Operationalising the definition of highly capable AI

Author: Guillem Bas
Affiliation: Observatorio de Riesgos Catastróficos Globales
Contact: gbasg@riesgoscatastroficosglobales.com

Summary

This research note proposes the establishment of a threshold, initially based on the computational resources used to train foundation models, as a mechanism to operationalize the definition of 'very capable foundation models'. Compute is suggested as a pivotal element due to its ability to predict the model's capabilities, as well as its measurability, verifiability, and traceability. However, the text acknowledges the need to update the threshold periodically to account for ongoing technological evolution, and to consider other metrics that could eventually complement the threshold to ensure it holds up over time. As such, an agnostic and future-proof definition is presented.

Background

In May 2023, the European Parliament approved its position on the AI Act, including a paradigm-shifting provision: Article 28b and its obligations for providers of foundation models. Though pertinent, the proposal had some flaws. For instance, the listed requirements were vague and insufficient to regulate the most capable models. Furthermore, several actors expressed concerns about some parts of the article being too burdensome for smaller providers.

A promising solution to ensure concreteness and address concerns about overregulation could be a tiered approach that distinguishes a subset of highly capable foundation models, which would be subject to the already envisioned obligations and even more stringent requirements, such as mandatory third-party model evaluations and red teaming, or an exhaustive risk management system. Though virtually non-existent at the beginning of the process, this idea recently gained traction among some stakeholders (Moës & Ryan, 2023; Zenner, 2023). Later, the presidency of the Council proposed the introduction of a new category named 'very capable foundation model' (Bertuzzi, 2023a) that received broad support during the last trilogue (Bertuzzi, 2023b). However, several doubts remain around the definition of that category.

In this note, we propose the creation of a threshold that could serve as a starting point to specify and operationalise that definition. In the first section, we present the advantages of computational resources (compute) as a promising governance node. Second, we discuss how a compute-based threshold could be established and updated. Third, we consider other metrics that could eventually complement compute thresholds. To conclude, we propose a future-proof definition of 'very capable foundation models' and list the institutions that could help operationalise that definition.

Compute as a governance node

Computing power used during the training process could be the primary metric of the initial threshold. As seen in Figure 1, there is a strong correlation between training compute and the performance of the resulting model (Owen, 2023), so compute is a relatively effective indicator of the emergence of strong capabilities that deserve careful oversight.

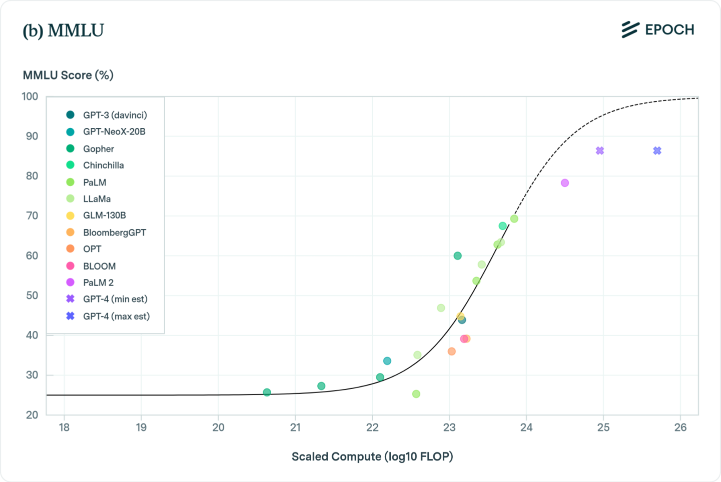

Figure 1. Relation between compute budgets and performance. Retrieved from Owen (2023)

Besides, compute gathers several advantages as a governance node, as it is easily measurable, verifiable, and traceable.

Measurable

Computer performance required to train an AI model is usually measured in floating-point operations (FLOP), i.e., the total number of floating-point arithmetic calculations performed by the end of the training process. There are mainly two methods to estimate the amount of compute used to train a model: (1) a procedure based on the architecture of the network and the number of training batches processed, and (2) a procedure based on the hardware setup and amount of training time (OpenAI, 2018; Sevilla et al., 2022).

The first method requires multiplying the total number of full passes (the number of times that information flows from an input to an output layer of the neural network and vice versa) by the number of operations performed in each pass. The developers can easily gather all this information. However, following this methodology properly requires first-hand knowledge about the architecture and, therefore, might be more challenging to verify by external auditors if the developers are unwilling to reveal its details.

As for the second method, the main variables needed are:

- Hardware used by adding the computing capability of their GPUs in FLOP/second.
- Utilization rate, i.e., the percentage of the maximum capability used, usually between 30% and 40% (Sevilla et al., 2022).
- Training time to estimate the total number of FLOP performed by the hardware.

The Parliament’s draft already requires providers of foundation models to include in the technical documentation a “description of the training resources used by the foundation model including computing power required, training time, and other relevant information related to the size and power of the model.” Labs could use this information to measure training compute following the second methodology. To improve regulatory visibility and facilitate a compute-based classification of models, the Act could also require including this information in the model registration in the EU database envisioned in Article 60.

Creating this precedent through the AI Act would be a milestone for compute governance elsewhere, as compute reports would help public authorities detect which actors have the resources to develop and deploy highly capable models and, therefore, identify where strong governance is needed (Whittlestone & Clark, 2021).

Verifiable

To prevent fraudulent self-reporting, labs’ claims could be verified by third-party auditors. To do so, these auditors could review the project’s planning and budget to identify the available infrastructure, as well as existing logs of training time and utilisation rates of GPUs. Verifiers could then contrast this information with model size and training dataset size to certify that the means are coherent with the results. For instance, unexpectedly good results resulting from a small amount of compute could be a possible indicator of dubious reporting.

Another possibility is measuring compute using publicly available information, which could help cross-check the results of auditing processes. In that regard, independent organisations such as Epoch have already developed a solid methodology to estimate models’ training compute. Finally, authorities could also audit providers of compute infrastructure. This measure is discussed in more depth in the next section.

Traceable

Some of the most notable advantages of hardware as a governance node are its physicality and exclusivity. Regarding the former, compute requires physical space and high energy demand (OECD, 2023), so it is relatively easy to keep track of large clusters. As for the latter, only a few providers, such as Microsoft Azure or Google Cloud have the necessary infrastructure to train frontier models, and only a few vendors, such as NVIDIA or TSMC, design and produce cutting-edge AI accelerators. As such, monitoring compute usage would only require considering a few companies.

To leverage this traceability, compute providers —including vendors of chips and cloud service providers— could be required to report who bought or accessed their resources above a certain threshold and for what purposes. In the context of the EU, tracking digital access to computational infrastructure seems to be more feasible. For that, cloud providers could implement know-your-customer checks and keep records of training runs above the specified compute threshold (Egan & Heim, 2023). The responsibility to access this information could be assumed by the AI Office as part of its envisioned task to “record and monitor known instances of large training runs.” However, since most providers of large infrastructure are outside of the EU, effective enforcement might depend on the ability of the AI Office or any responsible authority to coordinate with other relevant jurisdictions. In that sense, the U.S. Administration has recently approved an Executive Order that requires providers of Infrastructure as a Service to report the "acquisition, development, or possession [of large-scale computing clusters], including the existence and location of these clusters and the amount of total computing power available in each cluster."

The establishment of a compute threshold

Considering its accuracy at estimating capabilities and its relative simplicity, compute is sufficiently robust to be the metric defining the initial threshold. However, to be future-proof, compute thresholds should not be fixed by law but updated periodically considering several dynamics, including the following:

- Downward forces:
    - Algorithmic progress makes certain capabilities accessible with less compute. The introduction of better algorithms halves compute requirements of a given performance level every 4 to 25 months for computer vision (Erdil & Besiroglu, 2022) and every 5 to 13 months for language modeling (Besiroglu et al., forthcoming).
    - Increasing hardware efficiency makes compute cheaper. The amount of FLOP/second per dollar spent doubles every 2 to 3 years (Hobbhahn & Besiroglu, 2022).

- Upward forces:
    - The increasing ability to understand the risks of existing models and how to manage them. Governance of models should be progressively deprioritised —though not abandoned— as they get farther from the state-of-the-art.

The EP's text tasks the AI Office to "issue and periodically update guidelines on the thresholds that qualify training a foundation model as a large training run," which would be especially relevant if those thresholds have regulatory implications.

An initial compute threshold could be set, for example, at 1e24 floating-point operations (FLOP), above which it is estimated that there were about ten models developed by seven organisations by September 2024 (Epoch, 2022).

Model	Developer	Year	Compute (in FLOP)
GPT-4	OpenAI	2023	2.10E+25
PaLM 2	Google	2023	7.34E+24
Claude 2	Anthropic	2023	3.87E+24

Falcon 180B	TII	2023	2.78E+24
Minerva	Google	2022	2.74E+24
GPT-3.5	OpenAI	2022	2.58E+24
PaLM	Google	2022	2.53e+24
Megatron-Turing NLG 530B	Microsoft, NVIDIA	2021	1.17E+24
Ernie 3.0 Titan	Baidu	2021	1.04e+24

Source: Epoch.

When the AI Act starts to apply, the chosen threshold should account for all cutting-edge models that might pose sufficiently significant risks while remaining above a threshold that might be harder to enforce and include less risk-relevant models. As for the former, a certain model might not seem especially risky in its raw form, but it is worth adopting a conservative stance to account for the several improvements enabled after deployment, listed by the British DSIT (2023): better prompts (Wei et al., 2022), better tools (Boiko et al., 2023), better scaffolds (Yang et al., 2023), new fine-tuning data (Yoosuf & Yang, 2019), and interaction with other AI systems (Shen et al., 2023). As for the latter, a threshold currently set at one order of magnitude less —1e23 FLOP— would include around twenty more models (Epoch, 2022), including open-source models like BLOOM for which the requirements discussed for very capable foundation models would probably be unaffordable.

It is worth noting that different thresholds have already been proposed or approved in other jurisdictions. For example, in the recently enacted Executive Order, the U.S. Administration imposes several reporting requirements on models trained using more than 1E+26 FLOP or models using primarily biological sequence data and using more than 1E+23 FLOP. This provisional threshold shall be further defined and updated by the Secretary of Commerce.

Other benchmarks for a composite threshold

In the future, compute thresholds might have significant limitations that updates alone would not solve, including the fact that some individual tasks do not relate as well to compute scale (Owen, 2023) and some scale inversely with training compute (McKenzie et al., 2023). Besides, it is not clear how long scaling laws will apply. To address these limitations and ensure a future-proof regulation, compute thresholds could eventually be complemented with other criteria. In this section, we propose several benchmarks that could be considered for that purpose, classifying them according to the category they could be applied to:

Benchmarks for foundation models

1) Performance assessments based on benchmarks like BIG-bench (Srivastava et al., 2022), MMLU (Hendrycks et al., 2020), or HELM (Liang et al., 2022).

2) Number of tasks realizable by the model, based on lists of tasks such as the O*NET database (Eloundou et al., 2023). It is worth noting that there is no agreement on assessing the number of economically valuable tasks that are automatable (Owen, forthcoming).

Arguably, these two components provide the most accurate representation of the capabilities and limitations of a model. Moreover, there might already exist a basis for their use, as the Parliament's draft requires providers of foundation models to include, in the technical documentation, a "description of the model's performance, including on public benchmarks or state of the art industry benchmarks."

However, the main limitation is that qualitative benchmarks are hard to measure and compare in a standardised manner and prone to saturate in relatively short times (Ott et al., 2022). Besides, developers could fine-tune their models to underperform on specific benchmarks (Chen & Yang, 2023).

To avoid this problem, Moës and Ryan (2023) propose the inclusion of benchmarks that evaluate the ability to generalise across a range of tasks, such as the Abstraction and Reasoning Corpus (Chollet, 2019) or generality analysis (Hernández-Orallo et al., 2021). However, more work is still needed to develop these methods and make them readily available for regulatory purposes.

Given the lack of a robust methodology and the difficulty of assessing performance on such benchmarks (Owen, forthcoming), we currently advise against adopting them. However, we recommend the regulation allow for their use at the discretion of the regulatory body in charge of updating the benchmarks.

Benchmarks for general-purpose AI systems

1) Number of active recipients of the model, i.e., individuals that have engaged with or been exposed to the model, as defined by the DSA. In the case of AI systems, this could be measured by the number of inference API calls.

2) Affordances made available to the system, i.e., the resources and opportunities for influencing the world that are available to a system once deployed and could expand its capabilities (Sharkey, forthcoming). For example, if a model is given access to the Internet, it can access information in real time or use other applications, thus increasing overall capabilities and risk.

The first criterion does not provide information on the system's risk profile. Still, it might help anticipate the societal impact that any incident related to the system could have. The second criterion does affect the system's risk profile and, therefore, might require the implementation of additional measures such as model re-evaluations.

The main limitation of both criteria is that the total number of active recipients and affordances available are only known ex-post, which is less valuable if the objective is identifying regulatory targets. That is particularly problematic for the second benchmark, which would require continuous monitoring of the downstream systems that every highly capable foundation model is integrated into. As explained, this difficulty would justify adjusting thresholds and benchmarks for models downward, accounting for all foreseeable enhancements.

Benchmarks for dual-use narrow AI systems

Some narrow systems with lower compute requirements also carry significant risks, as in the cases of models for drug discovery (Urbina et al., 2022) or software vulnerability detection (Ferrag et al., 2023).

1) Capability-specific assessments, e.g., APPS for code generation (Hendrycks et al., 2021) or MACHIAVELLI for specific harmful behaviors (Pan et al., 2023).

2) Properties of the training data, e.g., using large datasets, including synthesizable molecules or cyber vulnerabilities.

The advantages and limitations of capability-specific assessments are the same as those of general performance assessments. As for data governance, the Act establishes several obligations in that regard for providers of high-risk systems (Article 10) and foundation models (Article 28b(2b) in the EP's draft). Moreover, the Parliament's draft requires providers of foundation models to include, in the technical documentation, a "description of the data sources used in the development of the foundational model." However, none of these provisions require reporting the utilization of datasets that could lead to dangerous capabilities. To address those shortcomings, the regulation could require providers of dual-use narrow systems to indicate the nature and size of datasets, including detailed information about biological sequencing and cyber vulnerabilities or attacks. In case of surpassing a certain threshold, those systems would also need to approve safety evaluations and red-teaming exercises focused on the relevant risk.

However, it is worth noting that the Act might not be fit for regulating some dual-use narrow systems that do not fall into any high-risk categories, as in the case of the cases above. Finally, the main limitation of this criterion is that it might be unfeasible to examine the content of a large dataset at a glance, and developers might be more reluctant to share this data.

Toward a future-proof definition

To ensure that the definition of a ‘very capable foundation model’ is future-proof and can be adapted to future technological progress, no specific thresholds should be established in the regulation's text. Notwithstanding, the definition could be based on two pillars: high performance in several critical tasks, especially those that could pose a risk to public safety, and the existence of a large training run, whether in terms of compute or other inputs. An example of such a definition could be the following:

> ‘Very capable foundation model’ means a foundation model possessing a wide range of high capabilities resulting from a resource-intensive training process, as defined by benchmarks and thresholds set by the Commission.

The European Commission should further specify the operationalisation of this definition in implementing acts. To update the threshold, the Commission should consult all relevant institutions, including the Joint Research Centre, the AI Office, benchmarking authorities, and CEN-CENELEC JTC-21.

References

Bertuzzi, L. (2023a, October 17). AI Act: EU countries headed to tiered approach on foundation models amid broader compromise. *EURACTIV*. https://www.euractiv.com/section/artificial-intelligence/news/ai-act-eu-countries-headed-to-tiered-approach-on-foundation-models-amid-broader-compromise/

Bertuzzi, L. (2023b, October 25). EU policymakers enter the last mile for Artificial Intelligence rulebook. *EURACTIV*. https://www.euractiv.com/section/artificial-intelligence/news/eu-policymakers-enter-the-last-mile-for-artificial-intelligence-rulebook/

Besiroglu, T., Erdil, E., Ho, A., Guo, C., Rahman, R., Owen, D., Cottier, B., & Wynroe, K. (forthcoming). *Algorithmic progress in language models*.

Boiko, D. A., MacKnight, R., & Gomes, G. (2023). *Emergent autonomous scientific research capabilities of large language models*. https://doi.org/10.48550/ARXIV.2304.05332

Chen, J., & Yang, D. (2023). *Unlearn What You Want to Forget: Efficient Unlearning for LLMs*. https://doi.org/10.48550/ARXIV.2310.20150

Chollet, F. (2019). *On the Measure of Intelligence*. https://doi.org/10.48550/ARXIV.1911.01547

DSIT. (2023). *Frontier AI: capabilities and risks – discussion paper*. https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper

Egan, J., & Heim, L. (2023). *Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers* (arXiv:2310.13625). arXiv. http://arxiv.org/abs/2310.13625

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*. https://doi.org/10.48550/ARXIV.2303.10130

Epoch. (2022). *Parameter, Compute and Data Trends in Machine Learning*. https://epochai.org/mlinputs/visualization

Erdil, E., & Besiroglu, T. (2022). *Algorithmic progress in computer vision*. https://doi.org/10.48550/ARXIV.2212.05153

Ferrag, M. A., Battah, A., Tihanyi, N., Debbah, M., Lestable, T., & Cordeiro, L. C. (2023). *SecureFalcon: The Next Cyber Reasoning System for Cyber Security*. https://doi.org/10.48550/ARXIV.2307.06616

Hendrycks, D., Basart, S., Kadavath, S., Mazeika, M., Arora, A., Guo, E., Burns, C., Puranik, S., He, H., Song, D., & Steinhardt, J. (2021). *Measuring Coding Challenge Competence With APPS*. https://doi.org/10.48550/ARXIV.2105.09938

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). *Measuring Massive Multitask Language Understanding*. https://doi.org/10.48550/ARXIV.2009.03300

Hernández-Orallo, J., Loe, B. S., Cheke, L., Martínez-Plumed, F., & Ó hÉigeartaigh, S. (2021). General intelligence disentangled via a generality metric for natural and artificial intelligence. *Scientific Reports*, *11*(1), 22822. https://doi.org/10.1038/s41598-021-01997-7

Hobbhahn, M., & Besiroglu, T. (2022). Trends in GPU price-performance. *Epoch*. https://epochai.org/blog/trends-in-gpu-price-performance

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., … Koreeda, Y. (2022). *Holistic Evaluation of Language Models*. https://doi.org/10.48550/ARXIV.2211.09110

McKenzie, I. R., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., McLean, E., Kirtland, A., Ross, A., Liu, A., Gritsevskiy, A., Wurgaft, D., Kauffman, D., Recchia, G., Liu, J., Cavanagh, J., Weiss, M., Huang, S., Droid, T. F., … Perez, E. (2023). *Inverse Scaling: When Bigger Isn't Better*. https://doi.org/10.48550/ARXIV.2306.09479

Moës, N., & Ryan, F. (2023). *Heavy is the Head that Wears the Crown: A risk-based tiered approach to governing General Purpose AI*. The Future Society.

OECD. (2023). *A blueprint for building national compute capacity for artificial intelligence* (OECD Digital Economy Papers 350; OECD Digital Economy Papers, Vol. 350). https://doi.org/10.1787/876367e3-en

OpenAI. (2018, May 16). AI and compute. *OpenAI*. https://openai.com/research/ai-and-compute

Ott, S., Barbosa-Silva, A., Blagec, K., Brauner, J., & Samwald, M. (2022). Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, *13*(1), 6793. https://doi.org/10.1038/s41467-022-34591-0

Owen, D. (2023). *Extrapolating performance in language modeling benchmarks*.

Owen, D. (forthcoming). *Challenges in predicting AI automation*.

Pan, A., Chan, J. S., Zou, A., Li, N., Basart, S., Woodside, T., Ng, J., Zhang, H., Emmons, S., & Hendrycks, D. (2023). *Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark*. https://doi.org/10.48550/ARXIV.2304.03279

Sevilla, J., Heim, L., Hobbhahn, M., Besiroglu, T., Ho, A., & Villalobos, P. (2022, January 20). Estimating Training Compute of Deep Learning Models. *Epoch*. https://epochai.org/blog/estimating-training-compute

Sharkey, L. (forthcoming). *An Auditing Framework for AI Safety*.

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., & Zhuang, Y. (2023). *HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face*. https://doi.org/10.48550/ARXIV.2303.17580

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2022). *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*. https://doi.org/10.48550/ARXIV.2206.04615

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, *4*(3), 189–191. https://doi.org/10.1038/s42256-022-00465-9

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. https://doi.org/10.48550/ARXIV.2201.11903

Whittlestone, J., & Clark, J. (2021). *Why and How Governments Should Monitor AI Development*. https://doi.org/10.48550/ARXIV.2108.12427

Yang, H., Yue, S., & He, Y. (2023). *Auto-GPT for Online Decision Making: Benchmarks and*

Additional Opinions. https://doi.org/10.48550/ARXIV.2306.02224

Yoosuf, S., & Yang, Y. (2019). Fine-Grained Propaganda Detection with Fine-Tuned BERT. *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, 87–91. https://doi.org/10.18653/v1/D19-5011

Zenner, K. (2023, July 20). A law for foundation models: The EU AI Act can improve regulation for fairer competition. *OECD*. https://oecd.ai/en/wonk/foundation-models-eu-ai-act-fairer-competition