

For a few years the papers announcing the largest artificial-intelligence models stated their size in the first paragraph. The largest American developers no longer state it for their flagship models. OpenAI stopped in March 2023, citing competition and safety. A year earlier a public experiment had already shown that the number, on its own, could not rank one model against another: a model a quarter the size of its rival, trained on far more text for the same budget, had beaten it. The regularities behind that result, empirical scaling laws, have roots in learning-curve research of the 1990s and a grammar-checking experiment of 2001. They are well tested and narrower than their reputation: they forecast a specified quantity, usually loss, under specified conditions; capabilities need forecasts of their own, which sometimes miss; and in 2020 the researchers who popularised the laws wrote that they had no solid theoretical understanding of them.

1. The number that stopped being printed

In May 2020 OpenAI posted the paper introducing GPT-3. The size of the model is in the abstract:

we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model

Nineteen months later DeepMind described "up to a 280 billion parameter model called Gopher" in its own abstract. Four months after that Google opened the PaLM paper by saying its authors "trained a 540-billion parameter, densely activated, Transformer language model". In those papers the parameter count was the headline number.

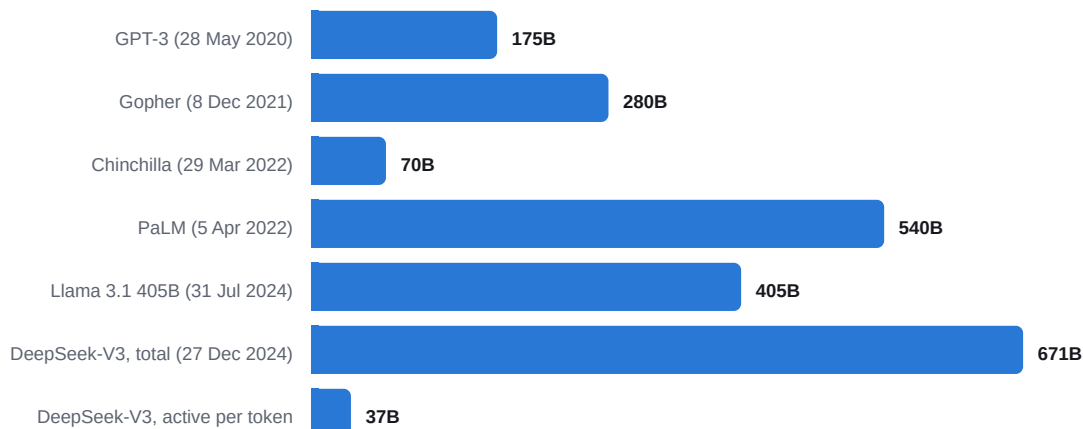
On 15 March 2023 OpenAI published the GPT-4 Technical Report. Section 2 of that document reads:

Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

Three and a half years later the silence is the norm among the largest closed developers. OpenAI's deployment-safety page for GPT-6 Astra, published on 3 September 2026, carries a section headed *Model Data and Training* that contains no quantity of any kind; the corresponding sections for GPT-5 (August 2025) and GPT-5.5 (April 2026) differ in wording and likewise give no parameter count, token total or training-compute figure. Anthropic's system card for Claude Opus 5 names its crawler, its robots-file posture, its post-training and a knowledge cutoff of May 2026, and no size. Google DeepMind's model card

for Gemini 3 Pro explains why total parameters and active parameters are now two different numbers — a sparse mixture-of-experts model "activate[s] a subset of model parameters per input token" — and states neither.

Figure 1. Parameter counts stated in the abstracts of headline model papers



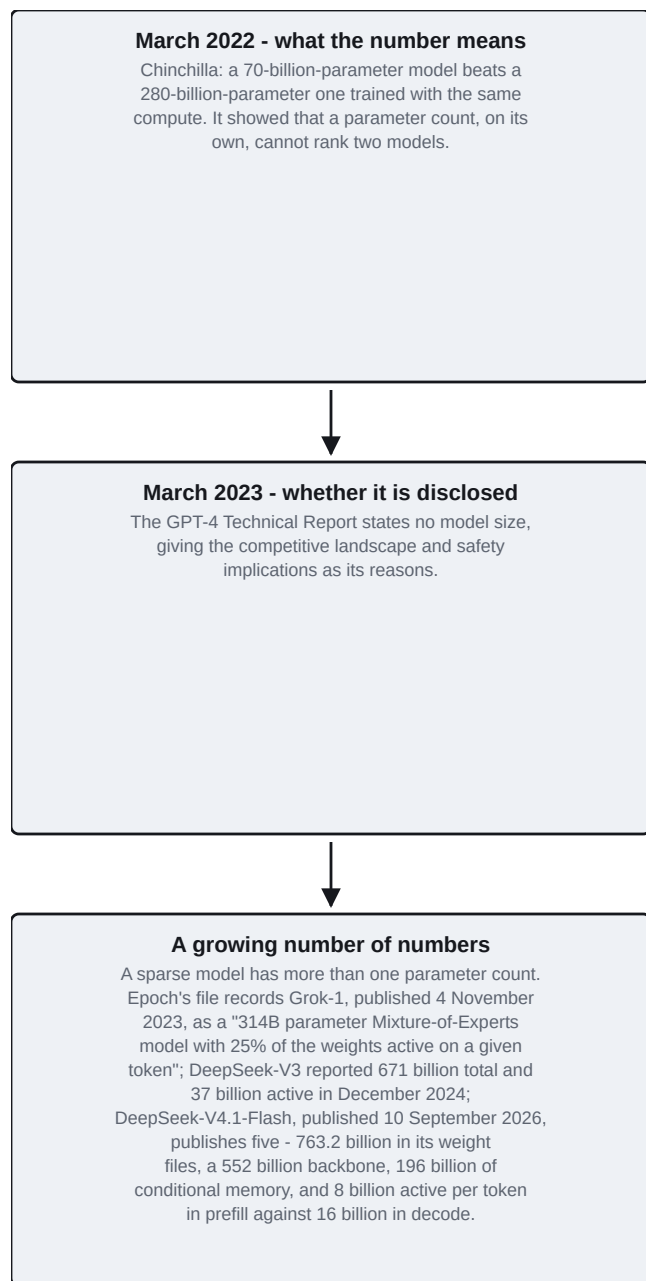
Each figure appears in the paper's own abstract on arXiv. The GPT-4 Technical Report (arXiv 2303.08774) states no parameter count, so it has no bar; its section 2 says the report contains no details about model size. DeepSeek-V3 carries two numbers because a mixture-of-experts model runs only part of itself for each token.

▼ Table view

Figure 1. Parameter counts stated in the abstracts of headline model papers

Document (date submitted to arXiv)	Parameters stated
GPT-3 (28 May 2020)	175B
Gopher (8 Dec 2021)	280B
Chinchilla (29 Mar 2022)	70B
PaLM (5 Apr 2022)	540B
Llama 3.1 405B (31 Jul 2024)	405B
DeepSeek-V3, total (27 Dec 2024)	671B
DeepSeek-V3, active per token	37B

Model capability, disclosure and the arithmetic of a sparse model are separate questions, and the dates below are routinely run together.

Figure 2. Three events, often merged into one

The first two dates are the arXiv submission dates of those documents. The three strands are independent of one another: the first concerns meaning, the second disclosure, the third arithmetic. The third carries no single date - the count of counts has been rising since at least 2021.

▼ Table view

Figure 2. Three events, often merged into one — stages

#	Stage	Note
1	March 2022 - what the number means	Chinchilla: a 70-billion-parameter model beats a 280-billion-parameter one trained with the same compute. It showed that a parameter count, on its own, cannot rank two models.

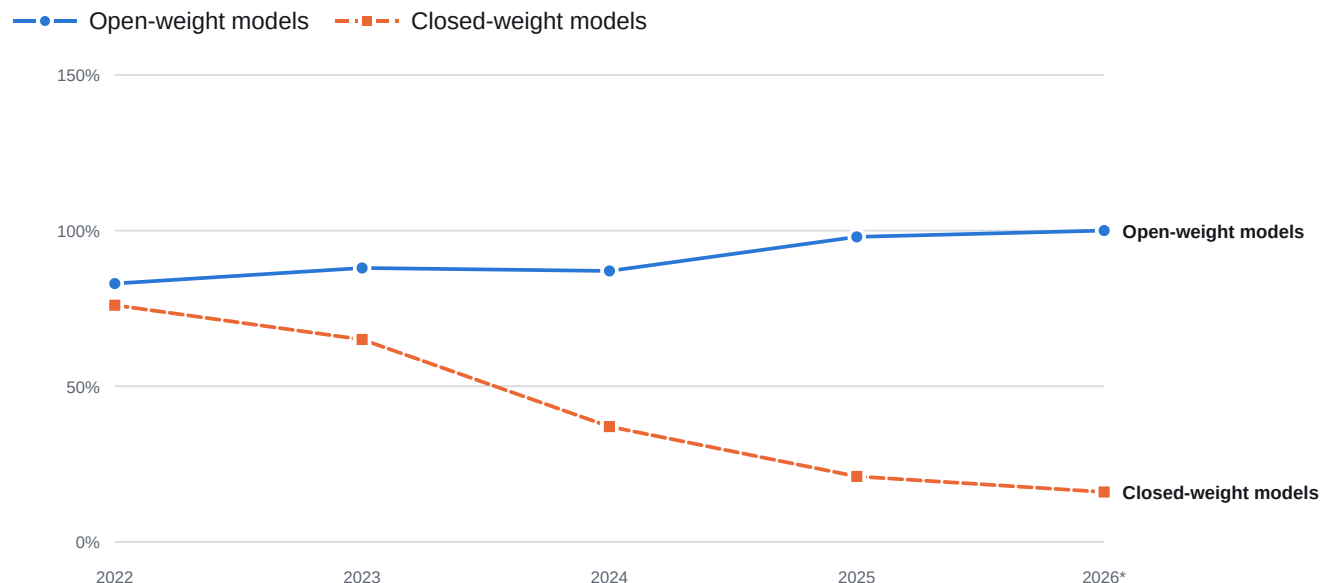
#	Stage	Note
2	March 2023 - whether it is disclosed	The GPT-4 Technical Report states no model size, giving the competitive landscape and safety implications as its reasons.
3	A growing number of numbers	A sparse model has more than one parameter count. Epoch's file records Grok-1, published 4 November 2023, as a "314B parameter Mixture-of-Experts model with 25% of the weights active on a given token"; DeepSeek-V3 reported 671 billion total and 37 billion active in December 2024; DeepSeek-V4.1-Flash, published 10 September 2026, publishes five - 763.2 billion in its weight files, a 552 billion backbone, 196 billion of conditional memory, and 8 billion active per token in prefill against 16 billion in decode.

Figure 2. Three events, often merged into one — connections

From	To	Label
March 2022 - what the number means	March 2023 - whether it is disclosed	
March 2023 - whether it is disclosed	A growing number of numbers	

2. What the disclosure record shows, measured

Epoch AI, a research organisation that maintains a public dataset of notable AI models, records for each entry a publication date, whether the weights are open, and a parameter figure and a training-compute figure where it judges one defensible. Grouped by year of publication, Epoch's records show a widening gap between open- and closed-weight models.

Figure 3. Share of Epoch AI's notable-model records carrying a parameter figure

Source: Epoch AI, Notable AI Models dataset (epoch.ai/data/notable_ai_models.csv), downloaded 13 September 2026 and re-downloaded on 16 September, when every figure here reproduced unchanged; the dataset page reads 'Updated Sep. 11, 2026' and the file holds 1,065 rows, the latest published 8 September 2026. Rows are grouped by Epoch's own 'Open model weights?' field; rows with that field blank are excluded. The counts behind the percentages are, for 2022 to 2026, open-weight 34/41, 52/59, 34/39, 44/45 and 26/26, and closed-weight 37/49, 39/60, 22/59, 13/62 and 7/45. A filled cell may be a developer's figure or Epoch's estimate, and may have been added after publication. *2026 covers publications to 8 September.

▼ Table view

Figure 3. Share of Epoch AI's notable-model records carrying a parameter figure

Year of publication	Open-weight models	Closed-weight models
2022	83%	76%
2023	88%	65%
2024	87%	37%
2025	98%	21%
2026*	100%	16%

In 2022 the two halves of the field were recorded almost identically: a parameter figure sat against 83% of open-weight models and 76% of closed ones. For 2026 the open half stands at 100% and the closed half at 16%. On training compute the divergence is sharper still: 62% against 4% for 2026, from 66% and 67% in 2022.

A filled cell is Epoch's record rather than necessarily a developer's statement, and a notes column gives the provenance of each figure. Of the seven closed-weight 2026 rows that carry a parameter figure, at least three take it from an open checkpoint the product is built on or shares; three more cite a developer's own page, table or public statement, though one of those three sits in the file beside an open row carrying the identical figure, and the note does not say which was the source; and one — xAI's Grok 4.20, at 500 billion — comes from a post on X by Elon Musk, as Epoch's note records. That last figure rests on no document, and the quoted post describes "our V8 small foundation model" rather than the model named in the row.

The four largest American developers contribute none of the seven. OpenAI's most recent parameter figures in the file are for gpt-oss-120b and gpt-oss-20b, both published on 5 August 2025 and both open-weight; Anthropic has no parameter figure anywhere in the dataset, in any year; Google DeepMind's last for a language model of its own is Gemini Nano, in December 2023; Meta's is Llama 4, in April 2025. GPT-6 Astra's row, dated 3 September 2026, has an empty parameter cell and an empty training-compute cell, as do those for Claude Opus 5, Gemini 3.8 Flash and Muse Spark 1.3. Epoch describes the list as non-exhaustive, assembled from literature reviews, conference proceedings, highly cited publications and suggestions, so the shares describe what it records rather than everything released.

The deeper asymmetry is about checkability rather than candour. A parameter count for an open-weight model is a property of a downloadable artefact, and summing the tensor shapes in the published files yields it whether or not the developer writes it down. The model card in Z.ai's GLM-5.3 repository, created in August 2026, gives benchmark protocols and sampling settings and no parameter statement at all; Hugging Face's own tally of the parameters in the published weight files is 753,329,940,480. Epoch's row for the same model reads 744 billion, taken from its note that GLM-5.3 shares a base model with GLM-5.2; the two numbers differ by about 1%, and they are measurements of different things by different parties rather than a correction of one by the other. For a closed model no equivalent route exists. The two lines in Figure 3 are therefore not two degrees of the same virtue: one group could not conceal the number if it wished to, and the other's number cannot be checked from outside.

How many counts a model has is itself no longer settled. DeepSeek-V4.1-Flash, published on Hugging Face on 10 September 2026, states five different parameter figures: its weight files total 763,205,315,794 by Hugging Face's own tally, and its card describes "552B backbone parameters", "Engram conditional memory (196B parameters, sparsely accessed via token-based lookup)", and an architecture that activates "only 8B parameters per token during prefill and 16B during decode". Only the first of those five is computable from the published files; the rest are DeepSeek's own account of its own model. Active-per-token, the figure a sparse model was supposed to need alongside its total, is here two figures rather than one, and which applies depends on the phase of inference. That model has no row in Epoch's dataset, whose page reads "Updated Sep. 11, 2026" — a concrete instance of the non-exhaustiveness Epoch states, and a caution against reading those shares as a census of what was released.

Open developers' reports disclose different parts of the training record. DeepSeek's V3 report of December 2024 states the training run's hardware time: 2.788 million H800 GPU-hours. Its V4 report, dated April 2026, states parameter counts and token counts and no hardware total. Moonshot's model card for Kimi K3 calls it "a 2.8T-parameter model", tabulates 104 billion activated parameters, credits its sparsity framework with "an approximate 2.5× improvement in overall scaling efficiency over Kimi K2", and gives no pre-training token count; its published weight files total 2,779,931,837,184 parameters by Hugging Face's tally, so that figure needs nobody's word. Epoch's note on the model has to reason from peer models and the rule of thumb that training compute is about six times parameters times tokens. In these reports the shape of the model is stated and the hardware total or the token count is not.

3. Two readings the documents do not support

The first reading is that a scaling law is a law in the sense that gravitation is a law, so improvement is guaranteed to continue on a schedule. The paper that put the phrase into general circulation in machine learning, by Jared Kaplan and colleagues at Johns Hopkins University and OpenAI, says otherwise in an appendix headed *Caveats*:

At present we do not have a solid theoretical understanding for any of our proposed scaling laws. The scaling relations with model size and compute are especially mysterious.

The same paper fixes a limit on its own extrapolation — "Our trends must eventually level off, though, since natural language has non-zero entropy" — and a section headed *Contradictions and a Conjecture* identifies what its authors call an apparent contradiction between two of their own fits at scales far beyond anything measured, which they say implies "our scaling laws must break down before this point". Theory has since caught up in part: in 2021 Yasaman Bahri and colleagues, Kaplan among them, proposed "a theory that explains the origins of and connects these scaling laws", identifying four distinct scaling regimes.

The second reading is the mirror image: that the curves were hype, stopped working and were quietly dropped. Withholding a number says nothing about whether a curve failed. The record contains a dated, prospective and public test of them, in the same GPT-4 report that withheld the model's size, and by that report's account the curve passed. Researchers were still fitting such curves to plan training runs in 2026. The curves relate loss to model size, data and compute together, so a count of one of those three inputs does not by itself specify the training conditions a forecast needs.

4. What a scaling law is, and where the idea came from

A learning curve records how often a method errs on examples held out of its training as the amount of training data increases, conventionally in tenfold steps so that a wide range fits on one chart. An empirical scaling law is such a curve measured over an unusually wide range and found regular enough to support forecasts at larger scales.

Learning curves were already being used to make decisions before language models. In 1993 Corinna Cortes and colleagues proposed, at the NIPS conference, a way to predict a classifier's suitability from its learning curve and so avoid "the costly procedure of training poor classifiers on the whole training set". An early and unusually clean demonstration in language processing followed eight years later. In 2001 Michele Banko and Eric Brill of Microsoft Research published *Scaling to Very Very Large Corpora for Natural Language Disambiguation* at the annual meeting of the Association for Computational Linguistics. The work, they wrote, "was partially motivated by the desire to develop an improved grammar checker", and the question they set themselves was budgetary: given a fixed amount of time, would effort be better spent improving the algorithm or enlarging the training text?

They chose a task on which labelled data costs nothing — choosing between words that are commonly confused, such as *then* and *than* or *among* and *between* — because "the correct answer is surface apparent in any collection of reasonably well-edited text". They assembled a billion-word training corpus from news articles, scientific abstracts, government transcripts, literature and other prose, "three orders of magnitude greater than the largest training corpus previously used for this problem", and held out a million words of *Wall Street Journal* text as the test set. Four standard methods of the day — winnow, perceptron, naïve Bayes and a very simple memory-based learner — were trained at cutoffs along the way, and each point on the resulting curves is an average over ten confusion sets.

Note that the curves appear to be log-linear even out to one billion words.

"Log-linear" is easily misread. It does not mean that improvement accelerates. On Banko and Brill's chart, accuracy rises by roughly the same step each time the training text grows *tenfold*, so each further step needs ten times as much text: the reward is additive and the data requirement multiplicative. That shape was visible in 2001 with learners as simple as a single perceptron. The laws later fitted to neural language models describe a related but different shape: within the range fitted, and provided the other two inputs do not become the bottleneck, loss falls by roughly a constant *proportion* with each tenfold increase in model size, data or compute, so the absolute gains shrink as scale grows. Where a fit carries a non-zero floor, as the GPT-4 report's $L(C) = aC^b + c$ does, it is the loss above that floor that shrinks proportionally.

On the task they tested, the methods were far from their limits:

At least for the problem of confusable disambiguation, none of the learners tested is close to asymptoting in performance at the training corpus size commonly employed by the field.

Figure 4. Accuracy on two confusion sets, with a small labelled seed and with a billion labelled words

Banko and Brill, *ACL 2001*, Table 3, first and last rows. The table belongs to the paper's committee-based unsupervised-learning experiment: the first row uses only the million-word labelled seed corpus, the last is ordinary supervised training on the full billion words. Colour marks the training data, not the confusion set. The paper plots the four learners' learning curves without publishing the underlying values needed for a numerical comparison.

▼ Table view

Figure 4. Accuracy on two confusion sets, with a small labelled seed and with a billion labelled words

Confusion set and training data	Test accuracy
then / than - labelled seed only	1.0
then / than - 10^9 words, supervised	1.0
among / between - labelled seed only	0.8
among / between - 10^9 words, supervised	0.9

The authors drew the budgetary lesson in hedged terms — the results "suggest that we may want to reconsider the trade-off between spending time and money on algorithm development versus spending it on corpus development" — and then more boldly, proposing "that a logical next step for the research community would be to direct efforts towards increasing the size of annotated training collections, while deemphasizing the focus on comparing different learning techniques trained only on small training corpora". They also recorded the limits. "Such gains in accuracy, however, do not come for free": the learned representations grew with the data. And because very few problems come with free annotated data at that scale, they judged that the result "may have somewhat limited ramifications".

By 2009 the argument had become an essay title. Alon Halevy, Peter Norvig and Fernando Pereira, all of Google, published *The Unreasonable Effectiveness of Data* in *IEEE Intelligent Systems*, conceding that a trillion-word collection of web text was in some ways a step backwards in quality from small edited corpora, and concluding:

But the fact that it's a million times larger than the Brown Corpus outweighs these drawbacks.

In December 2017 the observation was generalised beyond language. Joel Hestness and colleagues at Baidu Research published *Deep Learning Scaling is Predictable, Empirically*, reporting "power-law generalization error scaling" across machine translation, language modelling, image processing and speech recognition, four domains. Their abstract records that the exponents are "yet to be explained by theoretical

work". The same abstract reports that model improvements "only shift the error but do not appear to affect the power-law exponent": in those experiments a better model appeared to move the line down without tilting it.

On 23 January 2020 Kaplan and colleagues posted *Scaling Laws for Neural Language Models*, which measured the regularity for transformer language models against three inputs at once — model size, dataset size and training compute — "with some trends spanning more than seven orders of magnitude".

5. The second idea: how to split a budget

A learning curve says what more of something buys; it does not say what to buy. That is the question of compute-optimal training: how a fixed training budget is divided.

A large training run is, in practice, a fixed budget of arithmetic. In the words of a DeepMind paper of March 2022, the budget "is often known in advance: how many accelerators are available and for how long we want to use them", and "it is typically only feasible to train these large models once". That budget can buy a larger model or a smaller model run over more text. For the dense transformers in this literature, training compute is roughly six times the number of parameters times the number of training tokens, so doubling both requires about four times the arithmetic. The two are different purchases, and the ratio between them has to be chosen before the run begins.

Kaplan's answer was lopsided. The paper's fitted relations put the compute-efficient model size at roughly the budget to the power 0.73 and the data requirement at roughly the power 0.27. These are growth rates, not shares of the budget: as compute rises, the optimal model grows much faster than its data. The abstract states the consequence plainly: "optimally compute-efficient training involves training very large models on a relatively modest amount of data and stopping significantly before convergence." DeepMind later restated the prescription in concrete terms: a tenfold increase in compute budget should buy a model 5.5 times larger and only 1.8 times as much training text. The GPT-3 paper's own contributions section records that Jared Kaplan and Sam McCandlish "applied scaling laws to help predict and guide model and data scaling decisions for the research". By DeepMind's own account the field acted on it:

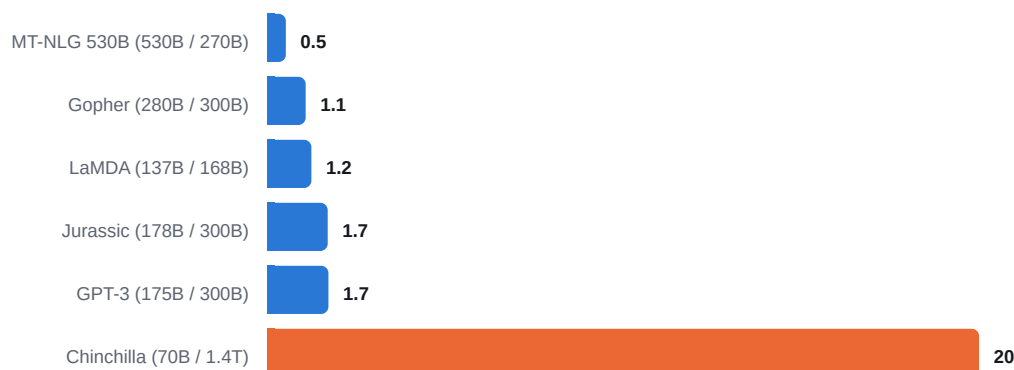
Following Kaplan et al. (2020) and the training setup of GPT-3 (Brown et al., 2020), many of the recently trained large models have been trained for approximately 300 billion tokens (Table 1), in line with the approach of predominantly increasing model size when increasing compute.

In March 2022 that DeepMind team, led by Jordan Hoffmann, tested the prescription by brute force, training "over 400 language models ranging from 70 million to over 16 billion parameters on 5 to 500 billion tokens", and reached a much more balanced recipe:

for compute-optimal training, the model size and the number of training tokens should be scaled equally: for every doubling of model size the number of training tokens should also be doubled.

They then built the model the new recipe implied. Chinchilla has 70 billion parameters and was trained on 1.4 trillion tokens with the same compute budget as Gopher, which has four times as many parameters and was trained on 300 billion tokens. The equality is the paper's own, from its full accounting of the runs; the six-times shortcut above is too coarse to reproduce it and puts the two within about a sixth of each other. Chinchilla, the paper reports, "uniformly and significantly outperforms Gopher (280B), GPT-3 (175B), Jurassic-1 (178B), and Megatron-Turing NLG (530B) on a large range of downstream evaluation tasks". That comparison is DeepMind's evaluation of two of its own models; what is open to outside scrutiny is the method, which three later groups took apart, rather than the head-to-head result.

Figure 5. Training tokens per parameter for the largest dense models of early 2022



Ratios are divisions of the parameter and token columns of Table 1 of Hoffmann et al., arXiv 2203.15556, which lists five of the largest dense transformer models of the time beside Chinchilla. The table's caption reads: 'Other than LaMDA (Thoppilan et al., 2022), most models are trained for approximately 300 billion tokens.' Chinchilla is in the second colour because it is the model built to the new recipe, not one of the five being assessed.

▼ Table view

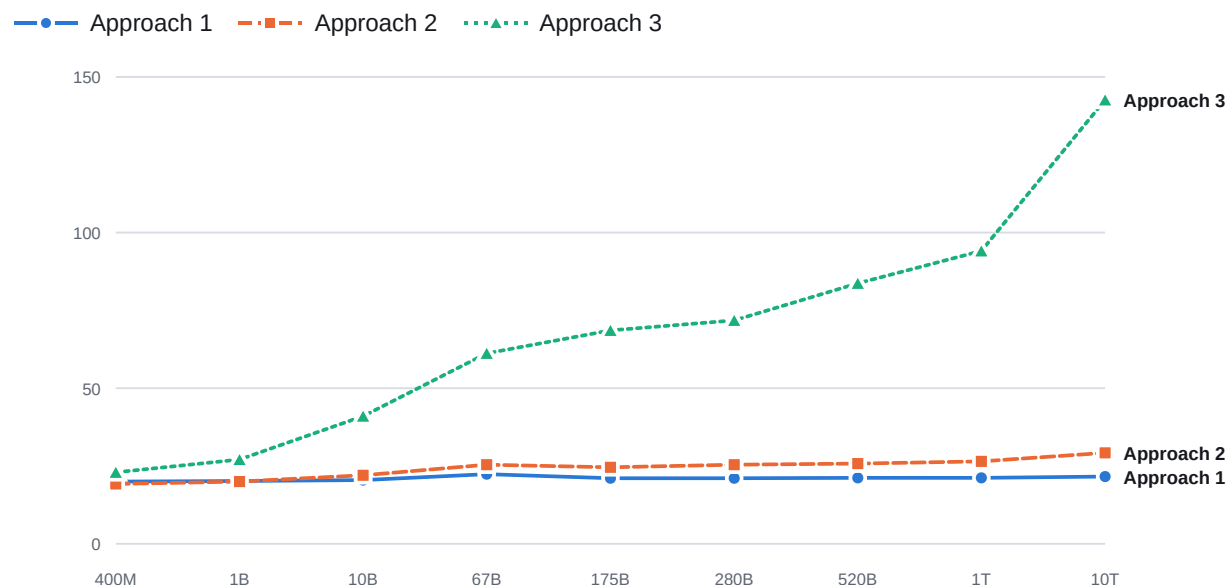
Figure 5. Training tokens per parameter for the largest dense models of early 2022

Model (parameters / training tokens)	Tokens per parameter
MT-NLG 530B (530B / 270B)	0.5
Gopher (280B / 300B)	1.1
LaMDA (137B / 168B)	1.2
Jurassic (178B / 300B)	1.7
GPT-3 (175B / 300B)	1.7
Chinchilla (70B / 1.4T)	20

That result made the insufficiency of a bare parameter count hard to ignore: a 70-billion-parameter model had beaten a 280-billion-parameter one on equal compute. The principal change was the split, though the paper also lists smaller departures from Gopher's recipe: the same dataset, MassiveText, but "a slightly different subset distribution" chosen to account for the larger number of training tokens; the AdamW optimiser in place of Adam; a slightly modified tokeniser; and a higher-precision copy of the weights in the optimiser state. The comparison is like-for-like on compute, not on every other setting. The count remained a real property of a model; it stopped being a verdict on it.

The folk version of Chinchilla holds that the correct ratio is twenty training tokens for every parameter. Twenty is exactly the ratio of Chinchilla's own run, 1.4 trillion tokens for 70 billion parameters, but the paper states no general ratio — its rule is the doubling relation, a doubling of training tokens for every doubling of model size — and although it describes its three estimation methods as giving comparable predictions, their numbers differ. Table 2 reports the exponent on model size as 0.50 for the first method, 0.49 for the second and 0.46 for the third, and the projected token counts diverge further than those exponents suggest.

Figure 6. Training tokens per parameter implied by each of Chinchilla's three estimation methods



Each value divides the projected token column by the parameter column of Hoffmann et al., arXiv 2203.15556: Table 3 for Approach 1 (minimum over training curves) and Table A3 for Approaches 2 (IsoFLOP profiles) and 3 (fitting a parametric loss function). Every value is a projection rather than a measurement, and the paper states no ratio. Approach 3 is the procedure a 2024 replication attempt found inconsistent with the other two.

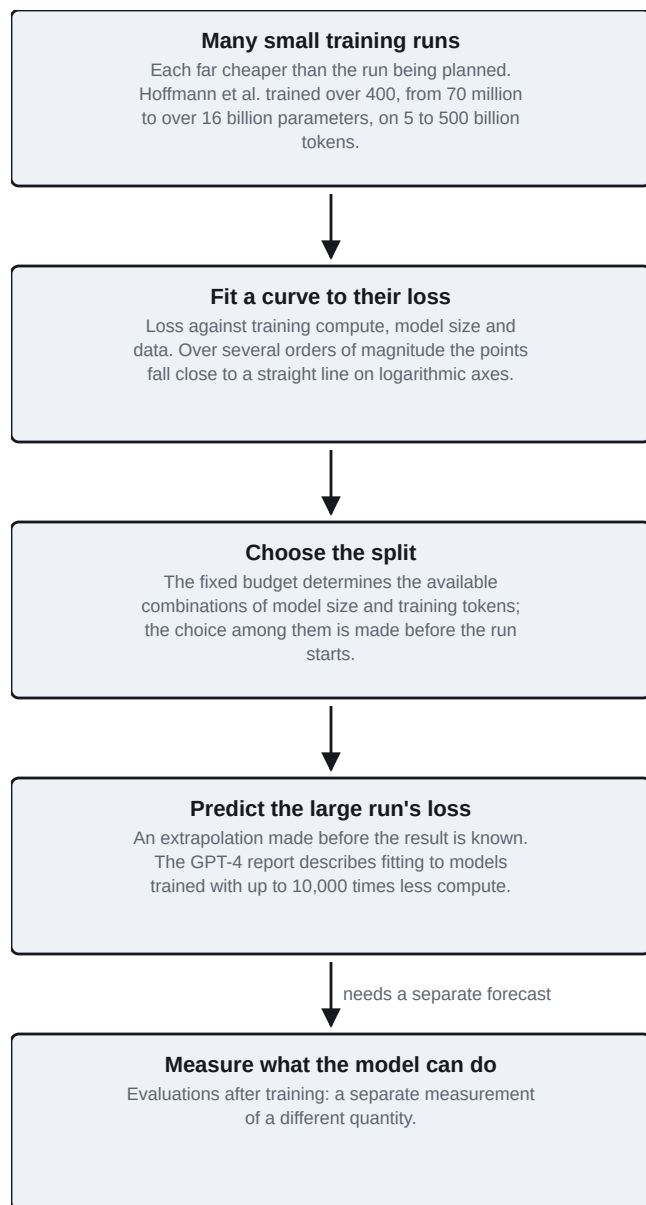
▼ Table view

Figure 6. Training tokens per parameter implied by each of Chinchilla's three estimation methods

Model size	Approach 1	Approach 2	Approach 3
400M	20	19.2	23
1B	20.2	20	27.1
10B	20.5	22	41
67B	22.4	25.4	61.2
175B	21.1	24.6	68.6
280B	21.1	25.4	71.8
520B	21.2	25.8	83.7
1T	21.2	26.5	94.1
10T	21.6	29.2	142.6

At 175 billion parameters the three methods project 3.7, 4.3 and 12.0 trillion tokens. The ratio of twenty matches the flattest of them. The paper's running text supplies a fourth figure for the same model size — a budget of 4.41×10^{24} operations and "over 4.2 trillion tokens" — so the document does not speak with one voice about this number, which is the reason the folk rule is not the paper's rule.

Figure 7. How a scaling law is used before a large run is paid for



Schematic. Loss and capability are different measurements. A loss curve forecasts loss; forecasting a capability needs a separate fit, which the GPT-4 report shows working for some measures and missing on others.

▼ Table view

Figure 7. How a scaling law is used before a large run is paid for — stages

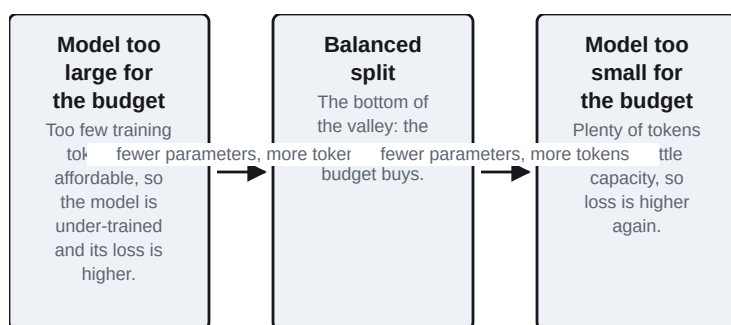
#	Stage	Note
1	Many small training runs	Each far cheaper than the run being planned. Hoffmann et al. trained over 400, from 70 million to over 16 billion parameters, on 5 to 500 billion tokens.
2	Fit a curve to their loss	Loss against training compute, model size and data. Over several orders of magnitude the points fall close to a straight line on logarithmic axes.
3	Choose the split	The fixed budget determines the available combinations of model size and training tokens; the choice among them is made before the run starts.

#	Stage	Note
4	Predict the large run's loss	An extrapolation made before the result is known. The GPT-4 report describes fitting to models trained with up to 10,000 times less compute.
5	Measure what the model can do	Evaluations after training: a separate measurement of a different quantity.

Figure 7. How a scaling law is used before a large run is paid for — connections

From	To	Label
Many small training runs	Fit a curve to their loss	
Fit a curve to their loss	Choose the split	
Choose the split	Predict the large run's loss	
Predict the large run's loss	Measure what the model can do	needs a separate forecast

Figure 8. One fixed budget, three ways to spend it



Schematic; no values are plotted. Hoffmann et al. held training compute fixed at nine budgets, varied model size, and report 'a clear valley in loss, meaning that for a given FLOP budget there is an optimal model to train' (arXiv 2203.15556, Figure 3). Under the dense approximation compute is about six times parameters times tokens, so each step along the row trades one for the other.

▼ Table view

Figure 8. One fixed budget, three ways to spend it — stages

#	Stage	Note
1	Model too large for the budget	Too few training tokens are affordable, so the model is under-trained and its loss is higher.
2	Balanced split	The bottom of the valley: the lowest loss this budget buys.
3	Model too small for the budget	Plenty of tokens but too little capacity, so loss is higher again.

Figure 8. One fixed budget, three ways to spend it — connections

From	To	Label
Model too large for the budget	Balanced split	fewer parameters, more tokens
Balanced split	Model too small for the budget	fewer parameters, more tokens

6. The clearest public test, and the misses printed beside it

Fitting a curve to runs already completed is retrospective and cheap. The stronger claim is that a curve fitted to small runs predicts a large one before the result is known, and a detailed public instance sits in the GPT-4 Technical Report of 15 March 2023, the document that declines to state the model's size.

The abstract describes the method: "A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4." Section 3 describes models "trained using $1,000\times - 10,000\times$ less compute" and the procedure for the headline test: a scaling law of the form $L(C) = aC^b + c$ was fitted to models trained with "at most 10,000x less compute than GPT-4", and the target was GPT-4's final loss on an internal codebase not included in the training set. In the report's words, "This prediction was made shortly after the run started, without use of any partial results. The fitted scaling law predicted GPT-4's final loss with high accuracy."

The same section extends the method to a quantity closer to usefulness — the mean log pass rate on a subset of the public HumanEval coding benchmark, predicted from models trained with "at most 1,000x less compute" — and describes its setup in detail. Predictions were registered before training completed, using only information available beforehand. The analysis was restricted to problems that every smaller model solved at least once given a large sample budget. All but the 15 hardest problems were sorted into six difficulty buckets by the performance of smaller models, and the figure the report shows covers the 23 problems of the third-easiest bucket. Predictions on the other five buckets "performed almost as well, the main exception being GPT-4 underperforming our predictions on the easiest bucket". The next sentence reads:

Certain capabilities remain hard to predict.

The example given is the Inverse Scaling Prize, a competition that collected tasks on which larger models do worse. On one of them, Hindsight Neglect, GPT-4 reversed that downward trend, a reversal the report likens to an earlier result by Jason Wei and colleagues. Extending the smaller models' trend on that task would have forecast the wrong direction.

OpenAI reports these results from its own training run. The loss target was a private codebase and the smaller models were the company's own, so the report does not give outsiders what they would need to repeat the exercise; what outsiders can audit is the published scaling literature.

7. Two correctives

Two postures are usually set against each other in public argument about scale, and the sharper correction to each came from within the research programme rather than from its critics.

The first posture is Richard Sutton's, in an essay dated 13 March 2019 and titled *The Bitter Lesson*:

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.

The essay's subject is the choice of method, not the purchase of parameters. Its stated mechanism is not the familiar chip-density rule itself — the 1965 observation that the number of components on an integrated circuit doubles at a regular interval — but, in the essay's own words, "its generalization of continued exponentially falling cost per unit of computation", and its recommendation is to prefer approaches that keep improving as computation grows. Chinchilla's smaller model does not contradict that preference; it contradicts only a cruder version of it that equates computation with parameters. A fixed-budget comparison does not test the essay's longer-run claim either way. Sutton has since put questions to the lesson himself. In a September 2025 interview with Dwarkesh Patel he asked of large language models, "Will they reach the limits of the data and be superseded by things that can get more data just from experience rather than from people?" Later in the same conversation, asked whether adding complexity would remain a false path in a world of billions of AI researchers, he set the lesson aside — "The bitter lesson, who cares about that?" — and went on:

That's an empirical observation about a particular period in history. 70 years in history, it doesn't necessarily have to apply to the next 70 years.

The second corrective is the audit, and it arrived three times inside eleven weeks in 2024.

On 15 April, Tamay Besiroglu, Ege Erdil, Matthew Barnett and Josh You of Epoch AI posted *Chinchilla Scaling: A replication attempt*, which addresses the third of Hoffmann's three estimation procedures. They report that its published estimates "are inconsistent with their first two estimation methods, fail at fitting the extracted data, and report implausibly narrow confidence intervals" — intervals so narrow, they calculate, that producing them would have required over 600,000 experiments where fewer than 500 were likely run. Their own rederivation by the same third approach gives results compatible with the other two.

On 12 June, Tim Pearce and Jinyeop Song posted *Reconciling Kaplan and Chinchilla Scaling Laws*, finding that "much of this discrepancy can be attributed to Kaplan counting non-embedding rather than total parameters, combined with their analysis being performed at small scale".

On 27 June, Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt and Yair Carmon posted *Resolving Discrepancies in Compute-Optimal Scaling of Language Models*, identifying "three factors causing the difference: last layer computational cost, warmup duration, and scale-dependent optimizer tuning", and reporting "excellent agreement" with the Chinchilla law once those are corrected.

Between them the audits locate the problems in parameter counting, training settings and statistical fitting. Epoch's authors report, for instance, that the optimiser Hoffmann's team used for its third fit "stopped before convergence due to a poor choice of loss scale", and record that the authors of the original paper confirmed the early stopping; that account appears in the replication's May 2024 revision rather than its first version. On DeepMind's own account, the prescription at issue had shaped how the largest models of the time were trained. The three groups converge from different directions; all are scaling-law researchers, and none includes an author of the papers in dispute.

8. What the ratio means now

Two things happened to "twenty tokens per parameter" after 2022, and they pull in opposite directions.

The first is that the phrase became ambiguous. Chinchilla studied dense models, in which, broadly, the whole network is used for every token, so total size and working size were the same number. Many of the models that now carry the largest parameter counts are sparse: they hold a large set of weights and activate a fraction of them for each token. DeepSeek's V4-Pro, described in its April 2026 report as 1.6 trillion parameters with 49 billion activated and trained on 33 trillion tokens, sits at 20.6 tokens per total parameter, almost exactly Chinchilla's figure; per activated parameter, a rough proxy for the arithmetic done per token, it sits near 670. The figure near twenty counts total parameters, and the resemblance does not show that this sparse model was trained to Chinchilla's rule.

The second is that the objective changed. Liquid AI's LFM2.5-350M, a 350-million-parameter model built to run on devices such as phones, records a training budget of 28 trillion tokens: about 80,000 tokens per parameter, some four thousand times the Chinchilla figure. Compute-optimal training minimises loss for a fixed training budget. A model intended for a phone is judged by its size when it runs, and for such a model it can make sense to spend far more training compute than the Chinchilla rule would allot in order to bring a smaller model to a given quality. *Test-Time Scaling Makes Overtraining Compute-Optimal* (arXiv 2604.01411, 1 April 2026) formalises a version of that argument, reporting that "when accounting for inference cost, optimal pretraining decisions shift radically into the overtraining regime".

Figure 9. Training tokens per parameter, as models are trained in 2026

20.6 / 673

DeepSeek-V4-Pro (Apr 2026), per total / per active parameter
1.6 trillion parameters, 49 billion activated per token, 33 trillion training tokens

~80,000

LFM2.5-350M, a model built to run on devices
350 million parameters against a stated training budget of 28 trillion tokens

Each ratio divides two figures the developer states in its own document: the DeepSeek-V4 report (dated 26 April 2026; arXiv 2606.19348), whose pre-training section gives 33 trillion tokens for V4-Pro where its abstract gives more than 32 trillion for the two models together, and Liquid AI's LFM2.5-350M model card. No developer states a ratio. Both the token totals and the parameter counts here are the developers' own statements about their own training runs, unaudited and not checkable from outside. Chinchilla's own figure, for a dense model in which total and active size are the same number, was 20.

▼ Table view

Figure 9. Training tokens per parameter, as models are trained in 2026

Measure	Value
DeepSeek-V4-Pro (Apr 2026), per total / per active parameter	20.6 / 673
LFM2.5-350M, a model built to run on devices	~80,000

The method itself is plainly still in production. Meta's Llama 3 report of July 2024 describes the same exercise on its own data, developing scaling laws with reference to both papers and extrapolating them to 3.8×10^{25} floating-point operations, which "suggests training a 402B parameter model on 16.55T tokens" — about 41 tokens per parameter.

Nor has the literature thinned. An arXiv search on 16 September 2026 for the exact phrase "scaling law" in the abstract, restricted to computer science and to a most recent submission date in 2026, returned 620 results — a crude count, since it includes revisions of older papers and misses work that fits curves without using the phrase, but not the signature of an abandoned idea. The same search run without quotation marks returns 1,167, because arXiv then matches the two words separately: in a sample of the first fifty of those, forty-two do not contain the phrase at all. What that literature does has changed. Two 2026 papers on data-constrained training open by naming an assumption the original fits made. *Prescriptive Scaling Laws for Data Constrained Training* (arXiv 2605.01640, 2 May 2026): "The widely adopted Chinchilla scaling law assumes every training token is unique. This limits its ability to guide pretraining decisions in data-constrained regimes." *Practical Scaling Laws* (arXiv 2605.09189, 9 May 2026) says the older laws were "calibrated for a single regime: data-rich, single-epoch pretraining". Neither paper says the curves stopped working. Each extends the earlier fits to training on repeated data, a regime in which the original assumption of fresh tokens does not hold, and each replaces the fit rather than the method.

9. The limits of public training text

Epoch AI's estimate of a ceiling on training data is a conditional forecast. In "*Will we run out of data? Limits of LLM scaling based on human-generated data*", Pablo Villalobos and colleagues at Epoch AI, posted in October 2022 and last revised in June 2024, put it this way:

Our findings indicate that if current LLM development trends continue, models will be trained on datasets roughly equal in size to the available stock of public human text data between 2026 and 2032, or slightly earlier if models are overtrained.

The forecast concerns only public human-generated text, spans six years and holds only if current trends continue. Epoch's accompanying page, dated 6 June 2024, puts the effective stock at about 300 trillion tokens with a 90% confidence interval from 100 trillion to 1,000 trillion, and gives the window as an 80% interval. The authors argue that synthetic data generation, transfer learning from data-rich domains and improvements in data efficiency "might support further progress", which describes what would have to change rather than predicting that progress stops.

2026 is the first year of the window, so not reaching the projected limit this year would not refute the forecast. The arXiv listing shows no version after 4 June 2024 and the Epoch page shows no update notice, while the same organisation's model and data-centre datasets carry September 2026 update dates.

Some of the arithmetic beneath the constraint has been measured rather than assumed. In *Scaling Data-Constrained Language Models*, Niklas Muennighoff and colleagues ran experiments with up to 900 billion training tokens and 9-billion-parameter models, and report that for a fixed compute budget "training with up to 4 epochs of repeated data yields negligible changes to loss compared to having unique data". Their fitted law then projects that with further repetition the value of adding compute approaches zero — a projection from the fit, not a measurement at those scales.

10. The unit of account in public

If a parameter count no longer ranks models and is no longer published for the largest closed ones, something else carries the weight in public documents, and two candidates are visible.

The first is law. Article 51(2) of Regulation (EU) 2024/1689, the AI Act, provides that a general-purpose AI model

shall be presumed to have high impact capabilities pursuant to paragraph 1, point (a), when the cumulative amount of computation used for its training measured in floating point operations is greater than 10^{25} .

The threshold is a quantity of arithmetic; parameters are not mentioned. Article 51(3) empowers the Commission to adopt delegated acts amending the thresholds "in light of evolving technological developments, such as algorithmic improvements or increased hardware efficiency, when necessary, for these thresholds to reflect the state of the art" — a duty conditioned on necessity, which the Commission's own guidelines describe as being "empowered" to act. Either way the statute acknowledges that the number will go stale and supplies a procedure for moving it.

The Commission has explained its choice of unit. Its non-binding guidelines on the scope of the obligations for providers of general-purpose AI models, C(2025) 7719 of 19 November 2025, set an indicative criterion for a model to count as general-purpose at all — training compute above 10^{23} floating-point operations together with the ability to generate language, images or video, subject to exceptions — and justify the unit as follows:

Training compute has the advantage of combining number of parameters and number of training examples into a single number that is reasonably straightforward for providers to estimate. This number is typically proportional to the number obtained by multiplying these two numbers, allowing a single threshold to be set rather than separate thresholds for model size and training data size. While training compute is an imperfect proxy for generality and capabilities, the Commission considers setting an indicative criterion which includes a training compute threshold to be the most suitable approach at present. Nevertheless, the Commission's approach may change in the future as technology and the market evolve.

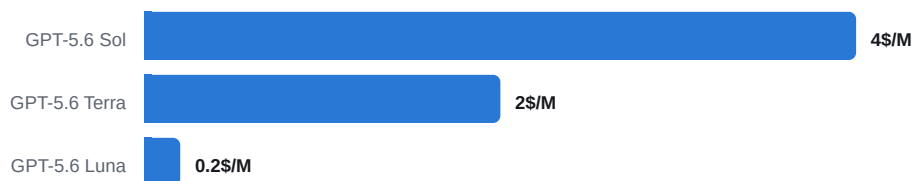
In that paragraph the product of model size and data, the quantity the scaling-law papers budget, becomes an administrative measure, described by the regulator itself as an imperfect proxy and as the most suitable approach "at present".

The threshold has not moved. The AI Act was amended by Regulation (EU) 2026/1744 of 8 July 2026, published in the Official Journal on 24 July 2026 and known as the Digital Omnibus on AI. The consolidated text carries 69 change markers, and the amendment moves the obligations for high-risk systems classified under Article 6(2) from 2 August 2026 to 2 December 2027, and those for systems classified under Article 6(1) from 2 August 2027 to 2 August 2028, citing "the delayed availability of standards, common specifications, and alternative guidance and the delayed establishment of national

competent authorities". Articles 51 to 55 carry no change marker. The amendment left the compute threshold where it was, and the Commission's November 2025 guidelines still describe the 10^{25} figure as the current one and the delegated-act power as something to be exercised in future.

The second candidate is commercial, and it appears on the developers' own model pages. OpenAI's model documentation lists eight fields for its flagship GPT-6 Astra: model identifier, reasoning settings, input price, output price, maximum output, context window, knowledge cutoff and available tools. None of the eight gives the model's size or its training compute. The tier ladder has changed units too. The three tiers of the GPT-5.6 family list identical context windows of 1.05 million tokens, identical maximum outputs of 128,000 tokens, identical knowledge cutoffs of 16 February 2026, identical reasoning settings and identical tools; among the fields describing what the model will do, only the price separates them. The one other difference is clerical: the top tier also carries a short alias. In the GPT-3 paper of May 2020 the family was presented as a ladder of eight model sizes, from 125 million to 175 billion parameters.

Figure 10. What separates one tier from the next in the GPT-5.6 family



Source: OpenAI model documentation (developers.openai.com/api/docs/models), read 13 September 2026 and re-read 16 September, unchanged. Output prices are \$20, \$12 and \$1.20 per million tokens respectively. All three tiers list the same context window (1.05M tokens), maximum output (128K tokens), knowledge cutoff (16 February 2026), reasoning settings and tools; the top tier additionally carries the alias gpt-5.6. No field describing the model itself is published for any of them.

▼ Table view

Figure 10. What separates one tier from the next in the GPT-5.6 family

Tier (same listed limits, reasoning settings and tools)	Input price per million tokens
GPT-5.6 Sol	4\$/M
GPT-5.6 Terra	2\$/M
GPT-5.6 Luna	0.2\$/M

Anthropic's model comparison table has the same shape — comparative latency, pricing, identifier, thinking mode, default effort, context window, maximum output and knowledge cutoff — and no size field. The shift says nothing about whether scaling laws hold; it shows that where a buyer once compared numbers describing what a model is, the published fields now describe what it costs and what it will accept.

11. Four dated markers

Two empty cells. Epoch AI's Notable AI Models file records, for GPT-6 Astra, published on 3 September 2026, no parameter figure and no training-compute figure. The same is true of Claude Opus 5, Gemini 3.8 Flash and Muse Spark 1.3. A newly filled cell for a flagship model from OpenAI, Anthropic, Google

DeepMind or Meta would indicate either a developer's disclosure or an outside estimate that Epoch considers defensible.

The threshold, and the marker beside it. The consolidated text of Regulation (EU) 2024/1689 lists one amending act, Regulation (EU) 2026/1744, and carries 69 change markers, none of them inside Articles 51 to 55. A delegated act adopted under Article 51(3) would appear in the Official Journal and is the plainest signal to watch; a change marker arriving inside Article 51 is the same event seen in the consolidated text, and nothing in that text shows the power exercised. Epoch AI's trends dashboard, updated 5 February 2026, estimates the growth of frontier training compute at $5\times$ a year since 2020; if that continues, a threshold left where it is sits progressively further below frontier training budgets, which is why the power to move it matters.

A number in the document whose silence is the subject. OpenAI's deployment-safety page for GPT-6 Astra carries a section headed *Model Data and Training* in which the only digit is a footnote marker. A quantity appearing there would be a disclosure by the developer, in the developer's own document, and needs no one else's judgement to interpret — which a filled cell in a third party's dataset does, since that may record an outside estimate instead.

A fit tested outside its original regime. Two May 2026 papers, Lovelace and colleagues' *Prescriptive Scaling Laws for Data Constrained Training* (arXiv 2605.01640) and Bryant and Liu's *Practical Scaling Laws* (arXiv 2605.09189), fit scaling laws for training on repeated data. The informative next document is a training report that states the distinct data available, the tokens processed, and a loss predicted in advance beside the loss observed.

12. Common readings the record does not support

Reading	What the documents say
Scaling laws are laws of nature, so improvement is guaranteed.	They are fitted regularities. Kaplan et al. record that they have no solid theoretical understanding of them, that the trends must eventually level off because natural language has non-zero entropy, and that two of their own fits imply the laws must break down beyond a certain point.
The curves stopped working, which is why nobody quotes them.	A detailed prospective public test — the GPT-4 report's loss prediction from models trained with 1,000 to 10,000 times less compute — is reported as accurate, and developers were still fitting such curves in 2026. What stopped being published was a model's size.
0.73 means 73% of the budget goes on parameters.	It is a growth exponent: in Kaplan et al.'s fit, compute-efficient model size grows roughly as compute to the power 0.73. It is not a spending share.
More compute means more parameters.	Extra training compute can buy a larger model, more tokens, or both; for dense models, doubling both takes roughly four times the compute.
A bigger model is a better model.	On equal compute a 70-billion-parameter model beat a 280-billion-parameter one; the principal change was how the budget was split between size and tokens.
Chinchilla established twenty training tokens per parameter.	Twenty is the ratio of Chinchilla's own run. The paper's rule is a doubling relation, and its three estimation methods project 3.7, 4.3 and 12.0 trillion tokens at 175 billion parameters.

Reading	What the documents say
"Log-linear" means progress accelerates.	It describes the axes. On Banko and Brill's chart each tenfold multiplication of the data bought about the same gain in accuracy; later power laws have loss shrinking by a steady proportion per tenfold. Neither describes speed over time.
The curve predicts what a model will be able to do.	A loss curve forecasts loss; predicting a capability needs a fit to a measure of that capability. The GPT-4 report's separate fit to a coding benchmark held for most difficulty buckets and fell short on the easiest, and the same section records a task on which GPT-4 reversed the smaller models' trend.
The world has run out of training data.	The forecast in circulation gives a window from 2026 to 2032, conditional on current trends continuing and restricted to public human-generated text. Its authors name synthetic data, transfer and data efficiency as routes if the constraint binds.
Repeating data solves the shortage.	In experiments running to 900 billion training tokens, up to four passes over repeated data changed loss negligibly against fresh data; the fitted law projects that with further repetition the value of adding compute approaches zero.
Kaplan was simply wrong and Chinchilla corrected the error.	Two 2024 re-examinations attribute much of the gap to parameter counting, small-scale analysis, last-layer cost, warmup duration and optimiser tuning; a third found that the published estimates from one of Chinchilla's own three methods did not fit the data.

13. What a scaling curve is for

Scaling laws are empirical: the curves are measured, not derived; in 2020 the paper that popularised them opened its caveats by saying that no solid theoretical understanding existed for any of them, and later theory explains them only in part.

In training, their use is a purchasing decision: how to divide a budget, fixed before the run, between a larger model and more text. On the public evidence — Chinchilla's more than four hundred training runs, the GPT-4 report's loss forecast and the 2024 audits of both founding papers — they have served that purpose well, and they were still being used for it in 2026.

What the training-budget fits forecast is loss, a measure of how surprised a model is by the next word of text. Whether a lower loss buys the capability anyone wanted is a separate question with a separate measurement. Some capabilities can be forecast with fits of their own, and the report that showed such a forecast working concedes, in the sentence after its own miss, that others are still hard to foresee. A claim that a system is enormous therefore invites two questions the parameter count cannot answer: how the budget was split, and what was measured afterwards. A scaling forecast is informative only when its target, the resources it counts, the conditions of training and the evidence behind it are all named.

Sources

Source	What it is	Date
Corinna Cortes, L. D. Jackel, Sara A. Solla, Vladimir Vapnik and John S. Denker, <i>Learning Curves: Asymptotic Values and Rate of Convergence</i>	NIPS 6; proceedings.neurips.cc	1993
Michele Banko and Eric Brill, <i>Scaling to Very Very Large Corpora for Natural Language Disambiguation</i>	ACL 2001; aclanthology.org/P01-1005	2001

Source	What it is	Date
Alon Halevy, Peter Norvig and Fernando Pereira, <i>The Unreasonable Effectiveness of Data</i>	<i>IEEE Intelligent Systems</i> , Expert Opinion	March/April 2009
Joel Hestness et al. (Baidu Research), <i>Deep Learning Scaling is Predictable, Empirically</i>	arXiv 1712.00409	v1, 1 December 2017
Richard Sutton, <i>The Bitter Lesson</i>	Essay, incompleteideas.net	13 March 2019
Jared Kaplan, Sam McCandlish et al., <i>Scaling Laws for Neural Language Models</i>	arXiv 2001.08361	v1, 23 January 2020
Tom B. Brown et al. (OpenAI), <i>Language Models are Few-Shot Learners</i>	arXiv 2005.14165; abstract, Table 2.1 and contributions	v1, 28 May 2020
Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee and Utkarsh Sharma, <i>Explaining Neural Scaling Laws</i>	arXiv 2102.06701	v1, 12 February 2021
Jack W. Rae et al. (DeepMind), <i>Scaling Language Models: Methods, Analysis & Insights from Training Gopher</i>	arXiv 2112.11446	v1, 8 December 2021
Jordan Hoffmann et al. (DeepMind), <i>Training Compute-Optimal Large Language Models</i>	arXiv 2203.15556; sections 1 and 3, Tables 1, 2, 3 and A3	v1, 29 March 2022
Aakanksha Chowdhery et al. (Google), <i>PaLM: Scaling Language Modeling with Pathways</i>	arXiv 2204.02311	v1, 5 April 2022
Pablo Villalobos et al. (Epoch AI), <i>Will we run out of data? Limits of LLM scaling based on human-generated data</i> , and the accompanying Epoch page	arXiv 2211.04325; epoch.ai	v1 26 October 2022, v2 4 June 2024; page 6 June 2024
OpenAI, <i>GPT-4 Technical Report</i>	arXiv 2303.08774; abstract, sections 2 and 3	v1, 15 March 2023
Niklas Muennighoff et al., <i>Scaling Data-Constrained Language Models</i>	arXiv 2305.16264	v1, 25 May 2023
Tamay Besiroglu, Ege Erdil, Matthew Barnett and Josh You (Epoch AI), <i>Chinchilla Scaling: A replication attempt</i>	arXiv 2404.10102; v2 carries the optimiser diagnosis	v1, 15 April 2024; v2 read
Tim Pearce and Jinyeop Song, <i>Reconciling Kaplan and Chinchilla Scaling Laws</i>	arXiv 2406.12907	v1, 12 June 2024
Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt and Yair Carmon, <i>Resolving Discrepancies in Compute-Optimal Scaling of Language Models</i>	arXiv 2406.19146	v1, 27 June 2024
Meta, <i>The Llama 3 Herd of Models</i>	arXiv 2407.21783	v1, 31 July 2024
DeepSeek-AI, <i>DeepSeek-V3 Technical Report</i> and <i>DeepSeek-V4</i>	arXiv 2412.19437; arXiv 2606.19348	27 December 2024; 26 April 2026
OpenAI, gpt-oss, GPT-5 and GPT-6 Astra pages, Deployment Safety Hub	deploymentsafety.openai.com	5 August 2025; 7 August 2025; 3 September 2026
European Commission, <i>Guidelines on the scope of the obligations for providers of general-purpose AI models</i> , C(2025) 7719 final	ec.europa.eu	19 November 2025
Google DeepMind, Gemini 3 Pro model card	deepmind.google	released November 2025, last updated May 2026
Roberts et al., <i>Test-Time Scaling Makes Overtraining Compute-Optimal</i>	arXiv 2604.01411	v1, 1 April 2026

Source	What it is	Date
Lovelace et al., <i>Prescriptive Scaling Laws for Data Constrained Training</i> ; Bryant and Liu, <i>Practical Scaling Laws</i>	arXiv 2605.01640; arXiv 2605.09189	2 May 2026; 9 May 2026
Liquid AI, LFM2.5-350M model card	huggingface.co/LiquidAI	read 13 September 2026
Regulation (EU) 2026/1744 (Digital Omnibus on AI), and the consolidated text of Regulation (EU) 2024/1689	eur-lex.europa.eu ; CELEX 02024R1689, consolidation of 27 July 2026	8 July 2026; OJ 24 July 2026
Moonshot AI, Kimi K3 model card; Z.ai, GLM-5.3 repository	huggingface.co/moonshotai ; huggingface.co/zai-org	read 16 September 2026
Richard Sutton, interview with Dwarkesh Patel	dwarkesh.com/p/richard-sutton ; transcript at 00:09:41 and 00:49:51	26 September 2025
Anthropic, Claude Opus 5 system card, and model comparison table	anthropic.com ; docs.claude.com	24 July 2026; read 13 September 2026
Epoch AI, Notable AI Models dataset, and Trends dashboard	epoch.ai/data/notable_ai_models.csv ; epoch.ai/trends	updated 11 September 2026; updated 5 February 2026
OpenAI, model documentation	developers.openai.com/api/docs/models	read 13 September 2026