Trends in Frontier AI Model Count: A Forecast to 2028

Iyngkarran Kumar [1] **Sam Manning** [2]

Abstract

Training compute thresholds are increasingly being used as a tool to regulate AI model development and deployment. We therefore forecast the number of models exceeding training compute thresholds in the coming years (2025-28), such as the 10^{25} FLOP threshold in the EU AI Act and the 10^{26} FLOP threshold in the US AI Diffusion Framework. We estimate that by the end of 2028, there will be between 103-306 foundation models exceeding a 10^{25} FLOP threshold and 45-148 models exceeding the 10^{26} FLOP threshold (90% CIs) with median predictions of 165 and 81 models, respectively. We also find that the number of models exceeding these thresholds grows superlinearly, but subexponentially. Compute thresholds that are defined with respect to the largest training run to date (for example, such that all models within one order of magnitude of the largest training run to date are captured by the threshold) see a more stable trend, with a median forecast of 14-16 models being captured by this definition annually from 2025-2028.

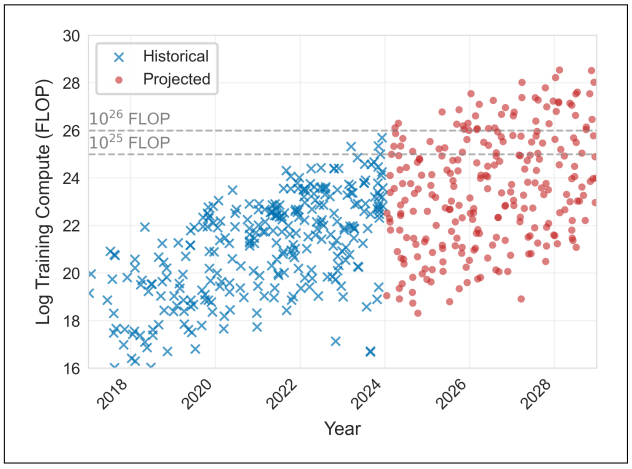

Figure 1. Historical data (2017-2023; blue) and a sample of our model's predictions (2024-2028; red) for the number of AI models exceeding 10^{25} and 10^{26} FLOP.

1. Introduction

Recent years in machine learning have seen the rise of foundation models – AI systems that exhibit powerful and general-purpose capabilities. Governments across the world are starting to impose requirements on the development and deployment of the most capable such systems, such as the GPT o-series (OpenAI, 2024). For example, the EU AI Act (European Union, 2023) and US AI Diffusion Framework (Federal Register, 2025) both subject models exceeding specific training compute thresholds (10^{25} and 10^{26} FLOP respectively) to additional requirements intended to mitigate risks and limit proliferation. However, it is well established that the amount of compute used to train foundation models has been increasing extraordinarily quickly, with recent estimates of 4-5x per year growth rates over the past decade (Sevilla, Jaime and Roldán, Edu, 2024). These trends have important implications for compute-based governance frameworks. In April 2025, estimates suggest that there are 2 publicly available models trained using more than 10^{26} FLOP and approximately 30 publicly available models trained using more than 10^{25} FLOP (EpochAI), yet if current trends continue and governments fail to take this growth into account, they may be left with insufficient capacity to implement their regulations and/or place regulatory burdens on an excessive number of actors.

With this in mind, we attempt to estimate the number of released models that will exceed various compute thresholds over the coming years. Extrapolating from current trends, we conclude that by the end of 2028 there could be between 103-306 models exceeding the 10^{25} FLOP threshold (90% confidence interval) with a median estimate of 165, and 45-148 models exceeding the 10^{26} FLOP threshold (90% confidence interval), with a median estimate of 81. We also study "frontier-connected thresholds" – thresholds that are defined relative to the largest training run at any one point in time rather than based on the absolute amount of training compute used – and estimate that in the coming years there will be between 6-35 models released within 1 order of magnitude (OOM) of the largest training run that has

[1]University of Edinburgh, Edinburgh, UK [2]Centre for the Governance of AI, Oxford, UK. Correspondence to: Iyngkarran Kumar <iyngkarrankumar@gmail.com>.

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada.

taken place (90% CI) with a stable median of 14-16 models captured by this definition. However, our analysis has limitations resulting from selection effects in the database that we extrapolate trends from, as well as uncertainty in key parameters that influence the projections.

Importantly, our estimates do not straightforwardly translate into the number of models in scope of the EU AI Act or the AI Diffusion Framework. Our numbers may provide an overestimate in that neither the EU or US would regulate models trained and made available in other jurisdictions (e.g: China). The EU AI Act specifically targets general purpose AI, potentially excluding image and video generation models; additionally, these regulations could disincentivize companies from releasing models above the thresholds. Conversely, we may underestimate affected models since regulations could extend to companies that simply modify models (Williams et al., 2025).

2. Method

Our aim is to forecast the number of models that will exceed different training compute thresholds in the coming years. To do this, we model scenarios for the distribution of AI model releases over training compute, and count models exceeding each threshold. We use Epoch AI's Notable Models databases (Epoch AI, 2025b) as the main dataset for our analysis, containing approximately 450 machine learning models with estimated training compute. To capture recent trends we analyse data from 2017-2023 (296 datapoints). Models included in this database are selected according to five "notability criteria" (see section A), which induces a selection effect that is discussed further in section 4.

To forecast the distribution of AI models over training compute, we:

1. Project total compute usage for AI workloads (training, inference, and other uses) with a growth rate of 4.1x per year.
2. Allocate this compute between model training and other uses (e.g: inference, compute for experiments), with 40% toward training in 2024-2026 and 30% in 2027-2028.
3. Allocate training compute across models of different sizes according to historical trends (Table 1).
4. Randomly sample models from each size category until the training compute allocation for that category is met.

In the first stage of our predictive model[1] we project the total compute usage for "AI workloads" at a rate of 4.1x per year. The term "AI workloads" is used to refer to compute used for model training, model inference, research experiments, and other uses. The 4.1x growth rate is a weighted average of two compute forecasts; the first comes from a recent analysis (Dean, 2024) estimating that compute for AI workloads grows at 3.4x per year, and the second comes from a 6.3x historical growth rate in the training compute stock estimated from the Notable Models database[2]. We give the first forecast three times as much weight as the second given the more detailed analysis that was used to arrive at this figure, but these weightings are subjective and Appendix B gives results for other reasonable choices of growth rate weighting.

The second stage of our model involves allocating the compute stock for AI workloads to be allocated to model training, inference, and other uses. Following (Dean, 2024), we use a 40% allocation towards model training for the years 2025 and 2026, and a 30% figure for 2027. (Dean, 2024) allocates 20% of compute towards model training by the end of 2027 (and presumably 2028) but we find their forecast to be aggressive with respect to the share of compute allocated to inference, so we instead maintain a 30% training compute share for 2028.

Next, the training compute stock is allocated across models of different size. An example of how this is done when retrodicting the model to 2023 is shown in Table 1. We define model size (in training compute) with respect to the largest training run in a given year, therefore, our predictive model must first make assumptions about the largest training run; this is done by allocating a fraction of the total training compute stock to the largest run. This parameter is called the "largest model share" (or LMS) in our model, and historical values of the LMS suggest it is uniformly distributed over the range $[0.05, 0.50]$ (Appendix C). However, a qualitative interpretation of the LMS parameter is the degree of concentration in the market of developers training AI models at the largest scale, and with several relatively new entrants (x.AI, Inflection, Mistral) joining established actors (OpenAI, Google DeepMind, Anthropic) in training large-scale models (Epoch AI, 2025a), we expect the number of actors training "frontier AI models" (Anderljung et al., 2023) to grow. To quantitatively account for this, we sample the LMS parameter log-normally from the bounds $[0.05, 0.50]$ when projecting the model forward.

The fraction of compute allocated to each model size is controlled by a model parameter referred to as the "allocation gradient" and denoted as k. The parameter is named as such because it is the gradient of a linear fit to the cumulative distribution function of training compute spending as a function of model size (normalised by the largest training

[1] We often refer to the model constructed to produce these forecasts as the "predictive model", to distinguish it from AI models.

[2] Which can be generalised to a 6.3x growth rate in the compute stock used for AI workloads under the assumption that allocation of compute between training and other uses has remained roughly constant in recent years.

run in a given year) - Appendix D shows k for the years 2020-2023 and further discusses it's quantitative interpretation. Varying the allocation gradient amounts to altering the distribution of compute amongst the largest models relative to smaller models; this is shown in Table 10. Historical values of the allocation gradient (Appendix D) indicate that k should be sampled from the uniformly from the range $[0.9, 1.1]$ for our projections. However, it should be noted that sampling k from a lower range than those found in the Notable Models database could serve as a mechanism to correct the notability selection effect previously discussed.

Table 1. Allocating 1.35×10^{26} FLOP of training compute amongst models in 2023. Largest training run: Gemini Ultra @ 5×10^{25} FLOP. Models that use up to 3 OOMs training compute less than Gemini Ultra are not shown, hence fractional allocations do not sum exactly to 1.0.

Model size relative to Gemini Ultra	**Within 2-3 OOM**	**Within 1-2 OOM**	**Within 1 OOM**
Model size (absolute)	5×10^{22}–5×10^{23}	5×10^{23}–5×10^{24}	5×10^{24}–5×10^{25}
Fractional allocation	0.98%	9.4%	90%
Compute allocation (FLOP)	1.32×10^{24}	1.27×10^{25}	1.22×10^{26}

Table 2. Compute allocations (%) for various values of the allocation gradient (k). Table shows percentages for model size ranges relative to the largest training run. For example, with k=1.1, 92% of the training compute stock is allocated to models within an OOM of the largest model, with 7.3% allocated to models within 1-2 OOMs of the largest run, etc. Models smaller than 1×10^{-3} of largest training run not shown (hence rows do not sum exactly to one.)

k	**Within 2-3 OOM**	**Within 1-2 OOM**	**Within 1 OOM**
0.75	2.6	15	82
0.9	1.4	11	87
1.0	0.9	9	90
1.1	0.58	7.3	92
1.25	0.3	5.3	94

Finally, models are randomly sampled from each size bin until the compute allocation for that bin is met. The model is run 1000 times to generate the confidence intervals shown in the next section.

3. Results and verification

Results for the number of models exceeding the absolute compute thresholds are show in Table 3. A median of 165 models exceeding the 10^{25} FLOP threshold in the EU AI Act is predicted, with a 90% confidence interval of 103-306 models. For the 10^{26} FLOP threshold in the AI Diffusion Framework a median of 81 models is predicted, with a 90% CI of 45-148 models.

Table 4 shows growth trends for the number of models exceeding the 10^{25} FLOP threshold. Superlinear growth is seen, with more models released that exceed the threshold in a given year than the year before, however this growth is subexponential. The implications of these growth rates are discussed in Section 4. These trends hold for all absolute compute thresholds.

To validate the model, we retrodict it for $10^{23}, 10^{24}$ and 10^{25} FLOP thresholds for which there exists data from 2020-2023. This is shown in Table 5. All historical datapoints are captured by the 90% confidence intervals, providing considerable evidence that the predicted intervals of Table 3 will capture the number of models above each threshold.

Finally, trends in compute thresholds that are defined with respect to the largest training run are also studied. Specifically, how many models are captured by a threshold that includes all models with training compute within one order of magnitude of the largest training run to date? The predictions of our model for this threshold, and variations, are shown in Table 6. More stable medians can be observed with the median number of models that fall within 1 OOM of the largest training run fixed at 14-16 from 2025-28, though the 90% confidence intervals are wide.

4. Limitations and Discussion

The key limitation of our analysis results from the selection effect applied to models in the Notable Models database. Models with larger training compute are more likely to satisfy the notability criteria, which leads to our median estimates being biased towards underestimating the true number of models that will exceed the compute thresholds. The predictive model presented in this paper does offer a potential correction - by sampling the allocation gradient (k) from a lower range than is seen in the Notable Models database, more compute can be allocated to smaller models relative to their larger counterparts, compensating for the over-representation of large models in the database. Model predictions when k is varied as such are shown in Appendix E. However, this is not done in the results presented in Section 3 as it is unclear how exactly to adjust k to compensate for the selection effect. The analysis is also limited by the limited historical data with which to calibrate key parameters of the model, such as the largest model share

Table 3. Results for absolute thresholds. The table presents 90% prediction intervals [5th, 50th, 95th percentile] for the number of models exceeding each compute threshold. Results are cumulative, showing estimates for the number of models released by the end of each year.

Threshold (FLOP)	**2025**	**2026**	**2027**	**2028**
$> \mathbf{10}^{25}$	[32, 45, 64]	[51, 77, 119]	[76, 117, 201]	[103, 165, 306]
$> \mathbf{10}^{26}$	[3, 7, 11]	[12, 24, 38]	[27, 47, 81]	[45, 81, 148]
$> \mathbf{10}^{27}$	[0, 0, 0]	[0, 2, 5]	[1, 10, 20]	[9, 27, 56]
$> \mathbf{10}^{28}$	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]	[0, 3, 8]
$> \mathbf{10}^{29}$	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]

Table 4. Projection of models exceeding 10^{25} FLOP threshold by year, showing predicted ranges [5th, 50th, 95th percentile], yearly differences in median values, and year-on-year median growth rates.

	2025	**2026**	**2027**	**2028**
$> \mathbf{10}^{25}$	[32, 45, 64]	[51, 77, 119]	[76, 117, 201]	[103, 165, 306]
Median Differences	22	32	40	48
Median Multipliers	1.96	1.71	1.52	1.41

Table 5. Absolute compute thresholds retrodiction. Each cell is formatted as O $(5, 50, 95)$ where O, $5, 50, 95$ are the historically observed values, 5th percentile, 50th percentile (median) and 95th percentile prediction.

Threshold (FLOP)	**2020**	**2021**	**2022**	**2023**
$> \mathbf{10}^{23}$	2 (0,1,4)	9 (8,13,27)	29 (19,29,60)	54 (36,54,128)
$> \mathbf{10}^{24}$	0 (0,0,0)	3 (0,2,3)	8 (3,8,10)	19 (13,22,44)
$> \mathbf{10}^{25}$	0 (0,0,0)	0 (0,0,0)	0 (0,0,0)	4 (0,4,5)

Table 6. Frontier-connected thresholds. Results show 90% prediction intervals [5th, 50th, 95th percentile] for models within specified distances of the frontier model. Results for each year represent new models released in that year only.

Distance from frontier model	**2025**	**2026**	**2027**	**2028**
Within 0.5 OOM	[3, 6, 14]	[2, 7, 16]	[3, 8, 15]	[3, 8, 17]
Within 1 OOM	[7, 14, 25]	[6, 15, 29]	[7, 16, 31]	[7, 15, 35]
Within 1.5 OOM	[12, 20, 39]	[11, 22, 46]	[11, 24, 50]	[11, 23, 55]

(LMS) and allocation gradient (k). The model is fit with data from 2017-2023, providing only six datapoints to set these parameters.

What are the implications of these results on compute-based regulatory thresholds? The superlinear growth in the number of models captured by absolute compute thresholds could strain regulatory capacity and/or impose excessive burdens on AI developers. Policymakers have two main options: 1) Make regulatory requirements more proportionate for lower-risk models while expanding enforcement capacity, or (2) update the thresholds over time to exclude certain models from the requirements. Both the EU AI Office and US Bureau of Industry and Security can adjust the respective frameworks' thresholds. Frontier-connected thresholds show more stable annual model counts, offering an alternative to absolute compute thresholds as a first pass filter for regulatory requirements.

References

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., et al. Frontier ai regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*, 2023.

Dean, R. Compute forecast — ai 2027. https://ai-2027.com/research/compute-forecast, 2024. Accessed April 2025.

Epoch AI. Models over 1e25 flop. https://epoch.ai/data-insights/models-over-1e25-flop, 2025a. Accessed April 2025.

Epoch AI. Notable ai models. https://epochai.org/data/notable-ai-models, 2025b. Accessed April 2025.

EpochAI. Update on large-scale ai models: 6 new models with 1e25 or more training flops were released in the past two months, bringing the total count to 30. URL https://x.com/EpochAIResearch/status/1910791160970555403.

European Union. The EU AI act. https://artificialintelligenceact.eu/, 2023. Accessed April 2025.

Federal Register. Framework for artificial intelligence diffusion. https://www.federalregister.gov/documents/2025/01/15/2025-00636/framework-for-artificial-intelligence-diffusion, 2025. Accessed January 2025.

OpenAI. Introducing OpenAI o1. https://openai.com/o1/, 2024. Accessed April 2025.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2022.

Sevilla, Jaime and Roldán, Edu. Training compute of frontier ai models grows by 4-5x per year. *Epoch AI Blog*, 2024.

Williams, S., Schuett, J., and Anderljung, M. On regulating downstream ai developers. *arXiv preprint arXiv:2503.11922*, 2025.

A. Notable Models Database

This appendix presents some basic information about the notable models database, the selection criteria that are used to populate the database, and the distribution of models in the database over training compute.

The Notable Models database, curated by EpochAI, contains over "over 900 models that were state of the art, highly cited, or otherwise historically notable." Models are considered notable if they satisfy any of the five criteria below:

1. highly cited (over 1000 citations);
2. large training cost (over $1,000,000, measured in 2023 USD);
3. significant use (over one million monthly active users);
4. state of the art performance (typically on a recognised ML benchmark);
5. indisputable historical significance.

Over 400 of these entries have training compute estimates.

Figures 2 and 3 show the distribution of notable models over training compute. Both plots show a deviation of 2024 data from previous years suggesting the data to be incomplete (this was also verified by a member of staff at EpochAI). Therefore we exclude 2024 data when fitting our model. We also do not use data of models released before 2017, as this corresponds to the era prior to the Transformer architecture that is at the heart of most frontier AI models today.

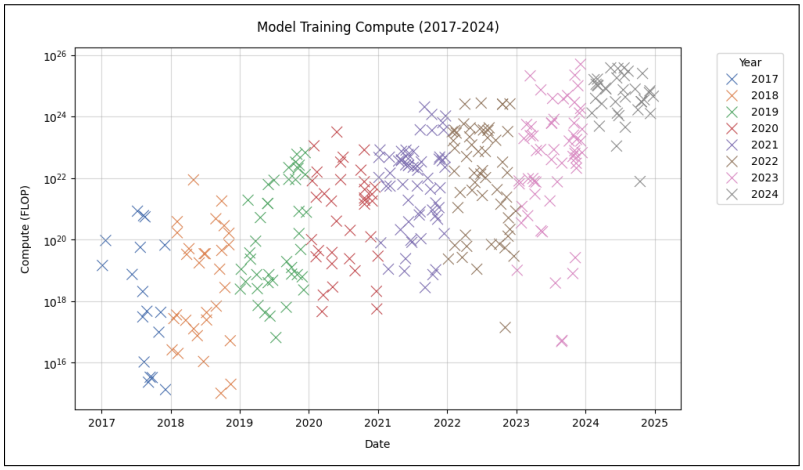

Figure 2. Historical distribution of models over training compute - scatter plot.

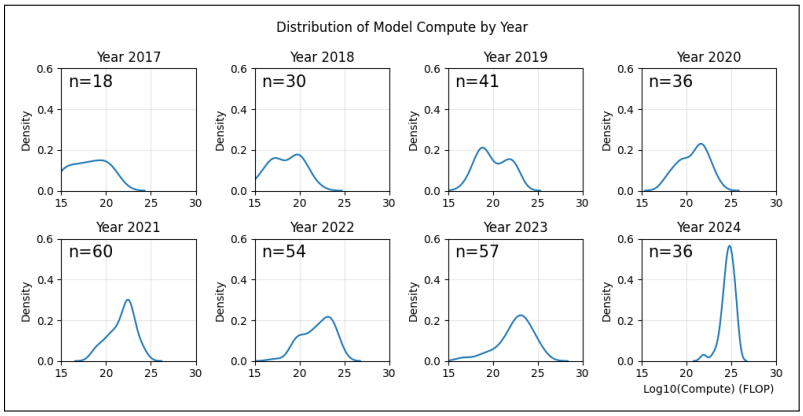

Figure 3. Historical distribution of models over training compute - KDEs

B. Results for alternate growth rate weightings

Our baseline results use a 4.1x growth rate in the compute stock used for "AI workloads". This figure is a weighted average between the 3.4x figure of (Dean, 2024) and a 6.3x figure derived from the Notable Models database, with a weighting of 3:1 between these forecasts. This is a subjective choice; therefore, we present results for alternate growth rate weightings in this appendix.

Table 7. Results for absolute thresholds with growth weighting of (0.1,0.9) between historical growth rate (6.3x) and forecasted growth rate (3.4x).

Threshold (FLOP)	**2025**	**2026**	**2027**	**2028**
$> \mathbf{10}^{25}$	[27, 41, 57]	[45, 71, 109]	[63, 106, 179]	[88, 150, 268]
$> \mathbf{10}^{26}$	[0, 5, 8]	[8, 18, 31]	[18, 36, 68]	[34, 64, 123]
$> \mathbf{10}^{27}$	[0, 0, 0]	[0, 0, 3]	[0, 4, 13]	[2, 15, 37]
$> \mathbf{10}^{28}$	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]	[0, 0, 4]
$> \mathbf{10}^{29}$	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]

Table 8. Results for absolute thresholds with growth weighting of (0.33,0.66) between historical growth rate (6.3x) and forecasted growth rate (3.4x).

Threshold (FLOP)	**2025**	**2026**	**2027**	**2028**
$> \mathbf{10}^{25}$	[33, 48, 69]	[54, 84, 128]	[78, 129, 207]	[106, 178, 305]
$> \mathbf{10}^{26}$	[5, 8, 11]	[16, 26, 38]	[30, 53, 87]	[50, 87, 152]
$> \mathbf{10}^{27}$	[0, 0, 0]	[0, 2, 5]	[3, 11, 25]	[13, 30, 57]
$> \mathbf{10}^{28}$	[0, 0, 0]	[0, 0, 0]	[0, 0, 1]	[0, 4, 9]
$> \mathbf{10}^{29}$	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]

Table 9. Results for absolute thresholds with growth weighting of (0.5,0.5) between historical growth rate (6.3x) and forecasted growth rate (3.4x).

Threshold (FLOP)	**2025**	**2026**	**2027**	**2028**
$> \mathbf{10}^{25}$	[37, 52, 78]	[59, 90, 146]	[83, 136, 239]	[115, 195, 363]
$> \mathbf{10}^{26}$	[6, 9, 14]	[19, 31, 48]	[34, 59, 107]	[58, 101, 189]
$> \mathbf{10}^{27}$	[0, 0, 0]	[0, 4, 7]	[7, 17, 30]	[21, 41, 77]
$> \mathbf{10}^{28}$	[0, 0, 0]	[0, 0, 0]	[0, 0, 2]	[0, 8, 17]
$> \mathbf{10}^{29}$	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]

C. Further discussion of largest model share parameter (LMS)

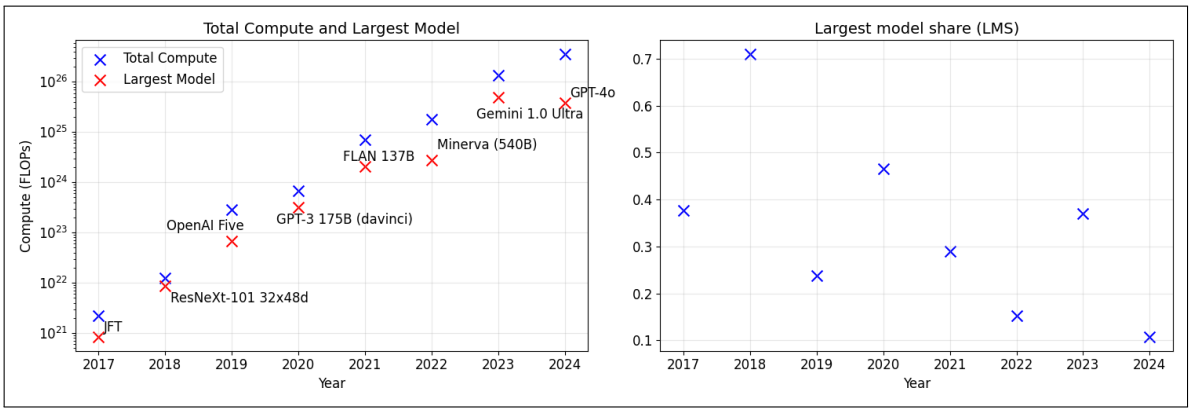

Figure 4. Left: Historical values of total training compute spending and largest training run. Right: Largest model share (LMS) derived from left-hand side plot.

Historical values of the LMS parameter and the total compute and the largest model each year that the LMS is derived from. The AlphaGo family of models have been removed from the dataset on the basis of being outliers, as done in similar analyses (Sevilla et al., 2022). We fit the model on data from 2017-2023 and also discount the LMS for 2018 as it appears to be an outlier. Our predictions sample the LMS uniformly from the range [0.05,0.5]. The upper bound is chosen to accommodate GPT-3 davinci accounting for ∼46% of training compute in 2020. The lower bound is chosen with the 2022 value of 0.15 in mind, however we incorporate a wide range underneath this value due to an increased sensitivity of the model's prediction to small values of the LMS parameter.

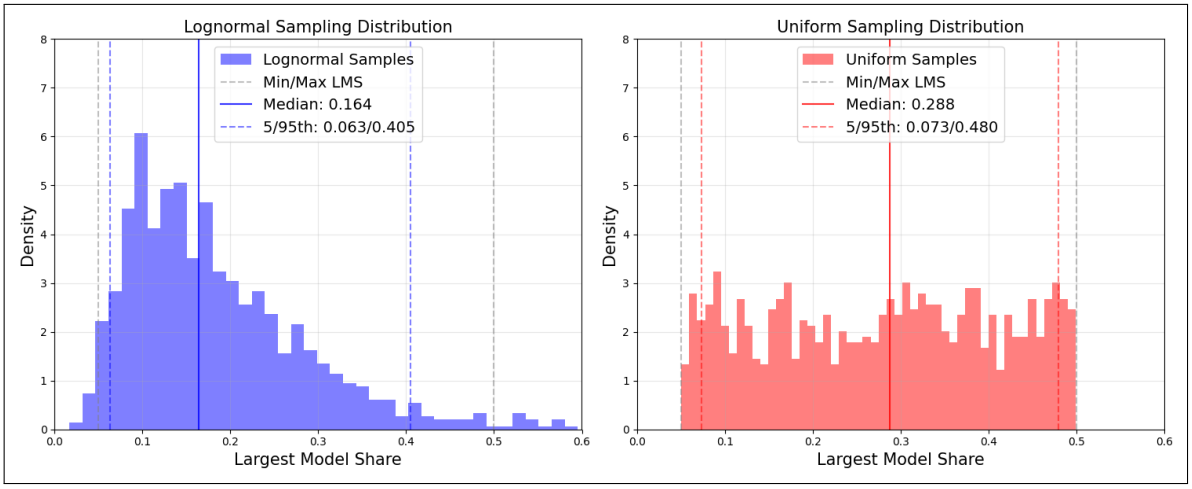

Figure 5. Lognormal and uniform sampling distributions of the LMS sampling parameter. The predictive model samples LMS values lognormally to model an increasing number of actors training frontier models.

D. Further discussion of allocation gradient parameter (k)

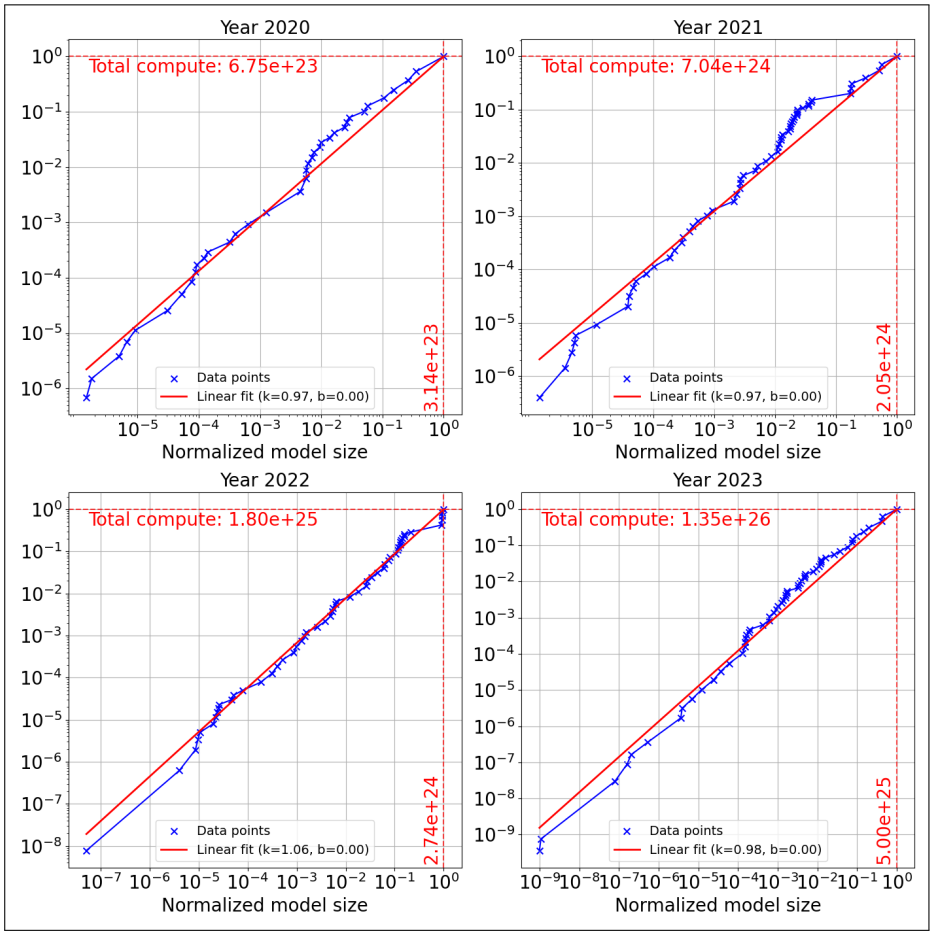

Figure 6. Linear fits to the cumulative distribution of annual training compute spending as a function of normalised model size (where models are normalised by the largest training run in a given year). These plots determine how training compute is allocated to models of various size, hence they are referred to as "allocation plots". Both axes are log-scaled.

Figure 6 shows (on a log-log scale) the cumulative distribution function of training compute spending as a function of normalised model size, alongside linear fits. To illustrate what these plots indicate, consider a datapoint from 2023 - we see that 1×10^{-4} of total compute spending was accounted for by models that were trained with up to 1×10^{-4} the compute of the largest training run that year (Gemini Ultra, 5×10^{25} FLOP). There is a consistent linear relationship between these two plots across the years 2017-2023 (though 2017-2017 data is not shown), hence this is a trend we assume remains constant in the coming years. The gradient of these plots is the allocation gradient (k) that is introduced in Section 2 - historical values of tis parameter are plotted in Figure 7. The mathematical details and interpretation of k are discussed in the following subsection.

D.1. Interpretation of linear fits to allocation plots

This subsection discusses the constraints on the linear fit to the compute allocation trends, and the interpretation of the allocation gradient parameter.

Observing historical data we see that the relationship between normalized model size (normalized by the largest model trained that year) - $\tilde{m}$ and the fraction of compute spent on models of size $\tilde{m}$ or less (the cumulative distribution function, denoted by $A(\tilde{m})$) is linear in log-space. Mathematically:

$$\log(A(\tilde{m})) = k \cdot \log(\tilde{m}) + b \tag{1}$$

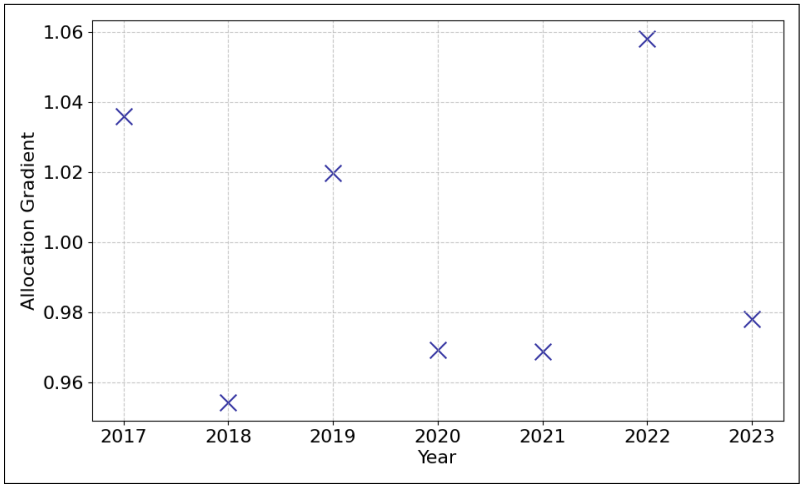

Figure 7. Historical values of the allocation gradient parameter k.

Let that largest model trained in a given year be m_{max}, then $\tilde{m}_{\text{max}} = 1$. Models of size m_{max} or smaller (i.e., all models) take up all compute spending that year, therefore $A(\tilde{m}_{\text{max}} = 1) = 1$. Enforcing this constraint on equation 1 means that $b = 0$ - and so equation (1) reduces to $A(\tilde{m}) = \tilde{m}^k$.

The parameter k determines how compute is allocated across models of different scales. To see this, let us first denote $a(m_1, m_2)$ as the amount of compute that is allocated to models in the range $[m_1, m_2)$. Consider also three sizes of models - m^*, $10m^*$, and $100m^*$. The compute allocated to models in the range $[m^*, 10m^*]$ is $a(m^*, 10m^*) = A(10m^*) - A(m^*) = (10^k - 1)m^k$ using equation (1). The compute allocated to models in the range $[10m^*, 100m^*)$ is $a(10m^*, 100m^*) = A(100m^*) - A(10m^*) = 10^k(10^k - 1)m^k$ after some simplification. Therefore, the relationship between $a(m^*, 10m^*)$ and $a(10m^*, 100m^*)$ is simply:

$$a(10m^*, 100m^*) = 10^k \cdot a(m^*, 10m^*) \tag{2}$$

In other words, scaling up model size by a factor of 10 leads to a factor of 10^k increase in compute allocated to models of this size. $k = 1$ means that these larger models get 10 times as much compute as their smaller counterparts. $k > 1$ means that they get a factor greater than 10, and $k < 1$ leads to a factor less than 10.

E. Results under alternate distributions of allocation gradient (k)

Table 10. Compute allocations (%) for various values of the allocation gradient (k). Table shows percentages for model size ranges relative to the largest training run. For example, with k=0.5, 68% of the training compute stock is allocated to models within an OOM of the largest model, with 22% allocated to models within 1-2 OOMs of the largest run, etc.

k	Within 4-5 OOM	Within 3-4 OOM	Within 2-3 OOM	Within 1-2 OOM	Within 1 OOM
0.5	0.68	2.2	6.8	22	68
0.6	0.3	1.2	4.7	19	75
0.7	0.13	0.64	3.2	16	80
0.8	0.053	0.34	2.1	13	84
0.9	0.022	0.17	1.4	11	87
1.0	0.009	0.09	0.9	9	90

Our baseline scenario samples the allocation gradient uniformly from the range [0.9, 1.1]. The median prediction in this scenario will therefore follow a compute allocation across model sizes as shown in the k=1 scenario in Table 10. This modeling choice is made from observations of the allocation plots for the notable models released in the years 2017-2023 (Appendix D).

However Section 4 discusses the limitations of the Notable Models database upon which these trends are based, specifically, the notability criterion (Appendix A) leads to models that use significant amounts of training compute being over-represented in the dataset. One potential way to account for the Notable Models selection effect is to allocate more compute to smaller models relative to their larger counterparts. This can be seen in the table above where the k=0.5 case allocates ∼68% of compute that year to the largest model category, whereas the k=1.0 case allocates 90% of compute. More generally, increasing model size by 10x leads to a 10^k times increase in compute allocated, as shown in Appendix D.1.

This appendix presents model predictions for allocation gradients that allocate relatively more compute to smaller model sizes. Specifically, Table 11 presents the results of the model when the allocation gradient (k) is sampled from the range [0.7,0.9] (corresponding to a median scenario in which k = 0.8), and Table 12 presents the results of the model when the allocation gradient is sampled from the range [0.5,0.7] (corresponding to a median scenario in which k = 0.6). Notably more aggressive medians can be observed in the later years of the projection for 10^{25} and 10^{26} FLOP thresholds compared to the baseline - this is because these scenarios allocate relatively more compute to smaller models, and in the years 2027 and 2028, 10^{25} and 10^{26} FLOP models are multiple orders of magnitude away from the largest training runs.

Table 11. Results for absolute compute thresholds when the allocation gradient is sampled uniformly from the range [0.7,0.9].

Threshold (FLOP)	**2025**	**2026**	**2027**	**2028**
$> \mathbf{10}^{25}$	[52, 74, 99]	[71, 114, 172]	[97, 160, 265]	[132, 221, 380]
$> \mathbf{10}^{26}$	[8, 13, 18]	[19, 34, 52]	[36, 63, 106]	[59, 105, 183]
$> \mathbf{10}^{27}$	[0, 0, 2]	[0, 4, 11]	[7, 17, 34]	[20, 40, 77]
$> \mathbf{10}^{28}$	[0, 0, 0]	[0, 0, 0]	[0, 0, 4]	[0, 7, 20]
$> \mathbf{10}^{29}$	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]

Table 12. Results for absolute compute thresholds when the allocation gradient is sampled uniformly from the range [0.5,0.7].

Threshold (FLOP)	**2025**	**2026**	**2027**	**2028**
$> \mathbf{10}^{25}$	[36, 49, 67]	[70, 106, 155]	[116, 199, 314]	[205, 359, 637]
$> \mathbf{10}^{26}$	[2, 5, 9]	[12, 21, 34]	[29, 55, 88]	[63, 113, 196]
$> \mathbf{10}^{27}$	[0, 0, 0]	[0, 1, 4]	[3, 7, 16]	[12, 26, 52]
$> \mathbf{10}^{28}$	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]	[0, 1, 6]
$> \mathbf{10}^{29}$	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]