Policy Primer

Cohere For AI

The Limits of Thresholds

Exploring the Role of Compute-Based Thresholds for Governing the Risks of AI Models

July 2024

Summary

Prominent AI governance frameworks around the world have specified thresholds based on the amount of computing power used to train an AI model, measured in floating-point operations (FLOPs). Models that exceed these thresholds are assumed to pose a level of risk that requires additional reporting and scrutiny.

However, the appropriateness of these approaches is under debate among scientific communities, in response to growing evidence that **increased training compute does not necessarily equate to increased risk**. This is due to several factors:

1. Model capabilities and performance are affected by factors beyond just training compute, including data quality, downstream model optimization techniques, and algorithmic architectures.
2. The risks associated with AI models are affected by factors that are not accounted for in training compute measures, such as characteristics of the datasets used in model training, deployment context, and safety optimization.

Further complexities add to the limitations of compute-based thresholds, including technical uncertainties about how FLOPs should be calculated, and the fact that current training compute thresholds are not likely to be met by any existing models, meaning that immediate and near-term risks may be overlooked.

The limitations of compute-based thresholds may have the following consequences that hinder the ultimate goal of managing AI risk (in no particular order, and not an exhaustive list):

- Incentives could be created for model developers to game compute thresholds rather than meaningfully address risks.
- Regulatory scrutiny could be applied disproportionately to models over the threshold that may not actually pose greater risk than models under the threshold.
- Resources and capacity of those working on AI safety could be diverted away from near-term, real-world risks.

To address these limitations, policymakers could consider: **adopting alternative or complementary approaches** to assessing which AI models should face greater scrutiny, developing **dynamic rather than static thresholds**, and **more clearly defining approaches to calculating and measuring FLOPs.**

This policy primer provides an overview of evidence about the limitations of compute-based thresholds, to support policymakers implementing risk-based governance of AI models. The primer references technical concepts that are explored in deeper detail in an essay by the Head of Cohere For AI, Sara Hooker.[1]

[1] Hooker, S. (2024) 'On the Limitations of Compute Thresholds as a Governance Strategy'. arXiv. http://arxiv.org/abs/2407.05694.

CONTENTS

1. Background: The Role of Compute Thresholds in AI Governance

As policymakers around the world work to develop appropriate governance frameworks for AI, one of the many challenges lies in balancing the need to assess the risk of AI models and systems with the required regulatory resources and developer capacity to undertake assessment processes. If emerging governance frameworks get that balance wrong, there's a chance that either the risks of AI models are insufficiently addressed, or so many AI models require a high-level of scrutiny that regulators become overwhelmed and AI developers do not have room to innovate. To strike the right balance, the regulatory scrutiny applied to AI models or systems must be proportionate to the risks associated with them.

In attempts to achieve this balance, both the White House Executive Order on AI Safety[2] and the EU AI Act[3] — two of the first policy instruments enacted to govern AI — detail thresholds that determine which AI models will be subject to higher levels of scrutiny. These thresholds are based on the amount of computing power (compute) required to train an AI model, measured in the number of integer or floating point operations (FLOPs).[4] The White House Executive order sets this as 10^{26} FLOPs or 10^{23} FLOPs for models using biological sequence data, while the EU AI Act specifies 10^{25} FLOPs for any general purpose AI (GPAI) model.[5] If an AI model exceeds these

[2] The White House (2023) *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, *The White House*. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.
[3] European Commission (2024). *The AI Act.* https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf.
[4] Floating point operations (FLOPs) refers to the mathematical operations a computer performs. The term arose in the 1950s and 60s to provide a unit of measure for computer processing speed and power.
[5] The EU AI Act defines GPAI as follows: "A GPAI model means an AI model, including when trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks, regardless of the way the model is placed on the market, and that can be integrated into a variety of downstream

thresholds, it will be subject to a higher level of scrutiny and regulatory obligation under these governance frameworks.

There are several reasons why policymakers setting thresholds to trigger extra scrutiny may have looked to the amount of training compute.[6] Training compute offers a quantifiable and externally verifiable metric that can be applied universally across different systems, it is a measure commonly used in the field of machine learning and computer science, enabling developers and researchers to assess hardware needs or compare models, often early in the AI model development lifecycle.[7, 8] These advantages of compute-based thresholds make them an appealing tool for policymakers seeking to manage the risks associated with AI systems.

However, compute-based thresholds are not without limitations. The following sections of this primer outline these limitations and the evidence which underpins them, in an effort to support policymakers develop and implement governance frameworks that effectively manage AI risk.

systems or applications. This does not cover AI models that are used before release on the market for research, development, and prototyping activities." https://artificialintelligenceact.eu/high-level-summary/.

[6] In addition to compute-based thresholds, other governance and policy documents from around the world reference thresholds as a key component in AI risk management. For example, the Seoul Frontier AI Safety Commitments require the signatories to "set out thresholds at which severe risks posed by a model or system, unless adequately mitigated, would be deemed intolerable." https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024. This scope of this policy primer, however, focuses on compute-based thresholds, and does not extend to explicitly address other types of risk threshold.

[7] Hooker, S. (2024) 'On the Limitations of Compute Thresholds as a Governance Strategy'. arXiv. http://arxiv.org/abs/2407.05694.

[8] Lennart Heim (2024) (Training) Compute Thresholds - Features and Functions in AI Governance. https://heim.xyz/documents/Training-Compute-Thresholds.pdf.

2. Limitations of Compute Thresholds

In risk-based approaches to technology governance, thresholds are a useful tool for setting the boundaries of risk tolerance and triggering appropriate and proportionate safeguards or mitigations. Such thresholds are applied with success in many techno-scientific fields, ranging from healthcare, such as infant low birth-weight thresholds,[9] to aviation, such as thresholds for the risk of collision between aircraft and obstacles near airports.[10] In these cases, there are clear, well-evidenced, and direct or linear relationships between the threshold measure and the risk of consequent harm.

The justification for current compute-based thresholds in AI is that they approximate the amount of compute used in training the most advanced models to date, and as such "the capabilities of the models above this threshold are not yet well enough understood [and] could pose systemic risks."[11] In other words, the rationale underpinning compute-based thresholds is that the amount of training compute serves as a proxy for model capabilities, which in turn indicate risk.

However, increasing evidence suggests that **the relationship between training compute and risk from AI models is not direct, and several limitations of compute-based thresholds lie within the uncertain and complex relationship between training compute, model capability, and risk.**[12]

The following section details some of these limitations and the evidence underpinning our understanding of them. This evidence relates to two crucial relationships that require greater scrutiny: first that training compute equals model capability, and second that model capability is the primary determinant of model risk.

2.1 Compute =/= Capability: Factors Beyond Compute Affect Model Capability

The notion that greater training compute results in greater model capability and performance is largely a product of the fact that this has been true for many recent developments in machine learning. When graphics processing units (GPUs) were repurposed for use in training deep neural networks, they led to significant advances. This was largely due to GPUs' ability to process more calculations per second than traditional CPUs (central processing units), and that they could be combined together to increase performance even further.[13, 14] The significance of this development was demonstrated when Google used 16,000 CPU cores in 2012 to classify

[9] The World Health Organization (no date) *Low birth weight*. https://www.who.int/data/nutrition/nlis/info/low-birth-weight.
[10] Moretti, L., Dinu, R. and Di Mascio, P. (2023) 'Collision risk assessment between aircraft and obstacles in the areas surrounding airports', *Heliyon*, 9(7), p. e18378. https://doi.org/10.1016/j.heliyon.2023.e18378.
[11] See: European Commission (2023) *Artificial Intelligence – Q&As*. https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683.
[12] Hooker, S. (2024) 'On the Limitations of Compute Thresholds as a Governance Strategy'. arXiv.http://arxiv.org/abs/2407.05694.
[13] Center for Security and Emerging Technology, Khan, S. and Mann, A. (2020) *AI Chips: What They Are and Why They Matter*. Center for Security and Emerging Technology. https://doi.org/10.51593/20190014.
[14] Hooker, S. (2020) The Hardware Lottery, arXiv. http://arxiv.org/abs/2009.06489.

cats in images, only for a paper to be published a year later that solved the same task with just two CPU cores and four GPUs.[15, 16] Since then, the use of GPUs has become mainstream in machine learning, and the basis for many advances on AI model capabilities has been associated with increasing, or "scaling", the volume of compute used to train AI models.[17]

However, the amount of training compute is not the only variable that determines an AI model's capabilities. Increasingly, optimization and algorithmic developments determine overall performance,[18] meaning there are many variables that affect and influence an AI model's capabilities, including:

- **Data quality**: A large body of research shows that improving the quality of data used to train an AI model — through methods such as de-duplicating[19, 20] or pruning[21, 22] — can increase model capability and performance without needing to increase the amount of compute.
- **Model optimization**: Recent progress in AI models has largely resulted from methods to improve pre-trained models through techniques such as instruction fine-tuning,[23, 24, 25] model distillation using synthetic data from more performant models,[26, 27] chain-of-thought reasoning,[28, 29] increased context-length,[30] enabled tool-use,[31, 32] retrieval

[15] Le, Q.V. *et al.* (2012) Building high-level features using large scale unsupervised learning, arXiv. Available at: https://doi.org/10.48550/arXiv.1112.6209.
[16] Coates, A. *et al.* (2013) 'Deep learning with COTS HPC systems', in *Proceedings of the 30th International Conference on Machine Learning*. *International Conference on Machine Learning*, PMLR, pp. 1337–1345. https://proceedings.mlr.press/v28/coates13.html.
[17] Amodei, D. and Hernandez, D. (2018) AI and compute. https://openai.com/index/ai-and-compute/.
[18] Hooker, S. (2024) 'On the Limitations of Compute Thresholds as a Governance Strategy'. arXiv.http://arxiv.org/abs/2407.05694.
[19] Hernandez, D. *et al.* (2022) 'Scaling Laws and Interpretability of Learning from Repeated Data'. arXiv. https://doi.org/10.48550/arXiv.2205.10487.
[20] Lee, K. *et al.* (2022) 'Deduplicating Training Data Makes Language Models Better'. arXiv. https://doi.org/10.48550/arXiv.2107.06499.
[21] Sorscher, B. *et al.* (2023) 'Beyond neural scaling laws: beating power law scaling via data pruning'. arXiv. http://arxiv.org/abs/2206.14486.
[22] Marion, M. *et al.* (2023) 'When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale'. arXiv. http://arxiv.org/abs/2309.04564.
[23] Mishra, S. *et al.* (2022) 'Cross-Task Generalization via Natural Language Crowdsourcing Instructions', in S. Muresan, P. Nakov, and A. Villavicencio (eds) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. *ACL 2022*, Dublin, Ireland: Association for Computational Linguistics, pp. 3470–3487. https://doi.org/10.18653/v1/2022.acl-long.244.
[24] Honovich, O. *et al.* (2023) 'Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor', in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. *ACL 2023*, Toronto, Canada: Association for Computational Linguistics, pp. 14409–14428. https://doi.org/10.18653/v1/2023.acl-long.806.
[25] Singh, S. *et al.* (2024) 'Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning'. arXiv. http://arxiv.org/abs/2402.06619.
[26] Gemma Team *et al.* (2024) 'Gemma: Open Models Based on Gemini Research and Technology'. arXiv. https://doi.org/10.48550/arXiv.2403.08295.
[27] Aryabumi, V. *et al.* (2024) 'Aya 23: Open Weight Releases to Further Multilingual Progress'. arXiv. https://doi.org/10.48550/arXiv.2405.15032.
[28] Wei, J. *et al.* (2023) 'Chain-of-Thought Prompting Elicits Reasoning in Large Language Models'. arXiv. https://doi.org/10.48550/arXiv.2201.11903.
[29] Hsieh, C.-Y. et al. (2023) 'Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes'. arXiv. https://doi.org/10.48550/arXiv.2305.02301.
[30] Xiong, W. et al. (2023) 'Effective Long-Context Scaling of Foundation Models'. arXiv. https://doi.org/10.48550/arXiv.2309.16039.
[31] Qin, Y. et al. (2023) 'ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs'. arXiv. https://doi.org/10.48550/arXiv.2307.16789.
[32] Wang, G. *et al.* (2023) 'Voyager: An Open-Ended Embodied Agent with Large Language Models'. arXiv. https://doi.org/10.48550/arXiv.2305.16291.

augmented generation,[33, 34] or reinforcement learning from human feedback.[35, 36, 37] These methods can yield improvements in model capability without necessarily requiring a concurrent increase in the amount of compute used to train a model.

- **Novel model architectures:** Major advances in model capabilities have been the product of advances in the underpinning algorithmic structure of an AI model, and not always reliant on increases in training compute. For example, the invention of convolutional neural networks in 2012 was fundamental to advances machine image recognition,[38, 39] as was the invention of transformer architecture, which led to the explosion in large language models in 2017.[40]

In addition to other variables that affect model performance, some research has begun to suggest that **increasing compute does not always lead to increasing model capability**. Studies have shown a level of redundancy in model weights: a significant proportion could be removed without affecting performance,[41] and a small subset of weights can accurately predict almost the full network.[42] Additionally, research has found instances where as training compute increases, model performance actually *decreases*.[43] Such findings further complicate understanding of the relationship between compute and model performance.

The notion that greater compute does not necessarily yield greater capability is demonstrated in the performance rates of recent models relative to their size, with an increasing number of smaller models matching or outperforming larger ones. For example, Falcon 180B is outperformed by far smaller open weights models such as Llama-3 8B, Command R+ 35B, and Gemma 27B; and Aya 23 8B outperforms the much larger BLOOM 176B.[44] These examples indicate an overall trend where larger models are not guaranteed to consistently outperform smaller models (see Figure 1. below).

[33] Lewis, P. *et al.* (2021) 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks'. arXiv. https://doi.org/10.48550/arXiv.2005.11401.
[34] Pozzobon, L. *et al.* (2023) 'Goodtriever: Adaptive Toxicity Mitigation with Retrieval-augmented Models', in H. Bouamor, J. Pino, and K. Bali (eds) *Findings of the Association for Computational Linguistics: EMNLP 2023*. *Findings 2023*, Singapore: Association for Computational Linguistics, pp. 5108–5125. https://doi.org/10.18653/v1/2023.findings-emnlp.339.
[35] Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs, https://arxiv.org/abs/2402.14740.
[36] Aakanksha *et al.* (2024) 'The Multilingual Alignment Prism: Aligning Global and Local Preferences to Reduce Harm'. arXiv. http://arxiv.org/abs/2406.18682.
[37] Dang, J. *et al.* (2024) 'RLHF Can Speak Many Languages: Unlocking Multilingual Preference Optimization for LLMs'. arXiv. https://doi.org/10.48550/arXiv.2407.02552.
[38] Valueva, M.V. *et al.* (2020) 'Application of the residue number system to reduce hardware costs of the convolutional neural network implementation', *Mathematics and Computers in Simulation*, 177, pp. 232–243. https://doi.org/10.1016/j.matcom.2020.04.031.
[39] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) 'ImageNet Classification with Deep Convolutional Neural Networks', in *Advances in Neural Information Processing Systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.
[40] Vaswani, A. *et al.* (2023) 'Attention Is All You Need'. arXiv. https://doi.org/10.48550/arXiv.1706.03762.
[41] Gale, T., Elsen, E. and Hooker, S. (2019) 'The State of Sparsity in Deep Neural Networks'. arXiv. http://arxiv.org/abs/1902.09574.
[42] Denil, M. *et al.* (2014) 'Predicting Parameters in Deep Learning'. arXiv. https://doi.org/10.48550/arXiv.1306.0543.
[43] McKenzie, I.R. *et al.* (2024) 'Inverse Scaling: When Bigger Isn't Better'. arXiv. https://doi.org/10.48550/arXiv.2306.09479.
[44] Hooker, S. (2024) 'On the Limitations of Compute Thresholds as a Governance Strategy'. arXiv.http://arxiv.org/abs/2407.05694.

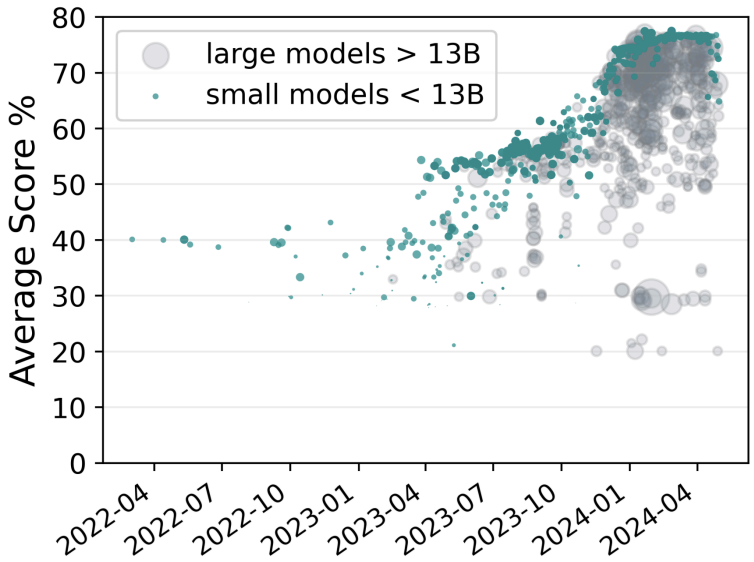

Figure 1. Models submitted to the OpenLLM leaderboard since April 2022. Smaller models (fewer than 13bn parameters) consistently match or outperform larger models.[45]

In sum, increasing the amount of compute to train an AI model has historically yielded advances in model capability, but evidence that it is the only or primary determinant of model capability is not robust, and an emerging body of research suggests that increasing compute *does not always* increase capability.

This means that in both the near- and long-term increasing numbers of models that fall well below a training compute-threshold may be more capable and performant than models above those thresholds, yet face less regulatory oversight. This may even create incentives for model developers to use techniques to increase performance without increasing compute, in order to remain below thresholds.

2.2 Compute =/= Risk: Factors Beyond Capability Affect AI Model Risk

The second assumed relationship underpinning the rationale for compute-based thresholds for risk is that as a model's capability increases, so do the associated risks. This is based on the notion that as AI models become more capable — i.e., able to perform a wider range of complex tasks or actions and with increasing success, scale, and speed — the risks of harm as a consequence of these capabilities amplifies.[46, 47]

[45] Hooker, S. (2024) p6.
[46] Sastry, G. et al. (2024) *Computing Power and the Governance of Artificial Intelligence*. Centre for the Governance of AI. https://cdn.governance.ai/Computing_Power_and_the_Governance_of_AI.pdf.
[47] Shevlane, T. *et al.* (2023) *Model evaluation for extreme risks*. Google DeepMind. http://arxiv.org/abs/2305.15324.

However, even if the amount of training compute were a reliable and dominant variable for predicting model *performance*, **there are several additional factors to understanding model *risk* that training compute and capability alone do not capture**:

- **Dataset composition**: AI models are trained on vast quantities of data to learn and understand patterns in that data that they can replicate in their behavior and outputs. This means that any risks inherent to the dataset — such as the presence of toxic language, hate speech, societal biases, stereotypical representations, or information that is sensitive in terms of safety to individuals, communities, organizations, or nations — may become embedded in the model and affect its behavior.[48, 49] These dataset risks are independent of the amount of compute and model capability; a model trained on harmful data can exhibit harmful outputs with comparatively low levels of compute or performance, including existing models that likely fall below the thresholds in either the White House Executive Order or the EU AI Act.[50]
- **Deployment context:** The risks of AI models manifest not only through the development of the model itself, but also through the application of that model into products and environments where real-world harms can be caused by errors or undesirable model behavior.[51] For example, the risks presented by a large language model used to search documents in a historical database might be very different to the risks of a similar model used to automate the summarisation of hospital patients' clinical notes. Moreover, these risks may occur regardless of the amount of compute used to train the model: a "low-compute" model can still cause harm if deployed irresponsibly. Compute-based thresholds for determining AI risk fail to account for this, and as such, they are not dissimilar to assessing the risks of electricity by measuring wattage — the risks associated with a certain amount of wattage vary depending on whether it is applied to a refrigerator, a lightbulb, or a dangling live wire. Those risks are further dependent on the many infrastructural components of these contexts, such as the quality of wiring, in-built safety features like circuit breakers, and so on. Indeed, many emerging regulatory frameworks — including the EU AI Act — recognise how AI systems may be deemed high-risk due to their deployment contexts, regardless of the amount of training-compute associated with them.
- **Optimizing models for safety:** Across different stages of model training development, various technical methods can be applied to address potential risks. There are many examples of such methods, including safety context distillation,[52, 53] training on values-targeted datasets,[54] reinforcement learning from human feedback,[55] or

[48] Suresh, H. and Guttag, J. (2021) 'A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle', in *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. New York, NY, USA: Association for Computing Machinery (EAAMO '21), pp. 1–9. https://doi.org/10.1145/3465416.3483305.
[49] Pozzobon, L. *et al.* (2024) 'From One to Many: Expanding the Scope of Toxicity Mitigation in Language Models'. arXiv. http://arxiv.org/abs/2403.03893.
[50] Solaiman, I. et al. (2023) 'Evaluating the Social Impact of Generative AI Systems in Systems and Society'. https://doi.org/10.48550/arXiv.2306.05949.
[51] Solaiman, I. et al. (2023) 'Evaluating the Social Impact of Generative AI Systems in Systems and Society'. arXiv. https://doi.org/10.48550/arXiv.2306.05949.
[52] Amanda Askell, et al. (2021) A general language assistant as a laboratory for alignment. https://arxiv.org/abs/2112.0.0861.
[53] Deep Ganguli et al (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. https://arxiv.org/abs/2209.07858.
[54] Solaiman, I. and Dennison, C. (2021) 'Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets'. https://cdn.openai.com/palms.pdf.
[55] A. Glaese, et al, (2022) "Improving alignment of dialogue agents via targeted human judgements" ; https://doi.org/10.48550/arXiv.2209.14375.

"un-learning" harmful behaviors.[56] These methods are part of the process of optimizing model performance for certain desired qualities — safety or otherwise — and they are a key component of developing capable and performant AI models across industry, academia, and beyond.

Compute-based thresholds by themselves are not able to account for the factors outlined above, meaning they fail to consider a full and comprehensive picture of risks associated with an AI model. This means that where compute-based thresholds are used they would benefit from being complemented with additional measures which can assess factors including dataset composition, deployment context, and safety optimization — and likely many other non-compute related factors not considered here; the above is not necessarily an exhaustive list.

2.3 Limitations of FLOPs as a Measure

The primary limitations of compute-based thresholds are associated with the unclear relationships between compute, capability, and risk, as outlined above. However, it is important to note that FLOPs as a measure also introduce several limitations, many of which are associated with technical questions associated with how they are calculated.

For example, **certain model structures complicate how FLOPs are calculated.**[57] For example, "mixtures of experts" combine independently trained models and enable inputs to be routed according to the characteristics of each "expert" model.[58] Similarly, cascading model systems that are common in existing AI applications combine models such that the output of one becomes the input of another.[59] Summing the total FLOPs used to train all component models in a mixture or cascade could lead to disproportionately high measures relative to the risk posed by an individual model or system, while counting the compute associated with each expert or model individually may fail to account for the overall risk posed by a mixture of experts or cascading system. This is not an insurmountable issue, but instead points to the need for much greater specificity in governance frameworks around how FLOPs should be calculated.[60]

It is also not clear how FLOP measures work with decentralized training and other downstream optimization. For example, where machine learning models are developed by multiple parties — such as finetuning an open source model — it is challenging to accurately measure the FLOPs across the entire life-cycle. This is especially true for models trained in a decentralized, privacy-preserving manner, where training hardware belongs to multiple participants. Additionally, many recent efforts show that performance gains can come from applying compute outside of model training, such as through best-of-n sampling techniques,[61]

[56] N. Li, A. Pan, A. Gopal, S., D. Hendrycks, et al. (2024) *The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning*, arXiv:2403.03218, https://doi.org/10.48550/arXiv.2403.03218.
[57] Hooker, S. (2024) 'On the Limitations of Compute Thresholds as a Governance Strategy'. arXiv. http://arxiv.org/abs/2407.05694.
[58] Zadouri, T. *et al.* (2023) 'Pushing Mixture of Experts to the Limit: Extremely Parameter Efficient MoE for Instruction Tuning'. arXiv. https://doi.org/10.48550/arXiv.2309.05444.
[59] Paleyes, A., Urma, R.-G. and Lawrence, N.D. (2023) 'Challenges in Deploying Machine Learning: a Survey of Case Studies', *ACM Computing Surveys*, 55(6), pp. 1–29. https://doi.org/10.1145/3533378.
[60] Frontier Model Forum (2024) 'Issue Brief: Measuring Training Compute'. https://www.frontiermodelforum.org/updates/issue-brief-measuring-training-compute/.
[61] Gemini Team. (2024) 'Gemini: A Family of Highly Capable Multimodal Models'. arXiv. https://doi.org/10.48550/arXiv.2312.11805.

chain-of-thought reasoning,[62] and model distillation using synthetic data.[63, 64] It is not yet clear how such approaches, which may alter a model's risk profile, would be accounted for by measuring training FLOPs.

Additionally, **few existing models are known to meet or exceed currently set FLOPs thresholds** (10^{26} FLOPs or 10^{23} for models using biological sequence data in the White House Executive Order and 10^{25} in the EU AI Act). This is sensible if the goal is to set prospective thresholds to address potential future risks. However, this fails to recognise near-term, real-world risks associated with current models that fall below these thresholds, such as confabulations, mis- or disinformation, information harms, bias, and toxicity.[65, 66]

FLOP counts vary significantly across AI model modalities. For example, code models and biological models have typically used less compute than large language models.[67, 68] And multilingual models — language models that work across multiple natural languages — tend to be larger as they require more compute for each additional language covered.[69, 70] Compute-thresholds which fail to account for these variations in the compute needs of different modalities may result in disproportionate regulatory burden across different types of AI model. The graphs in Figure 2., below, show these variations between the compute needs of some of the largest language models compared to many other models of different modalities.

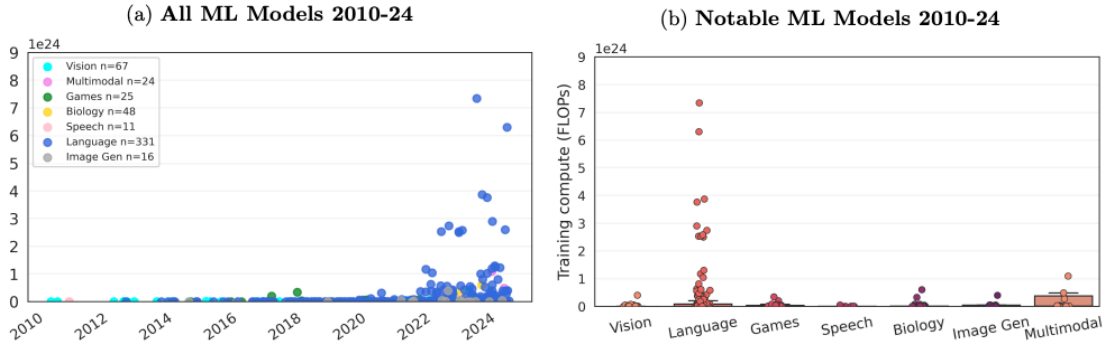

Figure 2. Two graphs showing that, of the notable AI models released between 2010 and 2024, language models have used considerably more training compute than any other modality.[71]

[62] Wei, J. *et al.* (2023) 'Chain-of-Thought Prompting Elicits Reasoning in Large Language Models'. arXiv. https://doi.org/10.48550/arXiv.2201.11903.
[63] Aryabumi, V. *et al.* (2024) 'Aya 23: Open Weight Releases to Further Multilingual Progress'. arXiv. https://doi.org/10.48550/arXiv.2405.15032.
[64] Shimabucoro, L. *et al.* (2024) 'LLM See, LLM Do: Guiding Data Generation to Target Non-Differentiable Objectives'. arXiv. https://doi.org/10.48550/arXiv.2407.01490.
[65] Privitera, D. *et al.* (2024) 'International Scientific Report on the Safety of Advanced AI - Interim Report'. https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai.
[66] Weidinger, L. *et al.* (2021) 'Ethical and social risks of harm from Language Models'. arXiv. http://arxiv.org/abs/2112.04359.
[67] Lin, J. *et al.* (2024) 'Scaling Laws Behind Code Understanding Model'. arXiv. http://arxiv.org/abs/2402.12813.
[68] Maug, N. (2024) *Biological Sequence Models in the Context of the AI Directives*, *Epoch AI*. https://epochai.org/blog/biological-sequence-models-in-the-context-of-the-ai-directives.
[69] Üstün, A. *et al.* (2024) 'Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model'. arXiv. http://arxiv.org/abs/2402.07827.
[70] Aryabumi, V. *et al.* (2024) 'Aya 23: Open Weight Releases to Further Multilingual Progress'. arXiv. https://doi.org/10.48550/arXiv.2405.15032.
[71] Hooker, S. (2024) 'On the Limitations of Compute Thresholds as a Governance Strategy'. arXiv. http://arxiv.org/abs/2407.05694. p12.

These considerations do not render FLOPs entirely invalid as a measure; indeed they remain commonly used within the field of computer science during model development and training to assess hardware needs or training timeframes. However, the ambiguities around FLOPs measures need to be resolved to ensure they are used in a fair, consistent, and effective way.[72] This could involve developing clear guidance and standards for how FLOPs should be calculated in practice.

[72] Frontier Model Forum (2024) 'Issue Brief: Measuring Training Compute'. https://www.frontiermodelforum.org/updates/issue-brief-measuring-training-compute/.

3. Conclusion: Consequences of Compute Thresholds

Compute-based thresholds offer a quantifiable and standardized metric that holds appeal for policymakers and regulators when assessing which AI models or systems pose greater risk. However, the relationship between training compute, model capability, and risk is complex and nonlinear. This not only undermines their effectiveness and suitability, but could lead to several undesirable consequences that weaken risk management and hinder innovation:

- **Gamification:** One potential consequence is that model developers are incentivised to discover ways to game the training compute threshold to avoid triggering additional regulatory burden. For example, keeping the volume of biological sequence data low to avoid classification as a biological model (and being subject to a lower compute threshold per the White House Executive Order), or throttling the amount of training compute where possible and using methods of model optimization to improve performance while keeping below thresholds.
- **Applying disproportionate scrutiny:** As smaller models continue to become more performant, models that could pose greater risk may face relatively less regulatory scrutiny because they fall underneath a static compute threshold. This means compute-based thresholds may result in false positives or negatives that incorrectly categorize models as higher or lower risk, leading to disproportionate allocation of regulatory and developer resources to models that involve large amounts of training compute, but may not actually pose greater risk.
- **Distracting regulatory and safety resources:** Compute thresholds that are based on the assumption that model capability is the primary determinant of risk ignore other key factors, such as training data, deployment context, and safety optimization practices. This means that rather than focusing on systemic risks such as bias, misinformation, or toxic representations, safety teams within organizations developing models may be forced to deal with arbitrary requirements that prioritize hypothetical and un-evidenced risks over more tangible ones. This misalignment of priorities and resources could hinder the progress of crucial science-backed and applied AI safety work.

These consequences do not mean that compute thresholds are entirely without merit, and nor do current approaches that rely on them fail to consider any limitations; indeed, many proponents recognize that compute-based thresholds are an imperfect proxy for risk. Nevertheless, the consideration of any threshold — proxy or otherwise — should occur with a full understanding of its limitations and their consequences.

4.1 Considerations for Policymakers

To account for the limitations of compute thresholds and their consequences, policymakers can consider:

- **Developing dynamic rather than static thresholds**, for example by focusing on models which score within the upper percentiles on a set of regularly-updated benchmarks for capability, safety, or risk. This would more appropriately capture highly capable and higher-risk models by comparing new models to existing top-performing ones. It would also reduce the need for policymakers to frequently revisit, revise, and redefine a static threshold as the state of the art develops. Categorizing such dynamic thresholds by modality would further increase their specificity and effectiveness, as they would discern between models with similar characteristics.
- **Developing alternative approaches to assessing which AI models should face greater levels of scrutiny, to complement training compute measures.** Multiple metrics or criteria should be considered to determine model risk, to reduce the impact of the limitations of any single metric. Examples of these could include assessments of a model's specific deployment context, qualitative assessments of risk based on taxonomies of different types of harm, assessments of how many people may be impacted by an AI model, and so on.
- **Specifying and more clearly defining approaches to calculating and measuring FLOPs** where they continue to be used, for example by establishing industry-wide standards and guidance that account for all stages of model training, quantized models, architectures such as mixture of experts, ensembles, or cascading models, and downstream post-training or fine-tuning.

Exploration of alternative and complementary approaches to risk thresholding remains an ongoing focus for researchers across industry, academia, and the public and non-profit sectors. This primer has focused on detailing some of the limitations of compute-based thresholds in order to bring additional evidence and perspectives to ongoing work of policymakers, regulators, researchers, and developers working to assess and manage the risks of AI models.

Contact:

Aidan Peppin
Policy and Responsible AI Lead
Cohere For AI
aidanpeppin@cohere.com