TECHNICAL REPORT | OCTOBER 2025

Inference Scaling and AI Governance

Toby Ord

GovAI

Inference Scaling and AI Governance

Toby Ord[1]

[1]Oxford Martin AI Governance Initiative, University of Oxford

The shift from scaling up the compute used to pre-train AI systems (pre-training compute) to scaling up the amount used to run them (inference compute) may have profound effects on AI governance. The nature of these effects depends crucially on whether this new inference compute will primarily be used to improve model performance during external deployment or as part of a more complex training programme within the lab. Rapid scaling of inference-at-deployment would somewhat lower the importance of open-weight models (and of securing the weights of closed models), reduce the impact of the first human-level models, change the business model for frontier AI, reduce the need for power-intensive data centres, and potentially undermine AI governance measures that rely on training-compute thresholds. Rapid scaling of inference-during-training would have more ambiguous effects that range from a revitalisation of pre-training scaling to a form of recursive self-improvement via iterated distillation and amplification.

Introduction

For years, AI progress has followed a predictable pattern: use more computing power to build bigger models, and performance improves accordingly. But recent developments suggest this era may be ending. AI progress is increasingly driven by scaling up inference compute: the amount of computing power a model of a given size can use to respond to a user's prompt.

The implications of inference scaling depend on whether AI developers focus more on scaling inference-at-deployment or inference-during-training. Scaling inference-at-deployment refers to the use of additional computational resources when serving a request. Scaling inference-during-training refers to a developer scaling up the inference used to complete some task, thus improving the model's performance, and then using the resulting data to train models.

If labs invest more of their resources scaling inference-at-deployment, this may:

- **Reduce the number of simultaneously served copies of each new model.** If model performance relies on increased inference compute, each model copy would require more compute to run.
- **Increase the cost of the first human-level AI systems.** If inference compute drives performance, then the highest-performing system could use many orders of magnitude more than less-performant ones.
- **Somewhat reduce the value of securing model weights.** If performance gains are derived at inference, then it matters less who has the model weights and more who obtains inference compute.
- **Somewhat reduce the benefits and risks of open-weight models.** If inference matters more for performance, open-weight models on their own matter less for both benefits and risks of AI.
- **Allow unequal performance for different tasks and for different users.** When inference drives performance, those with access to greater compute resources will have a more performant system.
- **Change the business model and industry structure.** Greater reliance on inference compute would increase marginal costs for the AI industry.
- **Reduce the need for monolithic data centres.** Inference compute does not require the type of large, centralized computing infrastructure needed for big pre-training runs.
- **Complicate the strategy of AI governance via compute thresholds.** As inference drives performance, compute thresholds may become less useful for triggering greater scrutiny and safeguards.

If companies instead focus on using inference compute during training, the consequences are less clear. This may lead to:

- **Less transparency about state-of-the-art models.** If labs scale inference-during-training, policymakers may have less insight into model capabilities, reducing readiness for advanced AI.
- **Shorter timelines to transformative AGI.** Iterated distillation may speed up progress toward AGI.

Next, I explain why I think we should consider the shift to inference scaling as a new paradigm, rather than a simple continuation of the familiar scaling era. Then, I examine how scaling inference-at-deployment could reshape the AI landscape, affecting everything from business models to regulatory frameworks. Finally, I consider the implications of scaling

inference-during-training, which could accelerate AI development in unexpected ways while reducing transparency into the most advanced systems. Throughout, I assess what these changes mean for policymakers seeking to govern AI systems effectively.

The end of an era – for both training and governance

The intense year-on-year scaling up of AI training runs has been one of the most dramatic and stable markers of the large language model (LLM) era. Indeed, it has been widely taken to be a permanent fixture of the AI landscape and the basis of many approaches to AI governance.

To date, researchers have found that AI systems tend to develop new capabilities with relative predictability as the models are trained using more computing power ("compute") and data. This trend, known as "scaling," was also useful for policymakers. As performance increased with computing power, governance mechanisms could focus on the most capable models at the "frontier" by using computing power as a proxy for performance. Because concerning capabilities are likely to emerge in the most capable models, governance mechanisms could look to computing-power thresholds ("compute thresholds") as a way to identify systems of concern, while limiting impacts on parts of the AI industry that fall below those thresholds. Researchers and policymakers could use the amount of compute used to train a model in order to define the scope of the rules and oversight mechanisms that apply to AI development. Researchers have fiercely debated the utility of using compute thresholds in this matter,[1] but in the absence of a viable alternative, compute thresholds have emerged as an important tool in AI governance.

Major regulatory efforts have built directly on these assumptions. The EU AI Act uses a threshold based on training compute – specifically 10^{25} floating point operations (FLOPs) – to define "general purpose AI systems with systemic risk", which face additional requirements. The Biden administration's (now rescinded) Executive Order on Artificial Intelligence required companies to report information about models trained using more than 10^{26} FLOPs, and California's recently enacted AI safety law relies on a compute (and revenue) threshold to identify firms subject to transparency requirements.

The Shift to Inference Scaling

Recent developments, though, suggest that these assumptions may be breaking down. Reports from leading labs,[2] supported by evidence about the capabilities of recent AI systems,[3] claim

[1] Lennart Heim and Leonie Koessler, "Training Compute Thresholds: Features and Functions in AI Regulation," arXiv:2405.10799, August 6, 2024, https://doi.org/10.48550/arXiv.2405.10799; Sara Hooker, "On the Limitations of Compute Thresholds as a Governance Strategy," arXiv:2407.05694, July 30, 2024, https://doi.org/10.48550/arXiv.2407.05694.
[2] Stephanie Palazzolo et al., "OpenAI Shifts Strategy as Rate of 'GPT' AI Improvements Slows," The Information, November 9, 2024, https://www.theinformation.com/articles/openai-shifts-strategy-as-rate-of-gpt-ai-improvements-slows.
[3] Nathan Lambert, "GPT-4.5: 'Not a Frontier Model'?," Interconnects, November 24, 2023, https://www.interconnects.ai/p/gpt-45-not-a-frontier-model.

that building bigger models – or scaling pre-training – substantially larger than GPT-4 has led to only modest gains in practical utility.[4] A possible reason for the slowdown in performance increases is that AI developers are running out of high-quality training data. While the scaling laws which predict how model performance improves with more data and compute might still be operating, the ability to harness them through rapid scaling of pre-training may not be. What was taken to be a fixture may instead have been just one important era in the history of AI development, an era which is now coming to a close.

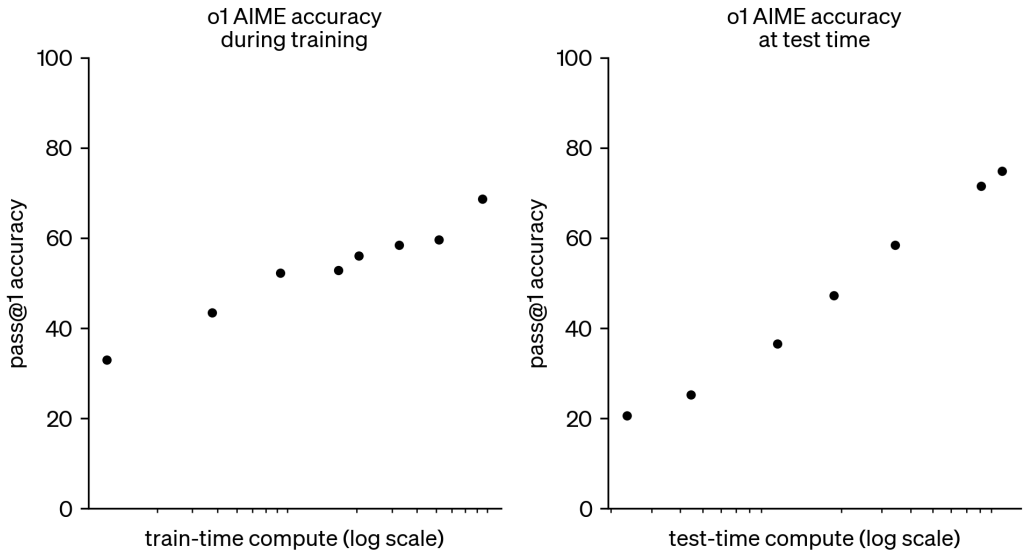


Figure 1. Reported model performance vs. increased compute of OpenAI's o1 system. The left chart shows the model's performance improving as a result of additional post-training reinforcement learning. The right chart shows performance improving as a result of increasing the amount of inference compute after deployment (Source: OpenAI, see footnote 5).

What will come next? Just before the reports of these difficulties emerged, OpenAI announced o1,[5] a breakthrough "reasoning" model that illustrated how labs are relying on techniques beyond scaling pre-training to deliver performance. Their announcement included a chart (Figure 1) showing how the model's performance on a difficult mathematics benchmark could be improved by increasing compute in two ways. The first was by dedicating more compute to post-training reinforcement learning (in which the model is fine-tuned through feedback to improve its overall performance); the second was by increasing the inference compute used on the current task (giving the model more computational resources to generate each response).

[4] Maxwell Zeff, "Current AI Scaling Laws Are Showing Diminishing Returns, Forcing AI Labs to Change Course," Techcrunch, November 20, 2024, https://techcrunch.com/2024/11/20/ai-scaling-laws-are-showing-diminishing-returns-forcing-ai-labs-to-change-course/.
[5] OpenAI, "Learning to Reason with LLMs," September 12, 2024, https://openai.com/index/learning-to-reason-with-llms/.

As Figure 1 shows, OpenAI claimed that using more inference compute led to impressive gains in model performance. Similarly, work on the trade-off between pre-training compute and inference compute suggests that, on the current margins, increasing inference compute on the task at hand by a factor of 10 can improve performance as much as increasing pre-training compute by 3–10x, with performance gains often plateauing after scaling up inference-at-deployment by a few orders of magnitude.[6]

These developments have led to intense speculation that the previous era of scaling pre-training compute could be followed by an era of scaling up inference compute. Some, including AI company executives, have suggested this represents a continuation of the previous paradigm.[7] To the contrary, I believe there are a number of key differences between scaling pre-training and scaling inference that have profound implications for both AI companies and AI governance.

Uncertainties of the Inference Era

Two key questions shape how we should think about this shift to inference scaling. One question is whether pre-training scaling has truly plateaued or if it will continue at a slower rate. Epoch AI suggests that the compute used in LLM pre-training grew at about 5x per year from 2020 to 2024.[8] Today, the rate seems to be lower – but how much lower remains unclear.

A second – and ultimately more important – question concerns where inference scaling will be applied. We can view the current AI pipeline as having three main stages (Figure 2):

1. **Pre-training:** teaching models to predict text through methods like next-token prediction
2. **Post-training:** refining models with techniques such as reinforcement learning from human feedback (RLHF) or reinforcement learning from AI feedback (RLAIF)
3. **Deployment:** making the trained model available to be deployed on various tasks through chat interfaces, API calls, or other platforms

[6] Pablo Villalobos and David Atkinson, Trading off Compute in Training and Inference (Epoch AI, 2023), https://epoch.ai/blog/trading-off-compute-in-training-and-inference.
[7] For example, Dario Amodei has said "Every once in a while, the underlying thing that is being scaled changes a bit, or a new type of scaling is added to the training process. From 2020–2023, the main thing being scaled was pretrained models: models trained on increasing amounts of internet text with a tiny bit of other training on top. In 2024, the idea of using reinforcement learning (RL) to train models to generate chains of thought has become a new focus of scaling." Dario Amodei, "On DeepSeek and Export Controls," January 2025, https://www.darioamodei.com/post/on-deepseek-and-export-controls.
[8] Jaime Sevilla and Edu Roldán, Training Compute of Frontier AI Models Grows by 4-5x per Year (Epoch AI, 2024), https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year.

Figure 2. Three stages of the AI pipeline: the first two stages, inside the box, take place during model development. (Source: author.)

The crucial question is whether scaled-up inference compute will primarily be used during deployment (like in o1 and DeepSeek's R1) or as part of a more complex post-training process. For example, reports suggest that OpenAI may have trained its o3 model by using many runs of o1 to generate training data, essentially using inference scaling to improve the training process itself. Similarly, xAI reportedly used roughly as much compute for reinforcement learning as for pre-training in order to develop Grok 4.[9]

Each possibility has important – but different – implications for AI governance. I argue that inference scaling means that many ideas in AI governance will need to be either adjusted or overhauled. Those of us in the field need to examine how this affects our existing approaches and assumptions.

Scaling inference-at-deployment

Consider first the scenario where most compute scaling is used to grow the amount of inference compute used during deployment. In this scenario, the capabilities of pre-trained systems remain at approximately GPT-5 level or only advance slowly, while new capabilities are unlocked via increasing inference compute. Some compute may be allocated to post-training aimed at having systems productively reason for longer (e.g. the reinforcement learning in the train-time compute graph in Figure 1), but this analysis assumes that the resulting performance gains remain relatively small compared to deployment compute scaling. Grok 4 provides some inconclusive support for this assumption: though it used approximately 10^{26} FLOP for post-training, its performance does not seem to have improved significantly, perhaps due to the extreme inefficiency of reinforcement learning for frontier models.[10]

This shift to inference-at-deployment would reshape several aspects of AI governance. It would affect how many copies of advanced models can be deployed simultaneously, alter the economics of AI systems, change the strategic importance of model weights and open-source releases, and potentially undermine current regulatory frameworks based on

[9] Nathan Lambert, "xAI's Grok 4: The Tension of Frontier Performance with a Side of Elon Favoritism," Interconnects, July 12, 2025, https://www.interconnects.ai/p/grok-4-an-o3-look-alike-in-search.
[10] Toby Ord, "The Extreme Inefficiency of RL for Frontier Models," September 19, 2025, https://www.tobyord.com/writing/inefficiency-of-reinforcement-learning.

training-compute thresholds. Each of these changes carries significant implications for how we govern AI systems.

Reducing the number of simultaneously served copies of each new model

It currently takes a vast number of chips to train a frontier model. Once the model is trained, those chips can be used for inference to deploy a large number of simultaneous copies of that model. Dario Amodei of Anthropic estimates this to be "millions" of copies.[11] This number of copies is a key parameter for AI governance as it affects the size of the immediate impact on the world the day the new model is ready. A shift to scaling inference-at-deployment would lower the number of copies that could be deployed with the same number of chips. For example, if inference-at-deployment is scaled by two orders of magnitude, then the number of copies drops by a factor of 100 and the new model can only be immediately deployed to 1% as many tasks as an equally powerful pre-trained model could be.[12]

Increasing the cost of first human-level AI systems

A related parameter is how expensive the first "human-level" AI systems will be to run. In the pre-training scaling paradigm, deploying such systems may well cost much less than human labour, meaning that they could be immediately deployed at a great profit. These profits could be ploughed back into acquiring more compute to run more copies of the system, creating a powerful feedback loop. But each additional order of magnitude that goes to inference-at-deployment may increase the cost of using these systems by up to an order of magnitude.

This increased inference-time cost will blunt the immediate impact of reaching any level of performance threshold and may even create an initial period where human-level AI systems are more expensive than equivalent human labour.[13] If so, such systems could be available for policymaker demonstrations or safety research before they have transformative effects on society.

[11] Dario Amodei, "Machines of Loving Grace," October 2024, https://www.darioamodei.com/essay/machines-of-loving-grace.
[12] Or, somewhat equivalently, it might be better thought of as slowing these systems down by that factor (e.g. 100x). Amodei's estimate is that AI systems are currently 10x–100x human speed, but if they reach intelligence via inference scaling, they may be slower than humans. Both ways of looking at it lead to the same reduction in the "human-days-equivalent of AI work each day" when the systems are switched from training to deployment.
[13] Obviously, the fact that AI is already much better than humans at some tasks while much worse at others complicates this idea of reaching "human-level", but I believe it is still a useful lens. For example, you can ask whether the first systems that can perform a particular job better than humans will cost more or less than human wages for that job.

Somewhat reducing the value of securing model weights

Consider a scenario where frontier model training compute plateaus at approximately the GPT-5 level while inference-at-deployment scales by a factor of 100. In this case, stealing model weights becomes much less appealing because the perpetrator still faces the full inference-at-deployment costs. Since these inference costs would dominate the total expense of operating the model at scale, obtaining the weights for free provides relatively little economic benefit. The value proposition of model theft diminishes when the largest costs cannot be avoided through theft.

On the other hand, inference scaling might increase certain misuse risks. If the actor stealing model weights does not need to deploy their model at scale, but are rather interested in high performance on a small number of tasks – such as acquiring information needed to access chemical or biological weapons – then having model performance scale with inference-at-deployment means a larger number of models can reach the requisite performance.

Somewhat reducing the benefits and risks of open-weight models

Inference scaling would also affect both the benefits and drawbacks of open-weight models. If open-weight models require vast amounts of inference-at-deployment from their users, then they are much less attractive to those users than are models of equivalent capability that were entirely pre-trained. So open-weight models could become both less valuable to users and less concerning from a capability proliferation perspective. They would become less strategically important overall. However, as noted above, certain misuse risks tied to achieving high performance on a small number of tasks could increase.

Unequal performance for different tasks and for different users

Since inference scaling affects how AI performance varies across different applications and user groups, it may create new forms of inequality in access to advanced capabilities.

Scaling inference-at-deployment helps most with tasks where the solution is objectively verifiable, such as certain kinds of maths and programming tasks. It can also be useful for tasks involving many steps. Two kinds of tasks that benefit from inference scaling are:

- Tasks that require methodical reasoning ("System 2 thinking") when performed by humans,
- Tasks that typically take humans a long time, indicating that they can benefit from a lot of thinking before diminishing marginal returns kick in.

Because some tasks benefit more from additional inference than others, it is possible to tailor the amount of inference compute to the task, spending, for example, 1,000x more on a hard mathematics problem than on a simpler, more intuitive task. This kind of tailoring is not possible with pre-training scaling, where scaling up by 10x increases the costs for everything.

The fact that performance can be increased by spending more on inference compute also changes the dynamics of AI accessibility: users with more financial resources can access greater AI capabilities. This trend is already evident at OpenAI, which now charges 10x more for access to the version of their models which use the most inference compute. The era in which all users received the same or similar AI services is over.

Changing the business model and industry structure

The LLM business model has had a lot in common with software: big upfront development costs and then comparatively low marginal costs per additional customer. When marginal costs per user are lower than average costs, companies benefit from economies of scale. This incentivizes them to set prices low to acquire customers, which in turn tends to create an industry with only a handful of players.

However, if the next two orders of magnitude of compute scaling go into inference-at-deployment instead of pre-training, this economic structure would change. The shift would disrupt existing business models and perhaps allow smaller players to compete in the industry.

Reducing the need for monolithic data centres

While training benefits from compute being localised in the same data centre, inference-at-deployment can be more easily distributed across different locations. Thus scaling inference-at-deployment by several orders of magnitude would reduce reliance on large centralized data centres. This would alleviate some current infrastructure bottlenecks, such as the challenge of securing a large amount of electrical power in one location.

This shift in required compute would complicate government oversight strategies that rely on monitoring and shaping infrastructure projects. It will also make it harder for governments to keep track of new frontier models simply by tracking activity in the largest data centres. As inference compute can be provided by a greater number of players in the compute ecosystem, know-your-customer rules, data centre monitoring, and chip export controls are likely to be less effective in controlling AI diffusion.

Complicate the strategy of AI governance via compute thresholds

A final implication of inference scaling is that it complicates current regulatory frameworks that rely on training-compute thresholds to identify potentially dangerous AI systems.

Many AI governance frameworks are based around regulating only those models above a certain threshold of training compute.[14] For example, parts of the EU AI Act focus on models trained using at least 10^{25} FLOP, while the (now rescinded) Biden-era US executive order and the recently passed SB53 in California used a threshold of 10^{26} FLOP. These thresholds allow regulators to draw a line around a handful of systems with especially significant or uncertain capabilities, without needing to regulate the great majority of AI models.

However, if capabilities can be increased via scaling inference-at-deployment, then a model trained using an amount of compute below these thresholds might be amplified to become as powerful as a model that would have exceeded them. For example, a model trained with 10^{24} FLOP might use 10,000 times more inference compute to perform at the level of a model trained with 10^{27} FLOP. This complicates the use of a training-compute threshold to trigger governance measures.

At first, the threat might be that someone scales up inference-at-deployment by a very large factor for a small number of important tasks. If the inference scale-up is only happening on a small fraction of all tasks the model is deployed on, one could use a very high scale-up factor (such as 100,000x) and suddenly operate at the level of a new tier of model.

Current techniques for inference scaling do face limitations, often hitting performance plateaus that cannot be exceeded by any amount of additional compute. Exceeding these plateaus requires substantial research and engineering efforts. However, AI companies are already developing better ways to drastically scale inference compute before performance plateaus. OpenAI's o3 model, for example, demonstrated the ability to use 10,000x more compute than their smallest reasoning model, o1-mini (Figure 3).

[14] Lennart Heim and Leonie Koessler, "Training Compute Thresholds: Features and Functions in AI Regulation," arXiv:2405.10799, August 6, 2024, https://doi.org/10.48550/arXiv.2405.10799.

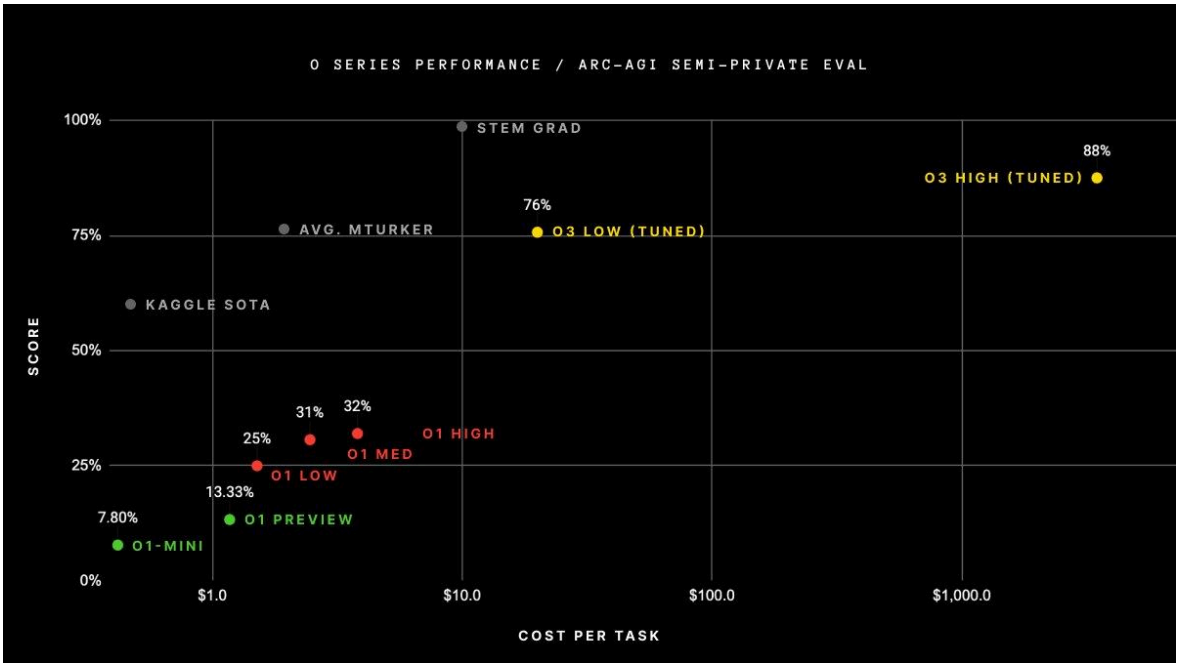


Figure 3 | Performance vs. cost for various OpenAI models showing performance gains from scaling up inference-at-deployment across three orders of magnitude. (Source: Arc Prize[15])

Leading companies have also been expanding their data centre capacity and improving algorithmic efficiency such that they may already have 100x the effective compute of the first data centres capable of serving GPT-4 to customers. This would allow them to offer wider access to large amounts of inference compute. For example, OpenAI's deep research model (based on o3) may well exceed the performance of a system pre-trained on 10^{26} FLOP, even if it is technically below that threshold.

However, while the increased reliance on inference scaling reduces the correlation between training compute and the concerning capabilities of AI models, this does not necessarily imply that compute thresholds should be abandoned. After all, models trained using large amounts of compute can still benefit from inference scaling, and the most capable models are still likely to be those that rely on large amounts of compute. Moreover, inference scaling techniques themselves face limitations and performance plateaus. Nonetheless, the shift toward inference scaling may require adjustments to how we use some tools in the AI governance toolbox and have implications for AI deployment.

Regulators need to make sure that inference scaling is taken into account when assessing the risk of models. If models are served or could be used with significant inference-at-deployment, it is not sufficient to just look at model performance from a single

[15] Francis Cholet. "OpenAI o3 Breakthrough High Score on ARC-AGI-PUB." Arc Prize Blog, December 2024. https://arcprize.org/blog/oai-o3-pub-breakthrough.

forward pass without reasoning tokens. Notably, the General-Purpose AI Code of Practice – which details the requirements of the EU AI Act on the most advanced models – requires that developers account for inference compute.[16]

Another way to update compute thresholds is to say that they cover both systems above 10^{26} FLOP of pre-training and systems above some smaller threshold (e.g. 10^{24} FLOP of pre-training) that have undergone post-training to enable high-inference deployment. But this would increase complexity and blur the clear demarcation lines that make current frameworks effective.

Capabilities-based thresholds represent another possible tool to identify models of concern. Rather than relying on proxy measures like computing power to identify potentially risky models, regulators might rely more heavily on evaluations of models' capabilities to carry out specific tasks, assuming significant inference compute budgets. Though making such a change would not be possible under the recently passed SB53 in California, it would be possible under the EU's AI Act.[17]

Scaling inference-during-training

AI labs may also be able to reap tremendous benefit from these inference-scaled models by using them as part of the training process. If so, the large scale-up of compute resources could go into post-training rather than deployment. This would have very different implications for AI governance.

In this section, we'll focus on the implications of a pure strategy of using inference scaling *only* during the training process. This will clarify its implications for AI governance, though realistically we will see inference scaling in both training and deployment.

Generating synthetic training data

An obvious approach to scaling inference-during-training is to use an inference-scaled model to generate large amounts of high-quality synthetic data – artificially generated data – on which to pre-train a new base model. This would make sense if the challenges in scaling up pre-training beyond GPT-4 stem from a lack of high-quality training data. For example, court documents have revealed that Meta trained models on a Russian repository of copyrighted books, LibGen, without permission because they were unable to reach GPT-4 level without it.[18]

[16] "EU AI Act: General-Purpose AI Code of Practice," EU AI Act: GPAI Code of Practice, 2025, https://code-of-practice.ai/.
[17] Article 51 gives the AI Office powers to designate models as general-purpose AI with systemic risk based on its capabilities. "Article 51: Classification of General-Purpose AI Models as General-Purpose AI Models with Systemic Risk," EU Artificial Intelligence Act, August 2, 2025, https://artificialintelligenceact.eu/article/51/.
[18] "Kadrey v. Meta, Document 391, Exhibit K, Vo Declaration," January 14, 2025, https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.391.24.pdf.

Anthropic recently agreed to pay $1.5 billion to settle a lawsuit brought by authors over the firm's reliance on datasets of pirated books.[19]

This strongly suggests that even though there are still many pieces of text on the internet that have not been used for AI training (about 30x as many as were used to pre-train GPT-4[20]), performance is limited by a lack of *high-quality* tokens. Developers have already tried to supplement the training data with synthetic data produced by an LLM, but if the issue is more about quality than quantity, then they need the best synthetic data they can get.

Inference scaling can help with this by making the model that produces the synthetic data more capable. This works particularly well in areas like mathematics or programming, where one can objectively verify the accuracy and efficiency of a model's answer. The training process could involve using advanced reasoning models to generate lots of proofs and computer programs, testing them for quality, and adding the best ones to the dataset used for pre-training the next base model.

Being able to verify correct answers in mathematics and coding is particularly important for getting the right training signal. But even for domains that are less black and white, it may be possible to use more inference compute to generate better synthetic data. For example, one could create many essays; intensively edit them; assess them for originality, insightfulness, and accuracy; and add the best ones to the stock of synthetic data.

One could also apply this technique to the stock of human-generated training data, assessing all documents in the training data and discarding low-quality ones. This could either improve the average quality of the existing training data or make some fraction of the unused data usable.

On its own, this approach of scaling inference-during-training to produce synthetic data for pre-training is not so interesting from an AI governance perspective. Its main direct effect is to allow the scaling of pre-training compute to recommence, reinvigorating the existing scaling paradigm.

Iterated distillation and amplification

But a modification of this approach may drive more rapid growth in AI capabilities. The idea is to repeatedly improve a model by:

1. Using inference scaling to boost its performance

[19] "What Authors Need to Know about the $1.5 Billion Anthropic Settlement," The Authors Guild, October 2, 2025, https://authorsguild.org/advocacy/artificial-intelligence/what-authors-need-to-know-about-the-anthropic-settlement/.
[20] Jaime Sevilla et al., Can AI Scaling Continue through 2030? (Epoch AI, 2024), https://epoch.ai/blog/can-ai-scaling-continue-through-2030.

2. Training a new model to replicate that boosted performance without the extra inference compute

3. Repeating this process many times

This process powered the advanced self-play in DeepMind's AlphaGo Zero (see Box 1) and was also independently discovered by Anthony et al. and, in the context of AI safety, by Christiano.[21]

Box 1. Iterated Distillation and AlphaGo Zero

- In the case of AlphaGo Zero, you start with a base model, M_0, that takes a representation of the Go board and produces two outputs: predictions about which moves a skilled player would choose, and an estimate of how likely the active player is to win the game.[22] This model will rely on an intuitive, fast mode of thinking – or "System 1" approach – to game playing, making quick decisions without systematically thinking through future moves.

- The training technique then plays 25,000 games of Go between two copies of M_0 that have been enhanced with additional inference compute and an algorithm called Monte Carlo Tree Search to search through possible moves. Both players use Monte Carlo Tree Search, with M_0 guiding the search by estimating which moves are most promising and how strong each position is. By repeatedly calling M_0 in the search (thousands of times), we get a form of inference scaling which amplifies the power of this model. We could think of it as taking the raw System 1 intuitions of the base model and embedding them in a System 2 reasoning process which thinks many moves ahead.

- This amplified model is better than the base model at predicting the move most likely to win in each situation, but it is also much more costly. So, we train a new model, M_1, to predict the outputs of M_0 + search. Following Christiano, I shall call this step distillation, though in the case of AlphaGo Zero, M_1 was simply M_0 with an additional stage of training. This trained its move predictions to be closer to what the enhanced M_0 would choose and to make its position evaluations closer to the actual game outcomes. While M_1 will not be quite as good at Go as the amplified version of M_0, it is better than M_0 alone.

- But why stop there? We can repeat this process, amplifying M_1 through inference scaling by using it to guide the search process, producing a level of play beyond any seen so far (M_1 + search). This

[21] Paul Christiano, "Benign Model-Free RL," AI Alignment, June 2, 2017, https://ai-alignment.com/benign-model-free-rl-4aae8c97e385; Thomas Anthony et al., "Thinking Fast and Slow with Deep Learning and Tree Search," arXiv:1705.08439, December 3, 2017, https://doi.org/10.48550/arXiv.1705.08439; David Silver et al., "Mastering the Game of Go without Human Knowledge," Nature 550, no. 7676 (2017): 354–59, https://doi.org/10.1038/nature24270.

[22] For AlphaGo Zero, the goal was to start with zero information about Go and learn everything, so M_0 was simply a randomly initialised network. But it is also possible to start with a more advanced network as M_0, such as one trained to imitate human behaviour.

then gets distilled into a new model, M_2, and we proceed onwards and upwards, climbing higher and higher along the ladder of Go-playing performance (Figure 4).

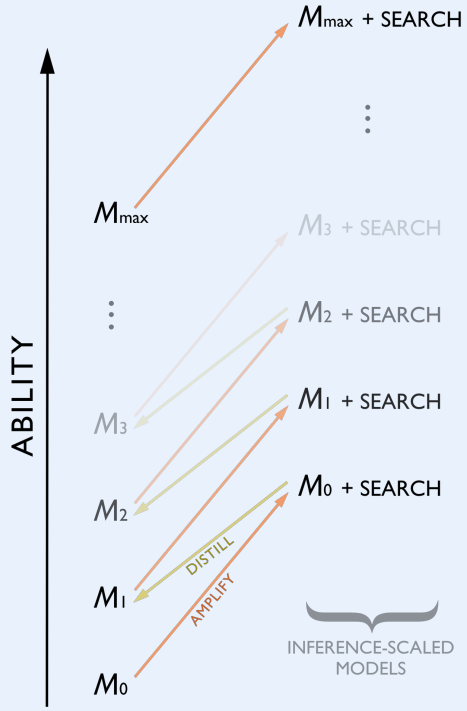


Figure 4. Iterated distillation and amplification to improve the performance of an inference-scaled AI model. (Source: author.)

- After just 36 hours, AlphaGo Zero had exceeded the ability of AlphaGo Lee, the version that beat world-champion Lee Sedol. Within 72 hours, it was beating AlphaGo Lee by 100 games to zero. And after 40 days of training (and 29 million games of self-play[23]), it reached its performance plateau, M_{max}, with an estimated Elo rating of 5,185 – far beyond the 3,739 of AlphaGo Lee or the low 3,000s of the world's best human players. Even when the final model was used without any search process (i.e. without any scaling of inference-at-deployment), it achieved a rating of 3,055, demonstrating professional-level play from pure "intuition".

It may be possible to use such a process of "iterated distillation and amplification" in training LLMs. The idea would be to take a model such as GPT-4o (which has powerful System 1 capabilities from pre-training) and use it as the starting model, M_0. Then, amplify it via inference scaling to simulate System 2–type reasoning before returning its final answer (as o1 and R1 do).[24] Then, distill this amplified model into a new model, M_1, that can produce answers

[23] Given 29 million games of self-play and a set-up with 25,000 games before each distillation, there were presumably 1160 iterations of amplification and distillation before it reached its plateau, such that M_{max} is M_{1160}.
[24] Like o1 and R1, we would presumably include additional RL post-training to prepare it for use in inference scaling.

of the same quality without doing extensive reasoning.[25] If this works, you now have a model that is more capable than GPT-4o without using extra compute during deployment.

By iterating this process of amplification followed by distillation, it may be possible for the LLM (just like AlphaGo Zero) to climb a very long way up this ladder before the process runs out of steam. And the time for each iteration may be substantially shorter than the time between major new pre-training runs. Like AlphaGo Zero, the final distilled model could display very advanced capabilities even without amplification. If this all worked, it would be a way of scaling inference-during-training to substantially quicken the rate of AI progress.

It is not at all clear whether this *will* work. The distillation process may plateau quickly, require increasingly large models at each step, take too long per iteration or too many iterations, or require years' worth of engineering effort to overcome the inevitable obstacles that will arise.[26] AlphaGo Zero provides a proof of concept, showing how a small team at a leading lab can achieve take-off with such a process and reach capabilities far beyond the former state of the art. However, the fact that we have so far not seen labs successfully use this method for LLMs should give pause regarding its usefulness.

So iterated distillation and amplification provides a plausible pathway for scaling inference-during-training to rapidly create much more powerful AI systems. Arguably, this would constitute a form of recursive self-improvement where AI systems are applied to the task of improving their own capabilities, leading to rapid escalation. While there have been earlier examples of this, they have often been on narrow domains (e.g. the game of Go) or have only applied to certain cognitive abilities (e.g. learning how to learn) and have therefore been bottlenecked on other abilities. An LLM scaled up with iterated distillation and amplification of LLMs could credibly learn to improve its own general intelligence.

Reduced governance transparency

What does this mean for AI governance? A key implication is that scaling inference-during-training could reduce transparency about the best current models. While this use of inference during the training process would reach the EU AI Act's compute threshold – both because inference-during-training counts as training compute and because it pushes the total compute over the limit – that threshold only requires oversight when the model is placed on the EU market.[27]

[25] Here M_1 could be a fresh model distilled from the inference-scaled M_0, or it could be M_0 with fine-tuning to make it behave more like the inference-scaled M_0.
[26] It is also possible that it will work in some domains (such as mathematics and coding) but not others, leading to superhuman capabilities in several new domains, but not across the board.
[27] And only when deployed inside the EU itself, where OpenAI's inference-scaled model deep research is conspicuously absent.

This means it may be possible for companies to substantially increase the intelligence of their leading models without anyone outside the organisation knowing. AI governance may then have to proceed with greater uncertainty about the state of the art. Relatedly, the lack of transparency would mean the public and policymakers would not be able to try these state-of-the-art models, making it harder for the range of publicly acceptable policies to shift in response. There would be less regulatory attention and a more abrupt shock to the world when the models at the top of the training ladder are deployed. The need to address transparency concerns raised by inference scaling may lend credence to pursuing an entity-based approach for AI governance.[28]

Shortened timelines to AGI

But perhaps most importantly, the possibility of training general models via iterated distillation and amplification could accelerate progress towards artificial general intelligence (AGI) systems with transformative global impacts. If this was combined with a lack of transparency about state-of-the-art models during internal scaling, policymakers could not know for sure whether progress was accelerating or not, making it hard to know whether emergency measures were required. All of this suggests that requiring companies to disclose the current capabilities of their systems – and their plans to improve them in the near-term – would be very valuable.

Conclusions

The shift from scaling pre-training compute to scaling inference compute may have substantial implications for AI governance.

If much of the remaining scaling comes from scaling inference-at-deployment, this could:

- Reduce the number of simultaneously served copies of each new model
- Increase the cost of first human-level AI systems
- Somewhat reduce the value of securing model weights
- Somewhat reduce the benefits and risks of open-weight models
- Allow unequal performance for different tasks and for different users
- Change the business model and industry structure
- Reduce the need for monolithic data centres

[28] Dean W. Ball and Ketan Ramakrishnan, Entity-Based Regulation in Frontier AI Governance (Carnegie Endowment for International Peace, 2025), https://carnegieendowment.org/research/2025/06/artificial-intelligence-regulation-united-states.

- Complicate AI governance via compute thresholds

If companies instead focus on using inference compute during training, then they may be able to use reasoning systems to create the high-quality training data needed to allow further gains from scaling pre-training. Inference-during-training could even accelerate scaling if companies use it to push their models up the ladder of distillation and amplification, as Google DeepMind did to create AlphaGo Zero. This possibility may lead to:

- Less transparency about state-of-the-art models
- Shorter timelines to transformative AGI

Either way, the shift to inference scaling also makes the future of AI less predictable than it was during the era of pre-training scaling. There is now more uncertainty about how quickly capabilities will improve and which longstanding features of the frontier AI landscape will persist. This uncertainty will make planning for the next few years more difficult for the frontier labs, investors, and policymakers. And it may place a premium on agility: the ability to first spot what is happening and pivot in response.

This analysis should be taken as a starting point for understanding the effects of inference scaling on AI governance. As this transition continues, it will be important for the field to track where inference compute is being employed and thus better understand which of these issues we are facing.

About the Authors

Toby Ord
Oxford Martin AI Governance Initiative, University of Oxford

Toby Ord is a Senior Researcher at Oxford University's AI Governance Initiative and research affiliate at Forethought. His work focuses on the big picture questions facing humanity. Toby has advised the United Nations, the World Health Organization, the World Economic Forum, and the UK Prime Minister's Office.

Acknowledgements

Special thanks to Seb Krier for discussions that inspired some of these ideas.

About GovAI

The Centre for the Governance of AI (GovAI) is a nonprofit based in London, UK. It was founded in 2018 at the University of Oxford, before becoming an independent research organization in 2021. GovAI's mission is to help decision-makers navigate the transition to a world with advanced AI, by producing rigorous research and fostering talent. The central focus of our research is threats that general-purpose AI systems may pose to security. We seek to understand the risks they pose today, while also looking ahead to the more extreme risks they could pose in the future.

Appendix | Comparing the costs of scaling pre-training vs inference-at-deployment

Scaling up pre-training by an order of magnitude and scaling up inference-at-deployment by an order of magnitude may have similar effects on the capabilities of a model, but they can have quite different effects on the total compute cost of the project. Which one is more expensive depends on the circumstances in a rather complex way.

Let's focus on the total amount of compute used for an AI system over its lifetime as the cost of that system (though this is not the only thing one might care about). The total amount of compute used for an AI system is equal to the amount used in training plus the amount used in deployment:

$$C = C_{\text{pre-training}} + C_{\text{post-training}} + C_{\text{deployment}}$$

Let N be the number of parameters in the model, D be the number of data tokens it is trained on, d be the number of times the model is deployed (e.g. the number of questions it is asked) and I be the number of inference steps each time it is deployed (e.g. the number of tokens per answer). Then this approximately works out to:

$$C \approx ND + C_{\text{post-training}} + dNI$$

Note that scaling up the number of parameters, N, increases both pre-training compute and inference compute, because you need to use those parameters each time you run a forward pass in your model. But scaling up D does not directly affect deployment costs. Some typical rough numbers for these variables in GPT-4-level LLMs are:

LLMs are:

$$N = 10^{12},\ D = 10^{13},\ I = 10^{3},\ d = ?$$

On this rough arithmetic, the deployment costs overtake the pre-training costs when the total number of tokens generated in deployment (dI) is greater than the total number of training tokens D. That would require $d > 10^{10}$. Apparently, this is usually the case, with deployment compute exceeding total training compute on commercial frontier systems.

The most standard way of training LLMs while minimising training compute involves scaling up N and D by the same factor. For example, if you scale up training compute by 1 OOM, that means 0.5 OOMs more parameters and 0.5 OOMs more data. So, scaling up training compute by 1 OOM also increases deployment compute by 0.5 OOMs. In contrast, scaling up

inference-at-deployment by an order of magnitude does not (directly) affect pre-training compute.

When either the pre-training compute (ND) or the deployment compute (dNI) is the bulk of the total (including $C_{post\text{-}training}$), there are some simple approximations for the costs of scaling. If $C_{pre\text{-}training} \gg C_{post\text{-}training} + C_{deployment}$, then scaling pre-training by 10x increases costs by nearly 10x, while scaling inference-at-deployment (I) by 10x does not affect the total much. Whereas if $C_{deployment} \gg C_{pre\text{-}training} + C_{post\text{-}training}$, then scaling pre-training by 10x increases costs by ~3x (from the larger number of parameters needed at deployment), while scaling inference-at-deployment by 10x increases costs by nearly 10x. So, there is some incentive to balance these numbers where possible.

It is important to note that the costs of scaling inference-at-deployment depend heavily on how much deployment you are doing. If you just use the model to answer a single question, then you could scale it all the way until it generates as many tokens as you pre-trained on (i.e. trillions) before it appreciably affects your overall compute budget. But if you are scaling up the inference used for every question, your overall compute budget could be affected even by a 2x scale-up.

References

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

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

Anthony, Thomas, Zheng Tian, and David Barber. "Thinking Fast and Slow with Deep Learning and Tree Search." arXiv:1705.08439. Preprint, arXiv, December 3, 2017. https://doi.org/10.48550/arXiv.1705.08439.

"Article 51: Classification of General-Purpose AI Models as General-Purpose AI Models with Systemic Risk." EU Artificial Intelligence Act. August 2, 2025. https://artificialintelligenceact.eu/article/51.

Ball, Dean W., and Ketan Ramakrishnan. Entity-Based Regulation in Frontier AI Governance. Carnegie Endowment for International Peace, 2025. https://carnegieendowment.org/research/2025/06/artificial-intelligence-regulation-united-states.

Christiano, Paul. "Benign Model-Free RL." AI Alignment, June 2, 2017. https://ai-alignment.com/benign-model-free-rl-4aae8c97e385.

EU AI Act: GPAI Code of Practice. "EU AI Act: General-Purpose AI Code of Practice." 2025. https://code-of-practice.ai.

Heim, Lennart, and Leonie Koessler. "Training Compute Thresholds: Features and Functions in AI Regulation." arXiv:2405.10799. Preprint, arXiv, August 6, 2024. https://doi.org/10.48550/arXiv.2405.10799.

Heim, Lennart, and Leonie Koessler. "Training Compute Thresholds: Features and Functions in AI Regulation." arXiv:2405.10799. Preprint, arXiv, August 6, 2024. https://doi.org/10.48550/arXiv.2405.10799.

Hooker, Sara. "On the Limitations of Compute Thresholds as a Governance Strategy." arXiv:2407.05694. Preprint, arXiv, July 30, 2024. https://doi.org/10.48550/arXiv.2407.05694.

"Kadrey v. Meta, Document 391, Exhibit K, Vo Declaration." January 14, 2025. https://storage.courtlistener.com/recap/gov.uscourts.cand.415175/gov.uscourts.cand.415175.391.24.pdf.

Lambert, Nathan. "GPT-4.5: 'Not a Frontier Model'?" Interconnects, November 24, 2023. https://www.interconnects.ai/p/gpt-45-not-a-frontier-model.

Lambert, Nathan. "xAI's Grok 4: The Tension of Frontier Performance with a Side of Elon Favoritism." Interconnects, July 12, 2025. https://www.interconnects.ai/p/grok-4-an-o3-look-alike-in-search.

OpenAI. "Learning to Reason with LLMs." September 12, 2024. https://openai.com/index/learning-to-reason-with-llms.

Ord, Toby. "The Extreme Inefficiency of RL for Frontier Models." September 19, 2025. https://www.tobyord.com/writing/inefficiency-of-reinforcement-learning.

Palazzolo, Stephanie, Erin Woo, and Amir Efrati. "OpenAI Shifts Strategy as Rate of 'GPT' AI Improvements Slows." The Information, November 9, 2024. https://www.theinformation.com/articles/openai-shifts-strategy-as-rate-of-gpt-ai-improvements-slows.

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

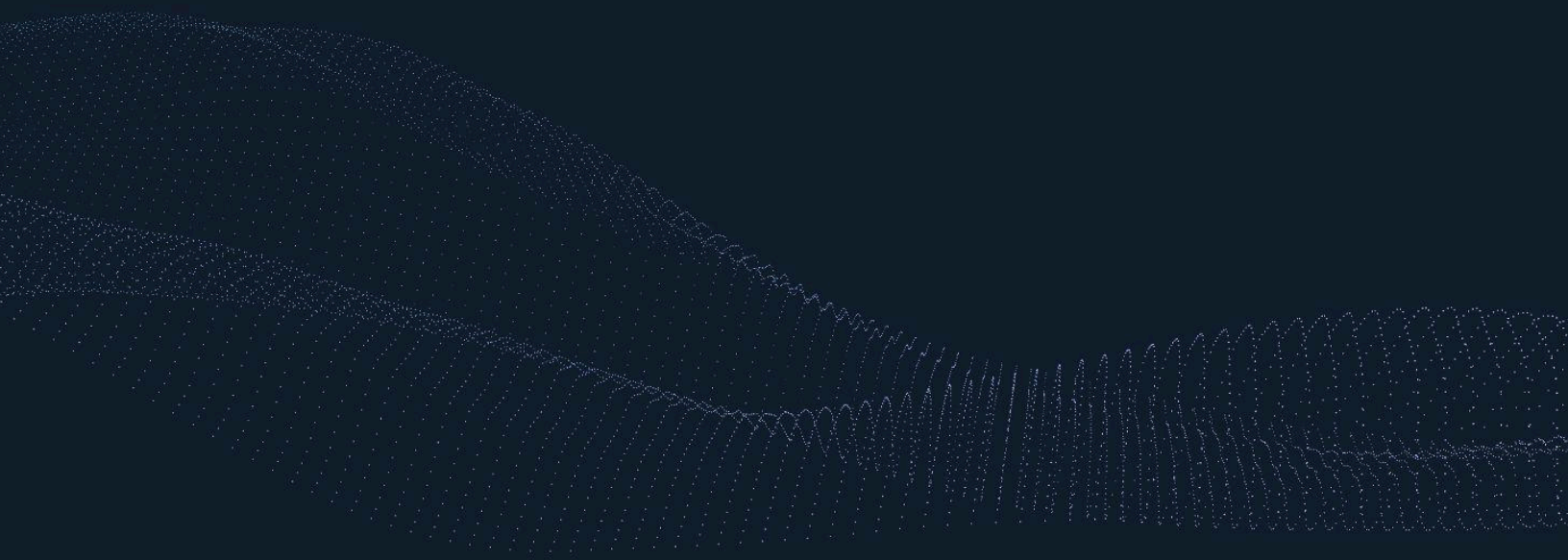
Sevilla, Jaime, and Edu Roldán. "Training Compute of Frontier AI Models Grows by 4-5x per Year." Epoch AI, 2024. https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year.

Silver, David, Julian Schrittwieser, Karen Simonyan, et al. "Mastering the Game of Go without Human Knowledge." Nature 550, no. 7676 (2017): 354–59. https://doi.org/10.1038/nature24270.

Villalobos, Pablo, and David Atkinson. Trading off Compute in Training and Inference. Epoch AI, 2023. https://epoch.ai/blog/trading-off-compute-in-training-and-inference.

"What Authors Need to Know about the $1.5 Billion Anthropic Settlement." The Authors Guild, October 2, 2025. https://authorsguild.org/advocacy/artificial-intelligence/what-authors-need-to-know-about-the-anthropic-settlement.

Zeff, Maxwell. "Current AI Scaling Laws Are Showing Diminishing Returns, Forcing AI Labs to Change Course." Techcrunch, November 20, 2024. https://techcrunch.com/2024/11/20/ai-scaling-laws-are-showing-diminishing-returns-forcing-ai-labs-to-change-course.



GovAI