in a single round for arbitrary targets (Dickinson, 2024). Given progress in existing protein docking benchmarks, it seems plausible that by 2030 interactions should be fairly predictable for randomly-chosen targets, similar to protein structure today.

[52] Arguably, there are already AI-enabled therapeutic drugs, if AI is defined more broadly to include pre-LLM methods. There are even studies suggesting such drugs may have higher approval rates than the industry average. However, such results do not seem highly relevant to the focus of this discussion (KP Jayatunga et al. 2024; Lowe 2024a).

on clinical trials focused on these properties, with failure rates of about 50% per phase across three phases (Sun et al. 2022). Even ignoring other benefits, reducing the frequency of expensive late stage failures could be transformative. There is also the possibility that entirely new biomedical processes may be facilitated by AI design and organisation, conceptually similar to processes such as mRNA vaccines, which can be safely renewed year-to-year without having to undergo approval from scratch (Brown et al. 2021). Hundreds of billions of dollars are spent on pharmaceutical R&D each year, and trillions on medicines, meaning downstream impacts would be highly valuable.

Researchers have frequently discussed two other important potential bottlenecks, such as the need for specialist biology data, or the ongoing importance of wet lab experiments (Lowe 2024b). Specialist data is crucially important, and it is an open question whether other problems will be as amenable to data collection as protein structure prediction.[53] One promising sign is that several initiatives are already collecting biological data in massive quantities for biology AI development.[54] Given the significant incentive to continue improving biology AI, data collection seems likely to expand further.[55] Conversely, wet lab experiments are almost certain to continue as an important part of day-to-day work, and here the uncertainty is how significant that bottleneck will be. This will largely depend on the extent to which improved AI methods can reduce the volume of experiments required. In several real world examples, protein structure prediction has substantially reduced experimental durations.[56] Despite

[53] Indeed, one pessimistic description comes from the National Academies of the Sciences' report on AI in the life sciences: "However, with the notable exception of nucleic acid sequences and structural data, data in the life sciences are fragmented, and robust, reliable, and well-curated data that are amenable to model training are scarce. Both top-down generation of high-quality datasets and bottom-up aggregation of diverse and smaller datasets alongside tools for data harmonization are valid approaches to address the paucity of biological data."

[54] For example, Leash Bio have even used some of their datasets for new AI challenges on small molecule - protein interaction.

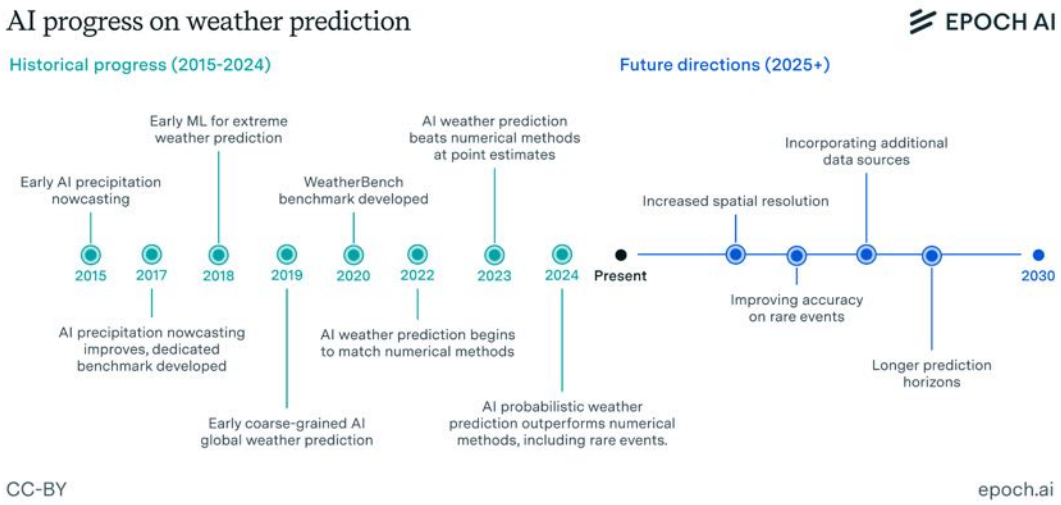
[55] As well as specialist data for prediction models, there is uncertainty whether data is also a bottleneck for literature research agents, for example due to publication bias. A counterargument is that the corpus of publications is also what is available to human researchers, who nevertheless learn to conduct literature research.

[56] For example, from malaria researcher Matthew Higgins, "We'd been battling with this problem for years, trying to get the details we needed. Then we added AlphaFold into the

concerns around data and the need for experiments, AI should meaningfully accelerate biology R&D.

mix. And when we combined our model with AlphaFold's predicted structure, we could suddenly see how the whole system worked." Quotes like these cannot determine the overall time saved on average across many projects, but do provide evidence that computational methods sometimes save significant amounts of time in practice.

Weather prediction



Timeline of key milestones in weather prediction AI, as well as anticipated developments for the future.

AI weather prediction can already improve on traditional methods across weather prediction tasks from hours up to weeks. Moreover, AI methods are cost-effective to run, and could improve further with the careful collection of more data. The next big challenges lie in improving existing predictions, predicting rarer events, potentially predicting further ahead, and making use of improved predictions to achieve benefits in the wider world.

Existing benchmarks for AI in weather prediction mostly show AI is on par with, or better than, state of the art ensembles of numerical models for horizons of hours to tens of days (Rasp et al. 2024). For prediction of key variables such as temperature, pressure, wind, and precipitation, AI methods can outperform by 10-30% (Price et al. 2024).[57] Such systems are trained on datasets of historic weather, but these in turn depend on

[57] This is true of individual forecasts for specific variables at specific locations and times, and more broadly true of the relative economic value of ensemble forecasts. (Relative economic value is not an all-things-considered estimate of the dollar value of improved forecasting, but a predefined metric of forecast value for a range of users across different decisions and tradeoffs.)

numerical weather models, so in an important sense AI methods augment numerical models rather than fully replacing them.[58]

It is unclear what further accuracy improvements can be achieved for short- and medium-term point forecasting beyond the numerical baseline, although researchers suggest they will continue, in addition to the possibility of finetuning on other data (Price et al. 2024). Perhaps more importantly, researchers expect further improvement in other key areas, such as better calibration of probabilistic models, particularly for rare weather events such as hurricanes.[59,60]

How will AI affect weather prediction R&D by 2030? We focus on narrow AI for weather prediction, although presumably there would be advances in general-purpose research agents somewhat like those discussed in previous sections. Numerical weather prediction research will doubtless continue, both as a tool for direct prediction and as a basis for AI augmentation. Potentially, empirical findings learned by AI systems could lead researchers towards important new effects for modelling. However, it also seems likely that the field will see a flourishing in empirical research: data collection, integrating new data sources into models, and validating their performance.[61] After collection, this research is likely to be relatively democratised: AI weather prediction models are generally cheap to experiment with, compared to numerical methods.[62]

---

[58] Numerical weather prediction models are used for reanalysis, converting observations of weather conditions back to best estimates of weather across the entire series of the surface and atmosphere gridded at regular intervals.

[59] A weather prediction researcher interviewed as background for this report emphasised that predicting extreme events is difficult, but a key area of interest and stated, "I've learned not to bet against ML."

[60] There is even early work arguing it may be possible to extend weather prediction horizons, where numerical methods were historically believed to be limited to around two weeks for point predictions (Shen et al. 2024; Chen et al. 2024). Note that this is quite distinct from the much longer term problem of climate prediction. AI for climate prediction is an area of active research interest, but remains more of an open question compared to the clear progress in weather prediction.

[61] A weather prediction researcher interviewed as background for this report described how more research effort into collecting and using better data could lead to a fundamentally "observation-driven approach" to weather forecasting.

[62] For example, GenCast training used 32 TPUv5 chips for five days. Generating a single forecast took eight minutes on a single chip, and was parallelisable. In comparison, leading numerical methods require entire supercomputers to be used for hours (Price et al. 2024).

Most of the obvious challenges to improving weather prediction are around data. Existing data is not always readily available, and collection latency may be suitable for R&D but not for real-time deployment.[63] Many proposed data sources are not collected systematically yet, or not publicly available (Bouallègue et al. 2024). Systematic data collection would face predictable bottlenecks: funding, institutional coordination, and in some cases even permissions to install data recording equipment.

Assuming these challenges are met, AI has the possibility of achieving significant real world impact through weather prediction. Already, research is exploring how prediction of extreme weather such as storms, floods and droughts may improve societal response (Camps-Valls et al. 2025; Cohen 2024). Additionally, day-to-day prediction of phenomena such as cloud cover, humidity, and rainfall can affect critical decisions in power infrastructure, agriculture, transport, and other areas (Talbott 2022; Google Research, n.d.-b). These applications have significant economic value: for example, improvements in hurricane prediction have been estimated as saving 70 billion dollars in the US between 2007 and 2020 (Molina and Rudik 2024). Existing weather forecasts for the general public and businesses in the UK have each been valued around tens of billions of dollars (Herr et al. 2024), suggesting that value globally could reach hundreds of billions.

By 2030, the capability will exist in theory to enrich weather prediction systems with more accurate, better calibrated, more frequently updated weather predictions. The challenge of figuring out how to make use of such predictions is ongoing, but even pessimistically, existing decisionmaking processes stand to benefit.

[63] For example, imagine that highly localised atmospheric recordings prove beneficial for short-term weather prediction. In some cases, currently such recordings might be recorded on field-deployed hardware, with readings uploaded every few weeks as part of a research project.

Discussion & conclusion

We have examined the trends that drive AI development, and how these are likely to unfold by 2030. We argue that continued scaling of training and inference compute makes the path forward somewhat predictable, provided it keeps improving downstream capabilities. For the most part, current trends are likely to persist until 2030. The largest AI models will cost hundreds of billions of dollars, and use about 1,000x more compute than today's leading models. This is worthwhile if they can generate trillions of dollars of economic value by increasing productivity – which seems plausible, in light of AI capabilities advances.

We have also examined how AI could accelerate parts of scientific R&D by 2030. In particular, we have looked at the areas where there is relatively compelling evidence: tasks where there are relevant benchmarks, and we can show that AI is on track to scale up to a high level of performance. These predictions have caveats, but they offer clear evidence about tasks that future AI will be able to perform. AI will help scientific R&D in two ways: specialised tools for specific high value tasks like biomolecule prediction, and general-purpose agents for research tasks like literature review. Evidence so far is strongest for the former, where existing AI tools are already helpful across several R&D areas. Meanwhile, general-purpose agents are seeing active development, and already exist in an early form, but with less evidence on how helpful they are so far.

We have not examined the risks that come with transformative technology. In the 2030 predicted in this work, there is obvious potential for misuse: many of the capabilities we discussed for scientific R&D have potential for dual use such as cyberattacks or creating biological weapons. The prospect of relatively autonomous agents, able to pursue goals in the wider world, complicates this picture even further. There are also broader societal risks in the rush to develop advanced AI, ranging from labour market disruption to the environment. The drive to establish sufficient electrical power and manufacture specialised AI hardware could lead to heightened political tensions and significant environmental impacts. We examined how,

depending on energy infrastructure, even a massive AI scale-up could lead to relatively small carbon emissions. However, as with the other risks, this requires societal choices about how to develop AI, how to control its use, and how to mitigate its hazards.

There are also important choices about how to *enable* AI development. For many of the key trends in development, decisions such as funding or regulation are crucial. One important example is power. If future AI training runs require gigawatts of power, approaching the demand of entire large cities, then regulation and investment for infrastructure will have important implications for where (and how easily) large-scale AI training can happen. Similarly, regulation could have large effects on where automation can happen in the economy, and could potentially lead to large differences in AI deployment between different jurisdictions. We take no stance on what form such regulation ought to take – but it will clearly be important, and may even shift the trajectory of AI development, for better or worse.

By 2030, AI is likely to be a key technology across the economy, present in every facet of people's interaction with computers and mobile devices. Less certain, but plausibly, AI agents might act as virtual coworkers for many, transforming their work through automation. If these predictions come to pass, then it is vitally important that key decisionmakers prioritise AI issues as they navigate the next five years and beyond.

Appendix: AI's potential to reduce GHG emissions

We conducted a shallow review of AI applications with the potential to reduce GHG emissions. For each of the five most carbon-intensive economic sectors (electricity and heat, transport, manufacturing and construction, agriculture, industry), we searched for the keywords "AI emissions [sector]". We then reviewed the results for consolidation into the below categories. This review is not intended to be exhaustive, but rather to provide a broad overview of different possible AI applications being discussed for these sectors.

Applications are divided into less speculative and more speculative categories. For the less speculative applications, we attempt to provide a back-of-the-envelope calculation for potential GHG reductions, relative to current emissions. This is intended to be more illustrative of applications' upper bound potential than a thorough projection. We include AI applications even when they can be fulfilled with older or lightweight AI models, noting that many of these applications do not rely on frontier AI systems, i.e. they might plausibly be achieved without substantial further scaling.

Less speculative, already being piloted or used

Example of downstream application	Potential to avert emissions
Energy management systems (e.g. for datacentres)	Industrial energy consumption is a complex control problem, where AI systems may improve on earlier efforts. Existing work resulted in 9-13% energy reduction in live testing (Luo et al. 2022). It is unclear whether further improvements are feasible, but deploying this across other datacentres could plausibly avert 10% of datacentre cooling emissions, which in turn make up 10-20% of total datacentre operational emissions, so could avert 0.01-0.02% of current global emissions, assuming full adoption.[64] More ambitiously, assuming space cooling makes up 2% of global emissions,[65] a similar effect would correspond to averting 0.2% of global emissions.
Condensation trail reduction	Condensation trails from aircraft are a significant contributor to global warming, estimated at 35% of aviation's total GHG emissions. Improved forecasting can allow airlines to adapt their routes to reduce contrails, and in a pilot study reduced emissions by half (Elkin and Sanekommu 2023). Aviation currently accounts for 2.5% of total world emissions (Ritchie 2024), so this could avert 0.4% of current global emissions under full adoption.
More efficient traffic routing	Drivers use map services to determine their routes. Efficient routing, using AI to process traffic data, can reduce emissions. It is uncertain to what extent this should be counted as a future development, given that this is already live within existing maps products. This feature averted 2.9 million tonnes CO2e in deployed countries across two years (Google Sustainability 2024), so below 0.01% of total emissions. Aggressively, if its usage could be increased 10x, this might reach 0.1%.

[64] Globally, datacentres are believed to account for about 1% of GHG emissions (Rozite et al. 2023).
[65] Space cooling was about 2% of global emissions in 2016 (International Energy Agency 2018).

	Relatedly, smart traffic control systems have reduced emissions while waiting at intersections by 10% in pilots (Google Research, n.d.-a). If we assume that anywhere between 5% to 50% of total transport emissions came from intersection waiting[66], and motor transport accounts for approximately 30% of total emissions, then this could avert 0.15-1.5% of current global emissions, under full adoption.
More efficient transport sharing and EV adoption	AI can potentially improve transport sharing and electric vehicle adoption by better solving allocation problems. Currently, something like 1% of annual car trips use the largest ride-sharing platform, Uber. A highly speculative estimate, but if motor transport accounts for 30% of total emissions, and better transport sharing could lead to 5% reduction of emissions, this might avert up to 1.5% of global emissions.[67]
Optimising the electrical grid to lower carbon intensity	Better prediction of demand and operation of complex control systems can improve coordination between green power generation and demand. A pilot study found that wind power's economic value could be boosted 20% through such methods (Elkin and Witherspoon 2019). Aggressively assuming that this leads to 20% more production of wind power than otherwise, and displaces an equal amount of non-renewables generation, this could avert up to 1.6% of global emissions.[68]
Industrial process optimisation	Some industrial processes such as oil and gas processing (Degot et al. 2021) (15% of global emissions), steel manufacturing (Degot et al. 2021) (7% of emissions), or cement-making (Carbon Re, n.d.; Ge et al. 2022) (7% of emissions) are particularly carbon-intensive. To the extent that AI can optimise these processes, it can have a large impact. It is unclear whether to count this as using AI, as existing case studies are more focused on analytics

[66] Broadly consistent with this back-of-the-envelope calculation, Wu et al. (2025) found a 6.7% reduction in road traffic emissions in simulations of China's largest 100 cities.
[67] This lines up well with a preliminary analysis by Stern and Romani (2025).
[68] Wind power is projected to become about 14% of total electricity generation by 2030 (IEA 2024). If it displaced an additional 2.8pp of non-renewable electricity, this would displace 2.8pp out of the projected 54% of non-renewable generation, which would by then comprise about 30% of global emissions, i.e. 1.6% of total emissions or 5.2% of power emissions. This is a factor of 3x smaller than an existing analysis of several potential applications of AI to the power sector (Stern and Romani 2025).

	platforms that might have been implementable with traditional methods. Case studies are claimed to reduce emissions by 2-5%, suggesting a global reduction of 0.6-1.5% under full adoption.
Optimising domestic heating systems	Similar to energy management systems for datacentres, domestic heating can be optimised to reduce emissions, and AI can optimise better than traditional approaches. Early evidence indicated a learning thermostat could reduce a home's heating demand by 5-8% (Park 2017). Current adoption of smart thermostats has been estimated at 13-17% (Parks 2024). Assuming that adoption increases to 40%, then heating emissions could be reduced by 1.2%. If building heating makes up 10% of global emissions (IEA 2022), then this could give up to a 0.15% reduction in total global emissions.

More speculative, currently suggested or early in research

Example of downstream application	**Potential to avert emissions**
AI-enabled breakthroughs for carbon capture	AI-assisted design of new materials can lead to improved materials for carbon capture and storage (Park et al. 2024). AI-assisted capture systems might also improve capture process efficiency (Fisher et al. 2024). In both cases, these are at an early state of research, without proof-of-concept at scale.
AI helping rollout of renewable energy through permitting and planning	Initiatives exist for using AI to help with planning submissions, and pilot studies suggested they reduce timelines, although it is unclear how much of this effect is from AI rather than a broader change in processes (CivCheck 2025).
AI helping monitor emissions (deforestation, regulatory non-compliance, etc)	Several initiatives exist to monitor processes such as deforestation or factory methane emissions (Food and Agriculture Organization of the United Nations 2025; Maguire 2024). The end result is highly dependent on regulation and enforcement, but AI can support monitoring.

Improving agriculture efficiency	Improving agricultural yields can correspondingly improve their carbon footprint, and tracking emissions throughout agriculture processes could prioritise workflow changes (Halper 2025).
Circular economy to reduce manufacturing emissions	To the extent that AI makes it easier to coordinate resale or donation of unwanted goods, it can reduce emissions from manufacture. It is unclear how large to treat this effect given the potential for induced demand.
More efficient sourcing of goods	AI can potentially make it more efficient to ensure that goods are sourced from sustainable suppliers. This is potentially more relevant for corporations deciding bulk purchase orders.
AI-enabled breakthroughs for nuclear power	AI is already used to advance basic scientific research for areas such as fusion research. It is highly uncertain what the end result would be, but if successful, this could lead to a breakthrough in clean energy (Degrave et al. 2022).
AI chemistry and biochemistry tools to develop new materials, fuels, feedstocks, processes, etc.	AI tools for discovery of materials and chemicals could potentially lead to greener alternatives (Merchant et al. 2023; Kuzhagaliyeva et al. 2022).

Appendix: benchmark extrapolation details

We select benchmarks with compelling relevance to an R&D field, e.g. SWE-bench for software engineering, and search for evaluation results from leaderboards and notable AI model releases.

We filter, keeping only scores from models that have plausibly been the best-scoring at the time of their publication. Sometimes there are edge cases, for example high scores from models that saw a long lag between publication and release, or with uncertain details for their elicitation and grading. We note these in relevant captions where they affect results.

We perform simple normalisation to 0-1 between random chance and perfect performance. Label noise could mean that a perfect score is unachievable in practice for many benchmarks, and for benchmarks such as RE-Bench it is unknown what the highest plausible score is. However, because most of our datapoints are below such a ceiling, we do not expect this to make a large difference in fitting.

We then fit a sigmoid versus time for the running best models. We do not include other details explicitly, for example training compute, data quality, etc. Although this approach is simple, it has a history of usage as a baseline for benchmark prediction, particularly once mid-range scores are achieved (Owen 2024a; Grattafiori et al. 2024).

We show indicative uncertainty bands via the 90% prediction interval, assuming normal residuals and standard errors of fitted parameters. All-things-considered uncertainty is higher, but this provides a sense of the range suggested by current trends.

Bibliography

Acemoglu, Daron, Fredric Kong, and Pascual Restrepo. 2024. *Tasks At Work: Comparative Advantage, Technology and Labor Demand*. Working Paper No. 32872. National Bureau of Economic Research. https://doi.org/10.3386/w32872.

Acemoglu, Daron, and Pascual Restrepo. 2018. *Artificial Intelligence, Automation and Work*. Working Paper No. 24196. National Bureau of Economic Research. https://doi.org/10.3386/w24196.

Altman, Sam. 2023. "Planning for AGI and Beyond." OpenAI, February 24. https://openai.com/index/planning-for-agi-and-beyond/.

Amodei, Dario. 2024. "Machines of Loving Grace." https://www.darioamodei.com/essay/machines-of-loving-grace.

Amodei, Dario. 2025. "On DeepSeek and Export Controls." https://www.darioamodei.com/post/on-deepseek-and-export-controls.

Amodei, Dario, and Danny Hernandez. 2018. "AI and Compute." OpenAI, May 16. https://openai.com/index/ai-and-compute/.

Armstrong, J. Scott, ed. 2001. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. International Series in Operations Research & Management Science. Springer US. https://doi.org/10.1007/978-0-306-47630-3.

Azerbayev, Zhangir, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. 2023. "ProofNet: Autoformalizing and Formally Proving Undergraduate-Level

Mathematics." arXiv:2302.12433. Preprint, arXiv. https://doi.org/10.48550/arXiv.2302.12433.

Baqaee, David Rezza, and Emmanuel Farhi. 2019. "The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten's Theorem." *Econometrica* 87 (4): 1155–203. https://doi.org/10.3982/ECTA15202.

Barnett, Matthew. 2025. *The Economic Consequences of Automating Remote Work*. Epoch AI. https://epoch.ai/gradient-updates/consequences-of-automating-remote-work.

Becker, Joel, Nate Rush, Elizabeth Barnes, and David Rein. 2025. "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity." arXiv:2507.09089. Preprint, arXiv. https://doi.org/10.48550/arXiv.2507.09089.

Big Sleep Team. 2024. "From Naptime to Big Sleep: Using Large Language Models to Catch Vulnerabilities In Real-World Code." *Project Zero*, November 1. https://googleprojectzero.blogspot.com/2024/10/from-naptime-to-big-sleep.html.

Bouallègue, Zied Ben, Mariana C. A. Clare, Linus Magnusson, et al. 2024. "The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-like Context." Bulletin of the American Meteorological Society. *Bulletin of the American Meteorological Society* 105 (6): E864–83. https://doi.org/10.1175/BAMS-D-23-0162.1.

Brown, Dean G., Heike J. Wobst, Abhijeet Kapoor, Leslie A. Kenna, and Noel Southall. 2021. "Clinical Development Times for Innovative

Drugs." *Nature Reviews Drug Discovery* 21 (11): 793–94. https://doi.org/10.1038/d41573-021-00190-9.

Burnham, Greg. 2025. "Mundane Utility." Substack newsletter. *Lemmata*, May 2. https://lemmata.substack.com/p/mundane-utility.

Camps-Valls, Gustau, Miguel-Ángel Fernández-Torres, Kai-Hendrik Cohrs, et al. 2025. "Artificial Intelligence for Modeling and Understanding Extreme Weather and Climate Events." *Nature Communications* 16 (1): 1919. https://doi.org/10.1038/s41467-025-56573-8.

Carbon Re. n.d. "Carbon Re: AI for Industrial Decarbonisation." https://carbonre.com/.

Chen, Lei, Xiaohui Zhong, Hao Li, et al. 2024. "A Machine Learning Model That Outperforms Conventional Global Subseasonal Forecast Models." *Nature Communications* 15 (1): 6425. https://doi.org/10.1038/s41467-024-50714-1.

Chui, Michael, Eric Hazan, Roger Roberts, et al. 2023. *Economic Potential of Generative AI: The Next Productivity Frontier*. McKinsey & Company. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#/.

CivCheck. 2025. "CivCheck: Reduce Building Permit Times with AI." https://www.civcheck.ai.

Cohen, Deborah. 2024. *An Improved Flood Forecasting AI Model, Trained and Evaluated Globally*. November 11. https://research.google/blog/a-flood-forecasting-ai-model-trained-and-evaluated-globally/.

Cottier, Ben, Robi Rahman, Loredana Fattorini, Nestor Maslej, Tamay Besiroglu, and David Owen. 2024. "The Rising Costs of Training Frontier AI Models." arXiv:2405.21015. Preprint, arXiv. https://doi.org/10.48550/arXiv.2405.21015.

Cottier, Ben, Ben Snodin, David Owen, and Tom Adamczewski. 2025. "LLM Inference Prices Have Fallen Rapidly but Unequally across Tasks." Epoch AI. https://epoch.ai/data-insights/llm-inference-price-trends.

Cui, Zheyuan (Kevin), Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz. 2025. "The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers." SSRN:4945566. Preprint, SSRN. https://doi.org/10.2139/ssrn.4945566.

Datta, Arnab, and Tim Fist. 2025. *Compute in America: A Policy Playbook*. IFP. https://ifp.org/special-compute-zones/.

Davies, Alex, Petar Veličković, Lars Buesing, et al. 2021. "Advancing Mathematics by Guiding Human Intuition with AI." *Nature* 600 (7887): 70–74. https://doi.org/10.1038/s41586-021-04086-x.

Davis, Ernest. 2021. "Deep Learning and Mathematical Intuition: A Review of (Davies et al. 2021)." arXiv:2112.04324. Preprint, arXiv. https://doi.org/10.48550/arXiv.2112.04324.

Degot, Charlotte, Sylvain Duranton, Michel Frédeau, and Rich Hutchinson. 2021. *Reduce Carbon and Costs with the Power of AI*. Boston Consulting Group. https://web-assets.bcg.com/28/f7/2ddf0628493d8cdd73dab4194e4a/reduce-carbon-and-costs-with-the-power-of-ai-new.pdf.

Degrave, Jonas, Federico Felici, Jonas Buchli, et al. 2022. "Magnetic Control of Tokamak Plasmas through Deep Reinforcement Learning." *Nature* 602 (7897): 414–19. https://doi.org/10.1038/s41586-021-04301-9.

Deloitte. 2013. *Measuring the Economic Benefits of Mathematical Science Research in the UK*. https://www.lms.ac.uk/sites/default/files/Report%20EconomicBenefits.pdf.

Denain, Jean-Stanislas. 2024. "Accuracy Increases with Estimated Training Compute." Epoch AI, November 27. https://epoch.ai/data-insights/compute-vs-accuracy.

Dickinson. 2024. "PPI Prediction Challenge 1." https://www.dickinsonlab.uchicago.edu/ppi-challenge.

Elkin, Carl, and Dinesh Sanekommu. 2023. "How AI Is Helping Airlines Mitigate the Climate Impact of Contrails." Google, August 8. https://blog.google/technology/ai/ai-airlines-contrails-climate-change/.

Elkin, Carl, and Sims Witherspoon. 2019. "Machine Learning Can Boost the Value of Wind Energy." Google Deepmind, February 26. https://deepmind.google/discover/blog/machine-learning-can-boost-the-value-of-wind-energy/.

Erdil, Ege. 2024. *Optimally Allocating Compute between Inference and Training*. Epoch AI. https://epoch.ai/blog/optimally-allocating-compute-between-inference-and-training.

Erdil, Ege, and Matthew Barnett. 2025. *Most AI Value Will Come from Broad Automation, Not from R&D*. Epoch AI.

https://epoch.ai/gradient-updates/most-ai-value-will-come-from-broad-automation-not-from-r-d.

Erdil, Ege, Tamay Besiroglu, and Anson Ho. 2024. "Estimating Idea Production: A Methodological Survey." SSRN:4814445. Preprint, SSRN, May 14. https://doi.org/10.2139/ssrn.4814445.

Erdil, Ege, Andrei Potlogea, Tamay Besiroglu, et al. 2025. "GATE: An Integrated Assessment Model for AI Automation." arXiv:2503.04941. Preprint, arXiv. https://doi.org/10.48550/arXiv.2503.04941.

Fisher, Oliver J., Lei Xing, Xingjian Tian, Xin Yee Tai, and Jin Xuan. 2024. "Responsive CO_2 Capture: Predictive Multi-Objective Optimisation for Managing Intermittent Flue Gas and Renewable Energy Supply." *Reaction Chemistry & Engineering* 9 (2): 235–50. https://doi.org/10.1039/D3RE00544E.

Food and Agriculture Organization of the United Nations. 2025. "REDD+ Reducing Emissions from Deforestation and Forest Degradation: Initiatives." https://www.fao.org/redd/initiatives/en/.

Frontier Economics. 2014. *Rates of Return to Investment in Science and Innovation*. https://assets.publishing.service.gov.uk/media/5a7f02a840f0b62305b8490b/bis-14-990-rates-of-return-to-investment-in-science-and-innovation-revised-final-report.pdf.

Ge, Xiou, Richard T. Goodwin, Haizi Yu, et al. 2022. "Accelerated Design and Deployment of Low-Carbon Concrete for Data Centers." arXiv:2204.05397. Preprint, arXiv. https://doi.org/10.48550/arXiv.2204.05397.

Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. 2023. “Gemini: A Family of Highly Capable Multimodal Models.” arXiv:2312.11805. Preprint, arXiv. https://doi.org/10.48550/arXiv.2312.11805.

Gil, Elad, and Sarah Guo. 2024. *With OpenAI Co-Founder & Chief Scientist Ilya Sutskever*. Episode 39. No Priors: AI, Machine Learning, Tech, & Startups. November 2. 41:59. https://www.youtube.com/watch?v=Ft0gTO2K85A.

Glazer, Elliot, Ege Erdil, Tamay Besiroglu, et al. 2024. “FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI.” arXiv:2411.04872. Preprint, arXiv. https://doi.org/10.48550/arXiv.2411.04872.

Google Deepmind. n.d. “About Google DeepMind.” https://deepmind.google/about/.

Google Research. n.d.-a. “Green Light.” http://sites.research.google/gr/greenlight/.

Google Research. n.d.-b. “Project Contrails: Preventing Contrails with AI.” https://sites.research.google/contrails/.

Google Sustainability. 2024. “2024 Environmental Report.” https://sustainability.google/reports/google-2024-environmental-report/.

Gottweis, Juraj, Wei-Hung Weng, Alexander Daryin, et al. 2025. “Towards an AI Co-Scientist.” arXiv:2502.18864. Preprint, arXiv. https://doi.org/10.48550/arXiv.2502.18864.

Grace, Katja, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner. 2024. “Thousands

of AI Authors on the Future of AI." arXiv:2401.02843. Preprint, arXiv. https://doi.org/10.48550/arXiv.2401.02843.

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. "The Llama 3 Herd of Models." arXiv:2407.21783. Preprint, arXiv. https://doi.org/10.48550/arXiv.2407.21783.

Griffin, Conor, Don Wallace, Juan Mateos-Garcia, Hanna Schieve, and Pushmeet Kohli. 2024. "A New Golden Age of Discovery." AI Policy Perspectives, November 26. https://www.aipolicyperspectives.com/p/a-new-golden-age-of-discovery.

Halper, Mark. 2025. "Farming with AI." *Communications of the ACM*, January 2. https://cacm.acm.org/news/farming-with-ai/.

Herr, Daniel, Tiffany Head, Clio von Petersdorff, et al. 2024. *The Economic Value of the Met Office*. London Economics. https://www.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/about-us/governance/met-office-evaluation-study---final-report-august-2024.pdf.

Ho, Anson, Ege Erdil, and Tamay Besiroglu. 2023. "Limits to the Energy Efficiency of CMOS Microprocessors." arXiv:2312.08595. Preprint, arXiv. https://doi.org/10.48550/arXiv.2312.08595.

Hu, Krystal. 2023. "ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note." *Reuters*, February 2. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

Huang, Kaixuan, Yuanhao Qu, Henry Cousins, et al. 2024. "CRISPR-GPT: An LLM Agent for Automated Design of Gene-Editing

Experiments." arXiv:2404.18021. Preprint, arXiv. https://doi.org/10.48550/arXiv.2404.18021.

IEA. 2022. *The Future of Heat Pumps: Executive Summary*. https://www.iea.org/reports/the-future-of-heat-pumps/executive-summary.

IEA. 2024. *Renewables 2024: Analysis and Forecast to 2030*. https://iea.blob.core.windows.net/assets/17033b62-07a5-4144-8dd0-651cdb6caa24/Renewables2024.pdf.

IEA. 2025a. *Aviation*. https://www.iea.org/energy-system/transport/aviation.

IEA. 2025b. *Electricity 2025: Demand*. https://www.iea.org/reports/electricity-2025/demand.

IEA. 2025c. *Energy and AI: Executive Summary*. https://www.iea.org/reports/energy-and-ai/executive-summary.

IEA. 2025d. *Global EV Outlook 2025*. https://www.iea.org/reports/global-ev-outlook-2025.

International Energy Agency. 2018. *The Future of Cooling: Opportunities for Energy-Efficient Air Conditioning*. OECD. https://doi.org/10.1787/9789264301993-en.

Isaac, Mike, Eli Tan, and Cade Metz. 2025. "A.I. Researchers Are Negotiating $250 Million Pay Packages. Just like N.B.A. Stars." *New York Times*, July 31. https://www.nytimes.com/2025/07/31/technology/ai-researchers-nba-stars.html.

Jimenez, Carlos E., John Yang, Alexander Wettig, et al. 2024. "SWE-Bench: Can Language Models Resolve Real-World GitHub

Issues?" arXiv:2310.06770. Preprint, arXiv. https://doi.org/10.48550/arXiv.2310.06770.

Kokotajlo, Daniel, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean. 2025. *AI 2027*. https://ai-2027.com/.

KP Jayatunga, Madura, Margaret Ayers, Lotte Bruens, Dhruv Jayanth, and Christoph Meier. 2024. "How Successful Are AI-Discovered Drugs in Clinical Trials? A First Analysis and Emerging Lessons." *Drug Discovery Today* 29 (6): 104009. https://doi.org/10.1016/j.drudis.2024.104009.

Kuzhagaliyeva, Nursulu, Samuel Horváth, John Williams, Andre Nicolle, and S. Mani Sarathy. 2022. "Artificial Intelligence-Driven Design of Fuel Mixtures." *Communications Chemistry* 5 (1): 111. https://doi.org/10.1038/s42004-022-00722-3.

Kwa, Thomas, Ben West, Joel Becker, et al. 2025. "Measuring AI Ability to Complete Long Tasks." arXiv:2503.14499. Preprint, arXiv, March 30. https://doi.org/10.48550/arXiv.2503.14499.

Laurent, Jon M., Joseph D. Janizek, Michael Ruzo, et al. 2024. "LAB-Bench: Measuring Capabilities of Language Models for Biology Research." arXiv:2407.10362. Preprint, arXiv. https://doi.org/10.48550/arXiv.2407.10362.

Leash Bio. 2024. "BELKA results suggest computers can memorize, but not create, drugs." https://leashbio.substack.com/p/belka-results-suggest-computers-can

Lowe, Derek. 2024a. "AI Drugs so Far." *In the Pipeline*, May 13. https://www.science.org/content/blog-post/ai-drugs-so-far.

Lowe, Derek. 2024b. “AI and Biology.” *In the Pipeline*, September 19. https://www.science.org/content/blog-post/ai-and-biology.

Lozano-Aguilera, Rubén. 2022. “More Ways to Drive Sustainably and Save Money with Google Maps.” Google, September 7. https://blog.google/around-the-globe/google-europe/eco-friendly-routing-in-europe/.

Lu, Chris, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. “The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery.” arXiv:2408.06292. Preprint, arXiv. https://doi.org/10.48550/arXiv.2408.06292.

Luo, Jerry, Cosmin Paduraru, Octavian Voicu, et al. 2022. “Controlling Commercial Cooling Systems Using Reinforcement Learning.” arXiv:2211.07357. Preprint, arXiv. https://doi.org/10.48550/arXiv.2211.07357.

Maguire, Yael. 2024. “How Satellites, Algorithms and AI Can Help Map and Trace Methane Sources.” Google, February 14. https://blog.google/outreach-initiatives/sustainability/how-satellites-algorithms-and-ai-can-help-map-and-trace-methane-sources/.

McNamee, Laura M., Michael Jay Walsh, and Fred D. Ledley. 2017. “Timelines of Translational Science: From Technology Initiation to FDA Approval.” *PLOS One* 12 (5): e0177371. https://doi.org/10.1371/journal.pone.0177371.

Merchant, Amil, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. “Scaling Deep Learning for Materials Discovery.” *Nature* 624 (7990): 80–85. https://doi.org/10.1038/s41586-023-06735-9.

Mirzadeh, Iman, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. “GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models.” arXiv:2410.05229. Preprint, arXiv. https://doi.org/10.48550/arXiv.2410.05229.

Miserendino, Samuel, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. 2025. “SWE-Lancer: Can Frontier LLMs Earn $1 Million from Real-World Freelance Software Engineering?” arXiv:2502.12115. Preprint, arXiv, May 2. https://doi.org/10.48550/arXiv.2502.12115.

Molina, Renato, and Ivan Rudik. 2024. “The Social Value of Hurricane Forecasts.” Working Paper No. 32548. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w32548.

Morris, Meredith Ringel, Jascha Sohl-Dickstein, Noah Fiedel, et al. 2024. “Levels of AGI for Operationalizing Progress on the Path to AGI.” arXiv:2311.02462. Preprint, arXiv, June 5. https://doi.org/10.48550/arXiv.2311.02462.

Novikov, Alexander, Ngân Vũ, Marvin Eisenberger, et al. 2025. “AlphaEvolve: A Coding Agent for Scientific and Algorithmic Discovery.” arXiv:2506.13131. Preprint, arXiv, June 16. https://doi.org/10.48550/arXiv.2506.13131.

Owen, David. 2024a. “How Predictable Is Language Model Benchmark Performance?” arXiv:2401.04757. Preprint, arXiv. https://doi.org/10.48550/arXiv.2401.04757.

Owen, David. 2024b. *Interviewing AI Researchers on Automation of AI R&D*. Epoch AI.

https://epoch.ai/blog/interviewing-ai-researchers-on-automation-of-ai-rnd.

Oxford Academic. 2024. "Researchers and AI: Survey Findings." https://academic.oup.com/pages/ai-survey-findings.

Park, Hyun, Xiaoli Yan, Ruijie Zhu, et al. 2024. "A Generative Artificial Intelligence Framework Based on a Molecular Diffusion Model for the Design of Metal-Organic Frameworks for Carbon Capture." *Communications Chemistry* 7 (1): 21. https://doi.org/10.1038/s42004-023-01090-2.

Park, Toby. 2017. *Evaluating the Nest Learning Thermostat: Four Field Experiments Evaluating the Energy Saving Potential of Nest's Smart Heating Control*. The Behavioural Insights Team. https://www.bi.team/wp-content/uploads/2017/11/311013-Evaluating-Nest-BIT-Exec-Tech-Summaries.pdf.

Parks, Elizbeth. 2024. "Empty Nest-ers Beware: Smart Thermostats Are Here to Stay." *Factor This*, January 22. https://www.renewableenergyworld.com/power-grid/smart-grids/empty-nest-ers-beware-smart-thermostats-are-here-to-stay/.

Patel, Dwarkesh. 2023. *Dario Amodei (Anthropic CEO): Scaling, Alignment, & AI Progress*. August 8. 01:58:43. https://www.dwarkesh.com/p/dario-amodei.

Pilz, Konstantin F., James Sanders, Robi Rahman, and Lennart Heim. 2025. "Trends in AI Supercomputers." arXiv:2504.16026. Preprint, arXiv. https://doi.org/10.48550/arXiv.2504.16026.

Price, Ilan, Alvaro Sanchez-Gonzalez, Ferran Alet, et al. 2024. "GenCast: Diffusion-Based Ensemble Forecasting for Medium-Range

Weather." arXiv:2312.15796. Preprint, arXiv. https://doi.org/10.48550/arXiv.2312.15796.

Rachitsky, Lenny. 2025. *OpenAI Researcher on Why Soft Skills Are the Future of Work | Karina Nguyen*. Lenny's Podcast. February 9. 01:14:34. https://www.youtube.com/watch?v=DeskgjrLxxs.

Rasp, Stephan, Stephan Hoyer, Alexander Merose, et al. 2024. "WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models." arXiv:2308.15560. Preprint, arXiv. https://doi.org/10.48550/arXiv.2308.15560.

Ritchie, Hannah. 2024. "What Share of Global CO_2 Emissions Come from Aviation?" Our World in Data, April 8. https://ourworldindata.org/global-aviation-emissions.

Romera-Paredes, Bernardino, Mohammadamin Barekatain, Alexander Novikov, et al. 2024. "Mathematical Discoveries from Program Search with Large Language Models." *Nature* 625 (7995): 468–75. https://doi.org/10.1038/s41586-023-06924-6.

Rozite, Vida, Emi Bertoli, and Brendan Reidenbach. 2023. "Data Centres and Data Transmission Networks." IEA. https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks.

Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. 2023. "Are Emergent Abilities of Large Language Models a Mirage?" arXiv:2304.15004. Preprint, arXiv. https://doi.org/10.48550/arXiv.2304.15004.

Sevilla, Jaime, Tamay Besiroglu, Ben Cottier, et al. 2024. "Can AI Scaling Continue through 2030?" Epoch AI. https://epoch.ai/blog/can-ai-scaling-continue-through-2030.

Sevilla, Jaime, Tamay Besiroglu, Owen Dudney, and Anson Ho. 2022. *The Longest Training Run*. Epoch AI. https://epoch.ai/blog/the-longest-training-run.

Sevilla, Jaime, and Edu Roldán. 2024. *Training Compute of Frontier AI Models Grows by 4-5x per Year*. Epoch AI. https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year.

Shen, Bo-Wen, Roger A. Pielke, Xubin Zeng, and Xiping Zeng. 2024. "Exploring the Origin of the Two-Week Predictability Limit: A Revisit of Lorenz's Predictability Studies in the 1960s." *Atmosphere* 15 (7): 837. https://doi.org/10.3390/atmos15070837.

Snodin, Ben, David Owen, and Luke Emberson. 2025. "The Combined Revenues of Leading AI Companies Grew by Over 9x in 2023-2024." Epoch AI. https://epoch.ai/data-insights/ai-companies-revenue.

Stern, Lord Nicholas, and Mattia Romani. 2025. "AI's Role in the Climate Transition and How It Can Drive Growth." World Economic Forum, January 16. https://www.weforum.org/stories/2025/01/artificial-intelligence-climate-transition-drive-growth/.

Stern, Nicholas, Mattia Romani, Roberta Pierfederici, et al. 2025. "Green and Intelligent: The Role of AI in the Climate Transition." *npj Climate Action* 4 (1): 56. https://doi.org/10.1038/s44168-025-00252-3.

Sun, Duxin, Wei Gao, Hongxiang Hu, and Simon Zhou. 2022. "Why 90% of Clinical Drug Development Fails and How to Improve It?" *Acta Pharmaceutica Sinica B* 12 (7): 3049–62. https://doi.org/10.1016/j.apsb.2022.02.002.

Sutton, Richard. 2019. “The Bitter Lesson”. http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Talbott, Chris. 2022. “Helping Farmers with Cloud Technology, Up Close and Global.” Google, June 3. https://blog.google/products/google-cloud/helping-farmers-with-cloud-technology-up-close-and-global/.

Tao, Terence. 2024. “I Have Played a Little Bit with OpenAI’s New Iteration of #GPT, GPT-O1, Which Performs an Initial Reasoning Step before Running the LLM.” Mathstodon, September 13. https://mathstodon.xyz/@tao/113132502735585408.

Todd, Benjamin. 2024. *The Market Expects AI Software to Create Trillions of Dollars of Value by 2027*. Substack newsletter. April 29. https://benjamintodd.substack.com/p/the-market-expects-ai-software-to.

US EPA. 2025. “Fast Facts on Transportation Greenhouse Gas Emissions.” Overviews and Factsheets. June 6. https://www.epa.gov/greenvehicles/fast-facts-transportation-greenhouse-gas-emissions.

Vendrow, Joshua, Edward Vendrow, Sara Beery, and Aleksander Madry. 2025. “Do Large Language Model Benchmarks Test Reliability?” arXiv:2502.03461. Preprint, arXiv, February 5. https://doi.org/10.48550/arXiv.2502.03461.

Wijk, Hjalmar, Tao Lin, Joel Becker, et al. 2025. “RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents against Human Experts.” arXiv:2411.15114. Preprint, arXiv. https://doi.org/10.48550/arXiv.2411.15114.

Wikipedia. 2025. “Vibe coding.” August 19. https://en.wikipedia.org/wiki/Vibe_coding.

Winters, Timothy, William Davie, and Deanna Weidenhamer. 2011. *Quarterly Retail E-Commerce Sales 4th Quarter 2010*. U.S. Census Bureau. https://www2.census.gov/retail/releases/historical/ecomm/10q4.pdf .

Wu, Carole-Jean, Ramya Raghavendra, Udit Gupta, et al. 2022. “Sustainable AI: Environmental Implications, Challenges and Opportunities.” arXiv:2111.00364. Preprint, arXiv. https://doi.org/10.48550/arXiv.2111.00364.

Wu, Kan, Jianrong Ding, Jingli Lin, et al. 2025. “Big-Data Empowered Traffic Signal Control Could Reduce Urban Carbon Emission.” *Nature Communications* 16 (1): 2013. https://doi.org/10.1038/s41467-025-56701-4.

Yang, John, Carlos E. Jimenez, Alexander Wettig, et al. 2024. “SWE-Agent: Agent-Computer Interfaces Enable Automated Software Engineering.” arXiv:2405.15793. Preprint, arXiv. https://doi.org/10.48550/arXiv.2405.15793.

Yang, Kaiyu, Gabriel Poesia, Jingxuan He, et al. 2024. “Formal Mathematical Reasoning: A New Frontier in AI.” arXiv:2412.16075. Preprint, arXiv. https://doi.org/10.48550/arXiv.2412.16075.

Yepis, Erin. 2024. “Developers Get by with a Little Help from AI: Stack Overflow Knows Code Assistant Pulse Survey Results.” Stack Overflow. https://stackoverflow.blog/2024/05/29/developers-get-by-with-a-little-help-from-ai-stack-overflow-knows-code-assistant-pulse-survey-results/.

You, Joshua, and David Owen. 2025. *Scaling Intelligence: The Exponential Growth of AI's Power Needs*. EPRI. https://www.epri.com/research/products/000000003002033669.

Yue, Yang, Zhiqi Chen, Rui Lu, et al. 2025. "Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs beyond the Base Model?" arXiv:2504.13837. Preprint, arXiv. https://doi.org/10.48550/arXiv.2504.13837.

Yuventi, Jumie, and Roshan Mehdizadeh. 2013. "A Critical Analysis of Power Usage Effectiveness and Its Use in Communicating Data Center Energy Consumption." *Energy and Buildings* 64: 90–94. https://doi.org/10.1016/j.enbuild.2013.04.015.

Zambaldi, Vinicius, David La, Alexander E. Chu, et al. 2024. "De Novo Design of High-Affinity Protein Binders with AlphaProteo." arXiv:2409.08022. Preprint, arXiv. https://doi.org/10.48550/arXiv.2409.08022.

About Epoch AI

Epoch AI is a multidisciplinary non-profit research institute investigating the future of artificial intelligence. We examine the driving forces behind AI and forecast its economic and societal impact.

We emphasize making our research accessible through our reports, models and visualizations to help ground the discussion of AI on a solid empirical footing. Our goal is to create a healthy scientific environment, where claims about AI are discussed with the rigor they merit.

To learn more about Epoch AI, contact us at info@epochai.org.