

# Training Compute-Optimal Large Language Models

Jordan Hoffmann\*, Sebastian Borgeaud\*, Arthur Mensch\*, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals and Laurent Sifre\*

\*Equal contributions

We investigate the optimal model size and number of tokens for training a transformer language model under a given compute budget. We find that current large language models are significantly under-trained, a consequence of the recent focus on scaling language models whilst keeping the amount of training data constant. By training over 400 language models ranging from 70 million to over 16 billion parameters on 5 to 500 billion tokens, we find that 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. We test this hypothesis by training a predicted compute-optimal model, *Chinchilla*, that uses the same compute budget as *Gopher* but with 70B parameters and 4× more more data. *Chinchilla* 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. This also means that *Chinchilla* uses substantially less compute for fine-tuning and inference, greatly facilitating downstream usage. As a highlight, *Chinchilla* reaches a state-of-the-art average accuracy of 67.5% on the MMLU benchmark, greater than a 7% improvement over *Gopher*.

## 1. Introduction

Recently a series of *Large Language Models* (LLMs) have been introduced (Brown et al., 2020; Lieber et al., 2021; Rae et al., 2021; Smith et al., 2022; Thoppilan et al., 2022), with the largest dense language models now having over 500 billion parameters. These large autoregressive transformers (Vaswani et al., 2017) have demonstrated impressive performance on many tasks using a variety of evaluation protocols such as zero-shot, few-shot, and fine-tuning.

The compute and energy cost for training large language models is substantial (Rae et al., 2021; Thoppilan et al., 2022) and rises with increasing model size. In practice, the allocated training compute budget is often known in advance: how many accelerators are available and for how long we want to use them. Since it is typically only feasible to train these large models once, accurately estimating the best model hyperparameters for a given compute budget is critical (Tay et al., 2021).

Kaplan et al. (2020) showed that there is a power law relationship between the number of parameters in an autoregressive language model (LM) and its performance. As a result, the field has been training larger and larger models, expecting performance improvements. One notable conclusion in Kaplan et al. (2020) is that large models should not be trained to their lowest possible loss to be compute optimal. Whilst we reach the same conclusion, we estimate that large models should be trained for many more training tokens than recommended by the authors. Specifically, given a 10× increase computational budget, they suggests that the size of the model should increase 5.5× while the number of training tokens should only increase 1.8×. Instead, we find that model size and the number of training tokens should be scaled in equal proportions.

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.

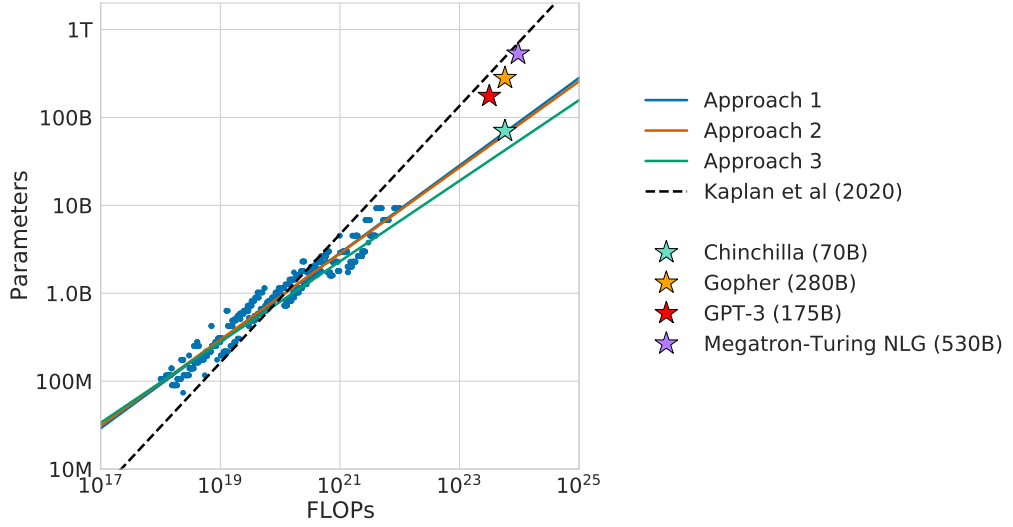


Figure 1 | **Overlaid predictions.** We overlay the predictions from our three different approaches, along with projections from Kaplan et al. (2020). We find that all three methods predict that current large models should be substantially smaller and therefore trained much longer than is currently done. In Figure A3, we show the results with the predicted optimal tokens plotted against the optimal number of parameters for fixed FLOP budgets. **Chinchilla outperforms Gopher and the other large models (see Section 4.2).**

In this work, we revisit the question: *Given a fixed FLOPs budget,<sup>1</sup> how should one trade-off model size and the number of training tokens?* To answer this question, we model the final pre-training loss<sup>2</sup>  $L(N, D)$  as a function of the number of model parameters  $N$ , and the number of training tokens,  $D$ . Since the computational budget  $C$  is a deterministic function  $\text{FLOPs}(N, D)$  of the number of seen training tokens and model parameters, we are interested in minimizing  $L$  under the constraint  $\text{FLOPs}(N, D) = C$ :

$$N_{\text{opt}}(C), D_{\text{opt}}(C) = \underset{N, D \text{ s.t. } \text{FLOPs}(N, D) = C}{\text{argmin}} L(N, D). \quad (1)$$

The functions  $N_{\text{opt}}(C)$ , and  $D_{\text{opt}}(C)$  describe the optimal allocation of a computational budget  $C$ . We empirically estimate these functions based on the losses of over 400 models, ranging from under 70M to over 16B parameters, and trained on 5B to over 400B tokens – with each model configuration trained for several different training horizons. Our approach leads to considerably different results than that of Kaplan et al. (2020). We highlight our results in Figure 1 and how our approaches differ in Section 2.

Based on our estimated compute-optimal frontier, we predict that for the compute budget used to train *Gopher*, an optimal model should be 4 times smaller, while being training on 4 times more tokens. We verify this by training a more *compute-optimal* 70B model, called *Chinchilla*, on 1.4 trillion tokens. Not only does *Chinchilla* outperform its much larger counterpart, *Gopher*, but its reduced model size reduces inference cost considerably and greatly facilitates downstream uses on smaller hardware. The energy cost of a large language model is amortized through its usage for inference and fine-tuning. The benefits of a more optimally trained smaller model, therefore, extend beyond the immediate benefits of its improved performance.

<sup>1</sup>For example, knowing the number of accelerators and a target training duration.

<sup>2</sup>For simplicity, we perform our analysis on the smoothed training loss which is an unbiased estimate of the test loss, as we are in the infinite data regime (the number of training tokens is less than the number of tokens in the entire corpus).

Table 1 | **Current LLMs.** We show five of the current largest dense transformer models, their size, and the number of training tokens. Other than LaMDA (Thoppilan et al., 2022), most models are trained for approximately 300 billion tokens. We introduce *Chinchilla*, a substantially smaller model, trained for much longer than 300B tokens.

Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion
<i>Chinchilla</i>	70 Billion	1.4 Trillion

## 2. Related Work

**Large language models.** A variety of large language models have been introduced in the last few years. These include both dense transformer models (Brown et al., 2020; Lieber et al., 2021; Rae et al., 2021; Smith et al., 2022; Thoppilan et al., 2022) and mixture-of-expert (MoE) models (Du et al., 2021; Fedus et al., 2021; Zoph et al., 2022). The largest dense transformers have passed 500 billion parameters (Smith et al., 2022). The drive to train larger and larger models is clear—so far increasing the size of language models has been responsible for improving the state-of-the-art in many language modelling tasks. Nonetheless, large language models face several challenges, including their overwhelming computational requirements (the cost of training and inference increase with model size) (Rae et al., 2021; Thoppilan et al., 2022) and the need for acquiring more high-quality training data. In fact, in this work we find that larger, high quality datasets will play a key role in any further scaling of language models.

**Modelling the scaling behavior.** Understanding the scaling behaviour of language models and their transfer properties has been important in the development of recent large models (Hernandez et al., 2021; Kaplan et al., 2020). Kaplan et al. (2020) first showed a predictable relationship between model size and loss over many orders of magnitude. The authors investigate the question of choosing the optimal model size to train for a given compute budget. Similar to us, they address this question by training various models. Our work differs from Kaplan et al. (2020) in several important ways. First, the authors use a fixed number of training tokens and learning rate schedule for all models; this prevents them from modelling the impact of these hyperparameters on the loss. In contrast, we find that setting the learning rate schedule to approximately match the number of training tokens results in the best final loss regardless of model size—see Figure A1. For a fixed learning rate cosine schedule to 130B tokens, the intermediate loss estimates (for  $D' \ll 130B$ ) are therefore overestimates of the loss of a model trained with a schedule length matching  $D'$ . Using these intermediate losses results in underestimating the effectiveness of training models on less data than 130B tokens, and eventually contributes to the conclusion that model size should increase faster than training data size as compute budget increases. In contrast, our analysis predicts that both quantities should scale at roughly the same rate. Secondly, we include models with up to 16B parameters, as we observe that there is slight curvature in the FLOP-loss frontier (see Appendix E)—in fact, the majority of the models used in our analysis have more than 500 million parameters, in contrast the majority of runs in Kaplan et al. (2020) are significantly smaller—many being less than 100M parameters.

Recently, Clark et al. (2022) specifically looked in to the scaling properties of Mixture of Expert

language models, showing that the scaling with number of experts diminishes as the model size increases—their approach models the loss as a function of two variables: the model size and the number of experts. However, the analysis is done with a fixed number of training tokens, as in [Kaplan et al. \(2020\)](#), potentially underestimating the improvements of branching.

**Estimating hyperparameters for large models.** The model size and the number of training tokens are not the only two parameters to choose when selecting a language model and a procedure to train it. Other important factors include learning rate, learning rate schedule, batch size, optimiser, and width-to-depth ratio. In this work, we focus on model size and the number of training steps, and we rely on existing work and provided experimental heuristics to determine the other necessary hyperparameters. [Yang et al. \(2021\)](#) investigates how to choose a variety of these parameters for training an autoregressive transformer, including the learning rate and batch size. [McCandlish et al. \(2018\)](#) finds only a weak dependence between optimal batch size and model size. [Shallue et al. \(2018\)](#); [Zhang et al. \(2019\)](#) suggest that using larger batch-sizes than those we use is possible. [Levine et al. \(2020\)](#) investigates the optimal depth-to-width ratio for a variety of standard model sizes. We use slightly less deep models than proposed as this translates to better wall-clock performance on our hardware.

**Improved model architectures.** Recently, various promising alternatives to traditional dense transformers have been proposed. For example, through the use of conditional computation large MoE models like the 1.7 trillion parameter Switch transformer ([Fedus et al., 2021](#)), the 1.2 Trillion parameter GLaM model ([Du et al., 2021](#)), and others ([Artetxe et al., 2021](#); [Zoph et al., 2022](#)) are able to provide a large effective model size despite using relatively fewer training and inference FLOPs. However, for very large models the computational benefits of routed models seems to diminish ([Clark et al., 2022](#)). An orthogonal approach to improving language models is to augment transformers with explicit retrieval mechanisms, as done by [Borgeaud et al. \(2021\)](#); [Guu et al. \(2020\)](#); [Lewis et al. \(2020\)](#). This approach effectively increases the number of data tokens seen during training (by a factor of  $\sim 10$  in [Borgeaud et al. \(2021\)](#)). This suggests that the performance of language models may be more dependant on the size of the training data than previously thought.

### 3. Estimating the optimal parameter/training tokens allocation

We present three different approaches to answer the question driving our research: *Given a fixed FLOPs budget, how should one trade-off model size and the number of training tokens?* In all three cases we start by training a range of models varying both model size and the number of training tokens and use the resulting training curves to fit an empirical estimator of how they should scale. We assume a power-law relationship between compute and model size as done in [Clark et al. \(2022\)](#); [Kaplan et al. \(2020\)](#), though future work may want to include potential curvature in this relationship for large model sizes. The resulting predictions are similar for all three methods and suggest that parameter count and number of training tokens should be increased equally with more compute<sup>3</sup>—with proportions reported in [Table 2](#). This is in clear contrast to previous work on this topic and warrants further investigation.

---

<sup>3</sup>We compute FLOPs as described in [Appendix F](#).

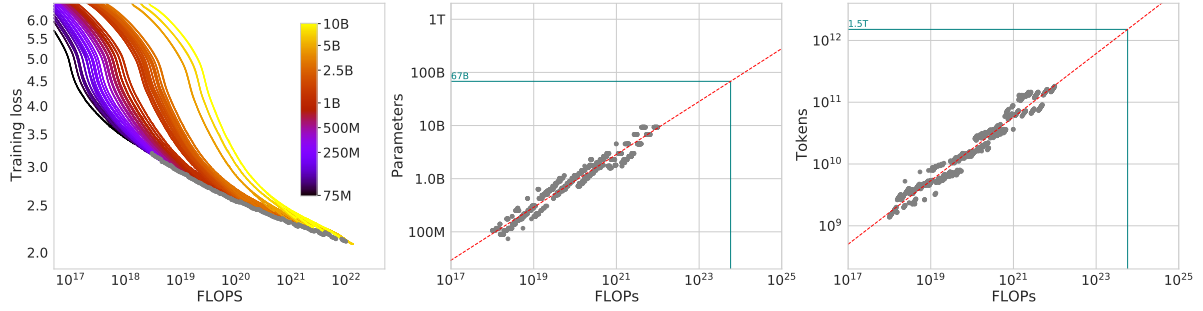


Figure 2 | **Training curve envelope.** On the **left** we show all of our different runs. We launched a range of model sizes going from 70M to 10B, each for four different cosine cycle lengths. From these curves, we extracted the envelope of minimal loss per FLOP, and we used these points to estimate the optimal model size (**center**) for a given compute budget and the optimal number of training tokens (**right**). In green, we show projections of optimal model size and training token count based on the number of FLOPs used to train *Gopher* ( $5.76 \times 10^{23}$ ).

### 3.1. Approach 1: Fix model sizes and vary number of training tokens

In our first approach we vary the number of training steps for a fixed family of models (ranging from 70M to over 10B parameters), training each model for 4 different number of training sequences. From these runs, we are able to directly extract an estimate of the minimum loss achieved for a given number of training FLOPs. Training details for this approach can be found in [Appendix D](#).

For each parameter count  $N$  we train 4 different models, decaying the learning rate by a factor of  $10\times$  over a horizon (measured in number of training tokens) that ranges by a factor of  $16\times$ . Then, for each run, we smooth and then interpolate the training loss curve. From this, we obtain a continuous mapping from FLOP count to training loss for each run. Then, for each FLOP count, we determine which run achieves the lowest loss. Using these interpolants, we obtain a mapping from any FLOP count  $C$ , to the most efficient choice of model size  $N$  and number of training tokens  $D$  such that  $\text{FLOPs}(N, D) = C$ .<sup>4</sup> At 1500 logarithmically spaced FLOP values, we find which model size achieves the lowest loss of all models along with the required number of training tokens. Finally, we fit power laws to estimate the optimal model size and number of training tokens for any given amount of compute (see the center and right panels of [Figure 2](#)), obtaining a relationship  $N_{\text{opt}} \propto C^a$  and  $D_{\text{opt}} \propto C^b$ . We find that  $a = 0.50$  and  $b = 0.50$ —as summarized in [Table 2](#). In [Section D.4](#), we show a head-to-head comparison at  $10^{21}$  FLOPs, using the model size recommended by our analysis and by the analysis of [Kaplan et al. \(2020\)](#)—using the model size we predict has a clear advantage.

### 3.2. Approach 2: IsoFLOP profiles

In our second approach we vary the model size<sup>5</sup> for a fixed set of 9 different training FLOP counts<sup>6</sup> (ranging from  $6 \times 10^{18}$  to  $3 \times 10^{21}$  FLOPs), and consider the final training loss for each point<sup>7</sup>. In contrast with Approach 1 that considered points  $(N, D, L)$  along the entire training runs. This allows us to directly answer the question: For a given FLOP budget, what is the optimal parameter count?

<sup>4</sup>Note that all selected points are within the last 15% of training. This suggests that when training a model over  $D$  tokens, we should pick a cosine cycle length that decays  $10\times$  over approximately  $D$  tokens—see further details in [Appendix B](#).

<sup>5</sup>In approach 2, model size varies up to 16B as opposed to approach 1 where we only used models up to 10B.

<sup>6</sup>The number of training tokens is determined by the model size and training FLOPs.

<sup>7</sup>We set the cosine schedule length to match the number of tokens, which is optimal according to the analysis presented in [Appendix B](#).

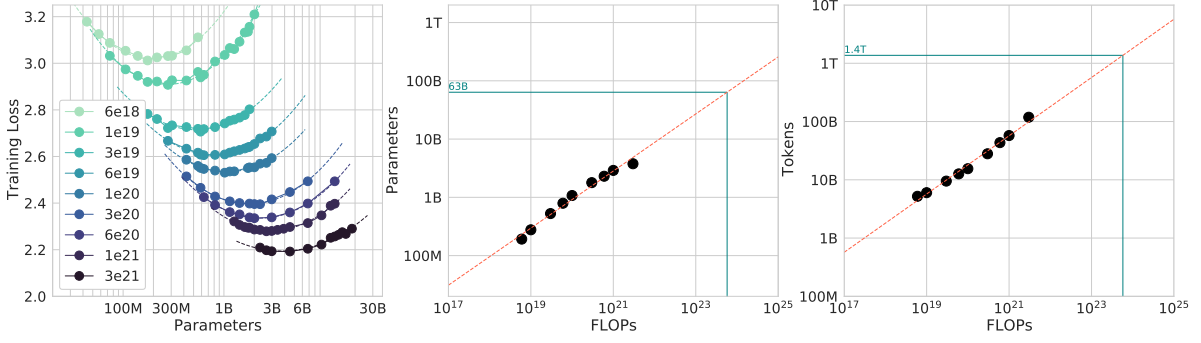


Figure 3 | **IsoFLOP curves.** For various model sizes, we choose the number of training tokens such that the final FLOPs is a constant. The cosine cycle length is set to match the target FLOP count. We find a clear valley in loss, meaning that for a given FLOP budget there is an optimal model to train (**left**). Using the location of these valleys, we project optimal model size and number of tokens for larger models (**center** and **right**). In green, we show the estimated number of parameters and tokens for an *optimal* model trained with the compute budget of *Gopher*.

For each FLOP budget, we plot the final loss (after smoothing) against the parameter count in Figure 3 (left). In all cases, we ensure that we have trained a diverse enough set of model sizes to see a clear minimum in the loss. We fit a parabola to each IsoFLOPs curve to directly estimate at what model size the minimum loss is achieved (Figure 3 (left)). As with the previous approach, we then fit a power law between FLOPs and loss-optimal model size and number of training tokens, shown in Figure 3 (center, right). Again, we fit exponents of the form  $N_{opt} \propto C^a$  and  $D_{opt} \propto C^b$  and we find that  $a = 0.49$  and  $b = 0.51$ —as summarized in Table 2.

### 3.3. Approach 3: Fitting a parametric loss function

Lastly, we model all final losses from experiments in Approach 1 & 2 as a parametric function of model parameter count and the number of seen tokens. Following a classical risk decomposition (see Section D.2), we propose the following functional form

$$\hat{L}(N, D) \triangleq E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}. \quad (2)$$

The first term captures the loss for an ideal generative process on the data distribution, and should correspond to the entropy of natural text. The second term captures the fact that a perfectly trained transformer with  $N$  parameters underperforms the ideal generative process. The final term captures the fact that the transformer is not trained to convergence, as we only make a finite number of optimisation steps, on a sample of the dataset distribution.

**Model fitting.** To estimate  $(A, B, E, \alpha, \beta)$ , we minimize the Huber loss (Huber, 1964) between the predicted and observed log loss using the L-BFGS algorithm (Nocedal, 1980):

$$\min_{A, B, E, \alpha, \beta} \sum_{\text{Runs } i} \text{Huber}_\delta \left( \log \hat{L}(N_i, D_i) - \log L_i \right) \quad (3)$$

We account for possible local minima by selecting the best fit from a grid of initialisations. The Huber loss ( $\delta = 10^{-3}$ ) is robust to outliers, which we find important for good predictive performance over held-out data points. Section D.2 details the fitting procedure and the loss decomposition.



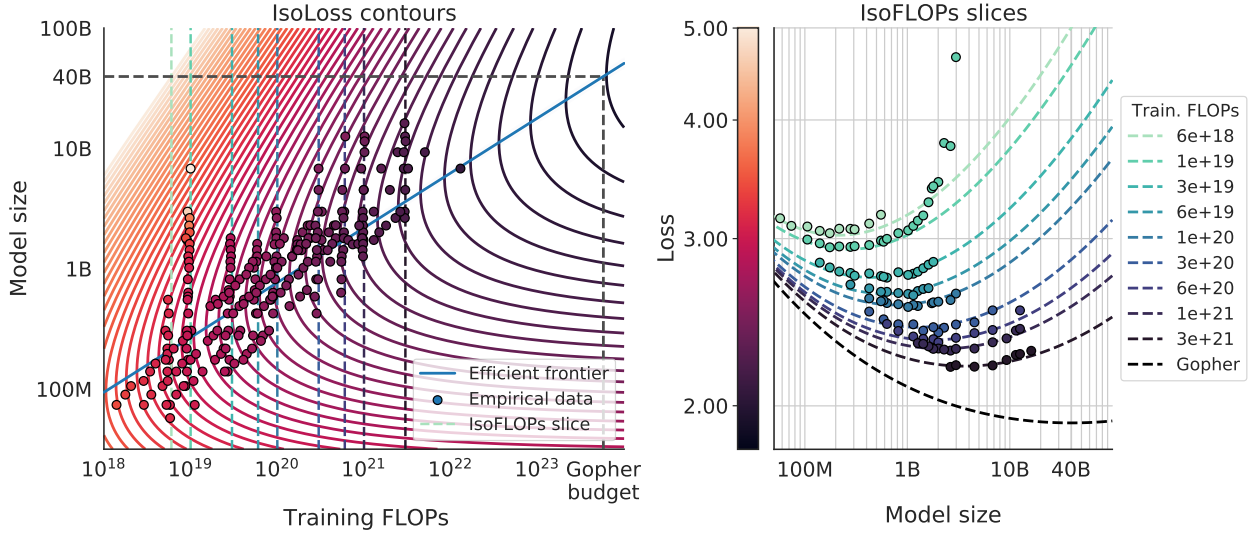


Figure 4 | **Parametric fit.** We fit a parametric modelling of the loss  $\hat{L}(N, D)$  and display contour (**left**) and isoFLOP slices (**right**). For each isoFLOP slice, we include a corresponding dashed line in the left plot. In the left plot, we show the efficient frontier in blue, which is a line in log-log space. Specifically, the curve goes through each iso-loss contour at the point with the fewest FLOPs. We project the optimal model size given the *Gopher* FLOP budget to be 40B parameters.

**Efficient frontier.** We can approximate the functions  $N_{opt}$  and  $D_{opt}$  by minimizing the parametric loss  $\hat{L}$  under the constraint  $FLOPs(N, D) \approx 6ND$  (Kaplan et al., 2020). The resulting  $N_{opt}$  and  $D_{opt}$  balance the two terms in Equation (3) that depend on model size and data. By construction, they have a power-law form:

$$N_{opt}(C) = G \left( \frac{C}{6} \right)^a, \quad D_{opt}(C) = G^{-1} \left( \frac{C}{6} \right)^b, \quad \text{where} \quad G = \left( \frac{\alpha A}{\beta B} \right)^{\frac{1}{\alpha+\beta}}, \quad a = \frac{\beta}{\alpha + \beta}, \quad \text{and} \quad b = \frac{\alpha}{\alpha + \beta}. \quad (4)$$

We show contours of the fitted function  $\hat{L}$  in Figure 4 (left), and the closed-form efficient computational frontier in blue. From this approach, we find that  $a = 0.46$  and  $b = 0.54$ —as summarized in Table 2.

### 3.4. Optimal model scaling

We find that the three approaches, despite using different fitting methodologies and different trained models, yield comparable predictions for the optimal scaling in parameters and tokens with FLOPs (shown in Table 2). All three approaches suggest that as compute budget increases, model size and the amount of training data should be increased in approximately equal proportions. The first and second approaches yield very similar predictions for optimal model sizes, as shown in Figure 1 and Figure A3. The third approach predicts even smaller models being optimal at larger compute budgets. We note that the observed points  $(L, N, D)$  for low training FLOPs ( $C \leq 1e21$ ) have larger residuals  $\|L - \hat{L}(N, D)\|_2^2$  than points with higher computational budgets. The fitted model places increased weight on the points with more FLOPs—automatically considering the low-computational budget points as outliers due to the Huber loss. As a consequence of the empirically observed negative curvature in the frontier  $C \rightarrow N_{opt}$  (see Appendix E), this results in predicting a lower  $N_{opt}$  than the two other approaches.

In Table 3 we show the estimated number of FLOPs and tokens that would ensure that a model of a given size lies on the compute-optimal frontier. Our findings suggests that the current generation of

Table 2 | **Estimated parameter and data scaling with increased training compute.** The listed values are the exponents,  $a$  and  $b$ , on the relationship  $N_{opt} \propto C^a$  and  $D_{opt} \propto C^b$ . Our analysis suggests a near equal scaling in parameters and data with increasing compute which is in clear contrast to previous work on the scaling of large models. The 10<sup>th</sup> and 90<sup>th</sup> percentiles are estimated via bootstrapping data (80% of the dataset is sampled 100 times) and are shown in parenthesis.

Approach	Coeff. $a$ where $N_{opt} \propto C^a$	Coeff. $b$ where $D_{opt} \propto C^b$
1. Minimum over training curves	0.50 (0.488, 0.502)	0.50 (0.501, 0.512)
2. IsoFLOP profiles	0.49 (0.462, 0.534)	0.51 (0.483, 0.529)
3. Parametric modelling of the loss	0.46 (0.454, 0.455)	0.54 (0.542, 0.543)
<a href="#">Kaplan et al. (2020)</a>	0.73	0.27

Table 3 | **Estimated optimal training FLOPs and training tokens for various model sizes.** For various model sizes, we show the projections from Approach 1 of how many FLOPs and training tokens would be needed to train compute-optimal models. The estimates for Approach 2 & 3 are similar (shown in [Section D.3](#))

Parameters	FLOPs	FLOPs (in <i>Gopher</i> unit)	Tokens
400 Million	1.92e+19	1/29,968	8.0 Billion
1 Billion	1.21e+20	1/4,761	20.2 Billion
10 Billion	1.23e+22	1/46	205.1 Billion
67 Billion	5.76e+23	1	1.5 Trillion
175 Billion	3.85e+24	6.7	3.7 Trillion
280 Billion	9.90e+24	17.2	5.9 Trillion
520 Billion	3.43e+25	59.5	11.0 Trillion
1 Trillion	1.27e+26	221.3	21.2 Trillion
10 Trillion	1.30e+28	22515.9	216.2 Trillion

large language models are considerably over-sized, given their respective compute budgets, as shown in [Figure 1](#). For example, we find that a 175 billion parameter model should be trained with a compute budget of  $4.41 \times 10^{24}$  FLOPs and on over 4.2 trillion tokens. A 280 billion *Gopher*-like model is the optimal model to train given a compute budget of approximately  $10^{25}$  FLOPs and should be trained on 6.8 trillion tokens. Unless one has a compute budget of  $10^{26}$  FLOPs (over 250× the compute used to train *Gopher*), a 1 trillion parameter model is unlikely to be the optimal model to train. Furthermore, the amount of training data that is projected to be needed is far beyond what is currently used to train large models, and underscores the importance of dataset collection in addition to engineering improvements that allow for model scale. While there is significant uncertainty extrapolating out many orders of magnitude, our analysis clearly suggests that given the training compute budget for many current LLMs, smaller models should have been trained on more tokens to achieve the most performant model.

In [Appendix C](#), we reproduce the IsoFLOP analysis on two additional datasets: C4 ([Raffel et al., 2020a](#)) and GitHub code ([Rae et al., 2021](#)). In both cases we reach the similar conclusion that model size and number of training tokens should be scaled in equal proportions.



## 4. Chinchilla

Based on our analysis in [Section 3](#), the optimal model size for the *Gopher* compute budget is somewhere between 40 and 70 billion parameters. We test this hypothesis by training a model on the larger end of this range—70B parameters—for 1.4T tokens, due to both dataset and computational efficiency considerations. In this section we compare this model, which we call *Chinchilla*, to *Gopher* and other LLMs. Both *Chinchilla* and *Gopher* have been trained for the same number of FLOPs but differ in the size of the model and the number of training tokens.

While pre-training a large language model has a considerable compute cost, downstream fine-tuning and inference also make up substantial compute usage ([Rae et al., 2021](#)). Due to being 4× smaller than *Gopher*, both the memory footprint and inference cost of *Chinchilla* are also smaller.

### 4.1. Model and training details

The full set of hyperparameters used to train *Chinchilla* are given in [Table 4](#). *Chinchilla* uses the same model architecture and training setup as *Gopher* with the exception of the differences listed below.

- We train *Chinchilla* on *MassiveText* (the same dataset as *Gopher*) but use a slightly different subset distribution (shown in [Table A1](#)) to account for the increased number of training tokens.
- We use AdamW ([Loshchilov and Hutter, 2019](#)) for *Chinchilla* rather than Adam ([Kingma and Ba, 2014](#)) as this improves the language modelling loss and the downstream task performance after finetuning.<sup>8</sup>
- We train *Chinchilla* with a slightly modified SentencePiece ([Kudo and Richardson, 2018](#)) tokenizer that does not apply NFKC normalisation. The vocabulary is very similar—94.15% of tokens are the same as those used for training *Gopher*. We find that this particularly helps with the representation of mathematics and chemistry, for example.
- Whilst the forward and backward pass are computed in bfloat16, we store a float32 copy of the weights in the distributed optimiser state ([Rajbhandari et al., 2020](#)). See *Lessons Learned* from [Rae et al. \(2021\)](#) for additional details.

In [Appendix G](#) we show the impact of the various optimiser related changes between *Chinchilla* and *Gopher*. All models in this analysis have been trained on TPUv3/TPUv4 ([Jouppi et al., 2017](#)) with JAX ([Bradbury et al., 2018](#)) and Haiku ([Hennigan et al., 2020](#)). We include a *Chinchilla* model card ([Mitchell et al., 2019](#)) in [Table A8](#).

Model	Layers	Number Heads	Key/Value Size	$d_{\text{model}}$	Max LR	Batch Size
<i>Gopher</i> 280B	80	128	128	16,384	$4 \times 10^{-5}$	3M $\rightarrow$ 6M
<i>Chinchilla</i> 70B	80	64	128	8,192	$1 \times 10^{-4}$	1.5M $\rightarrow$ 3M

Table 4 | ***Chinchilla* architecture details.** We list the number of layers, the key/value size, the bottleneck activation size  $d_{\text{model}}$ , the maximum learning rate, and the training batch size (# tokens). The feed-forward size is always set to  $4 \times d_{\text{model}}$ . Note that we double the batch size midway through training for both *Chinchilla* and *Gopher*.

<sup>8</sup>Interestingly, a model trained with AdamW only passes the training performance of a model trained with Adam around 80% of the way through the cosine cycle, though the ending performance is notably better—see [Figure A7](#)

	# Tasks	Examples
Language Modelling	20	WikiText-103, The Pile: PG-19, arXiv, FreeLaw, ...
Reading Comprehension	3	RACE-m, RACE-h, LAMBADA
Question Answering	3	Natural Questions, TriviaQA, TruthfulQA
Common Sense	5	HellaSwag, Winogrande, PIQA, SIQA, BoolQ
MMLU	57	High School Chemistry, Astronomy, Clinical Knowledge, ...
BIG-bench	62	Causal Judgement, Epistemic Reasoning, Temporal Sequences, ...

Table 5 | **All evaluation tasks.** We evaluate *Chinchilla* on a collection of language modelling along with downstream tasks. We evaluate on largely the same tasks as in [Rae et al. \(2021\)](#), to allow for direct comparison.

## 4.2. Results

We perform an extensive evaluation of *Chinchilla*, comparing against various large language models. We evaluate on a large subset of the tasks presented in [Rae et al. \(2021\)](#), shown in [Table 5](#). As the focus of this work is on optimal model scaling, we included a large representative subset, and introduce a few new evaluations to allow for better comparison to other existing large models. The evaluation details for all tasks are the same as described in [Rae et al. \(2021\)](#).

### 4.2.1. Language modelling

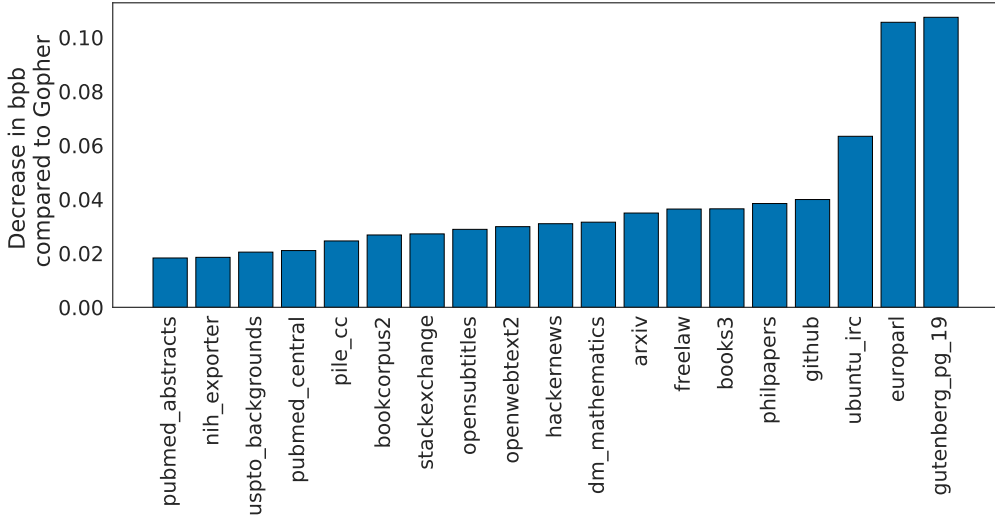


Figure 5 | **Pile Evaluation.** For the different evaluation sets in The Pile ([Gao et al., 2020](#)), we show the bits-per-byte (bpb) improvement (decrease) of *Chinchilla* compared to *Gopher*. On all subsets, *Chinchilla* outperforms *Gopher*.

*Chinchilla* significantly outperforms *Gopher* on all evaluation subsets of The Pile ([Gao et al., 2020](#)), as shown in [Figure 5](#). Compared to Jurassic-1 (178B) [Lieber et al. \(2021\)](#), *Chinchilla* is more performant on all but two subsets– *dm\_mathematics* and *ubuntu\_irc*– see [Table A5](#) for a raw bits-per-byte comparison. On Wikitext103 ([Merity et al., 2017](#)), *Chinchilla* achieves a perplexity of 7.16 compared to 7.75 for *Gopher*. Some caution is needed when comparing *Chinchilla* with *Gopher* on these language modelling benchmarks as *Chinchilla* is trained on 4× more data than *Gopher* and thus train/test set leakage may artificially enhance the results. We thus place more emphasis on other

Random	25.0%
Average human rater	34.5%
GPT-3 5-shot	43.9%
<i>Gopher</i> 5-shot	60.0%
<b><i>Chinchilla</i> 5-shot</b>	<b>67.6%</b>
Average human expert performance	89.8%
June 2022 Forecast	57.1%
June 2023 Forecast	63.4%

Table 6 | **Massive Multitask Language Understanding (MMLU)**. We report the average 5-shot accuracy over 57 tasks with model and human accuracy comparisons taken from [Hendrycks et al. \(2020\)](#). We also include the average prediction for state of the art accuracy in June 2022/2023 made by 73 competitive human forecasters in [Steinhardt \(2021\)](#).

tasks for which leakage is less of a concern, such as MMLU ([Hendrycks et al., 2020](#)) and BIG-bench ([BIG-bench collaboration, 2021](#)) along with various closed-book question answering and common sense analyses.

#### 4.2.2. MMLU

The Massive Multitask Language Understanding (MMLU) benchmark ([Hendrycks et al., 2020](#)) consists of a range of exam-like questions on academic subjects. In [Table 6](#), we report *Chinchilla*’s average 5-shot performance on MMLU (the full breakdown of results is shown in [Table A6](#)). On this benchmark, *Chinchilla* significantly outperforms *Gopher* despite being much smaller, with an average accuracy of 67.6% (improving upon *Gopher* by 7.6%). Remarkably, *Chinchilla* even outperforms the expert forecast for June 2023 of 63.4% accuracy (see [Table 6](#)) ([Steinhardt, 2021](#)). Furthermore, *Chinchilla* achieves greater than 90% accuracy on 4 different individual tasks— `high_school_gov_and_politics`, `international_law`, `sociology`, and `us_foreign_policy`. To our knowledge, no other model has achieved greater than 90% accuracy on a subset.

In [Figure 6](#), we show a comparison to *Gopher* broken down by task. Overall, we find that *Chinchilla* improves performance on the vast majority of tasks. On four tasks (`college_mathematics`, `econometrics`, `moral_scenarios`, and `formal_logic`) *Chinchilla* underperforms *Gopher*, and there is no change in performance on two tasks.

#### 4.2.3. Reading comprehension

On the final word prediction dataset LAMBADA ([Paperno et al., 2016](#)), *Chinchilla* achieves 77.4% accuracy, compared to 74.5% accuracy from *Gopher* and 76.6% from MT-NLG 530B (see [Table 7](#)). On RACE-h and RACE-m ([Lai et al., 2017](#)), *Chinchilla* greatly outperforms *Gopher*, improving accuracy by more than 10% in both cases—see [Table 7](#).

#### 4.2.4. BIG-bench

We analysed *Chinchilla* on the same set of BIG-bench tasks ([BIG-bench collaboration, 2021](#)) reported in [Rae et al. \(2021\)](#). Similar to what we observed in MMLU, *Chinchilla* outperforms *Gopher* on the vast majority of tasks (see [Figure 7](#)). We find that *Chinchilla* improves the average performance by 10.7%, reaching an accuracy of 65.1% versus 54.4% for *Gopher*. Of the 62 tasks we consider, *Chinchilla* performs worse than *Gopher* on only four—`crash_blossom`, `dark_humor_detection`,

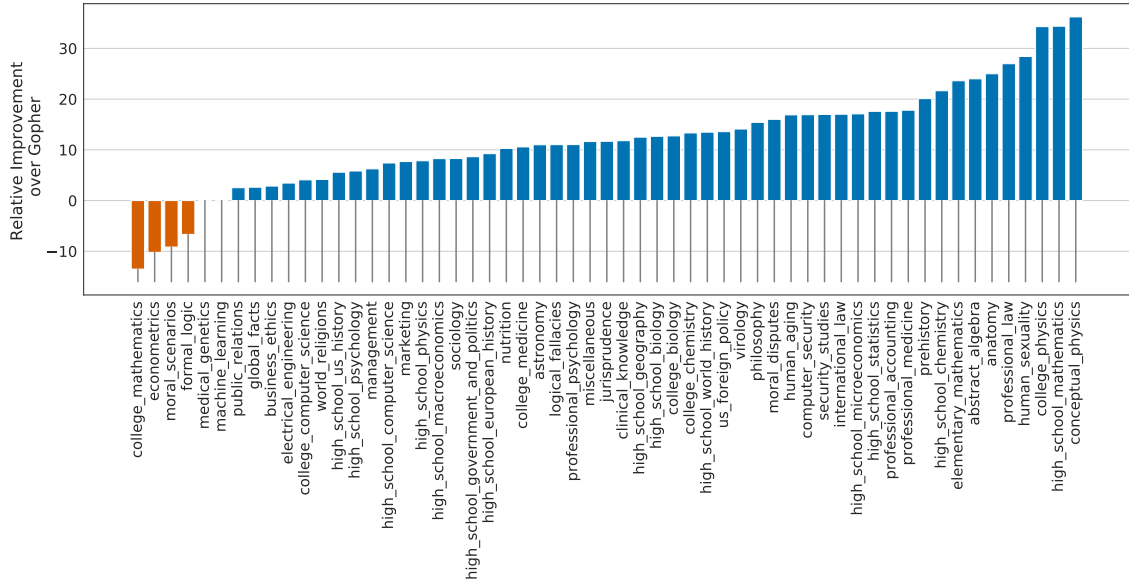


Figure 6 | **MMLU results compared to *Gopher*** We find that *Chinchilla* outperforms *Gopher* by 7.6% on average (see Table 6) in addition to performing better on 51/57 individual tasks, the same on 2/57, and worse on only 4/57 tasks.

	<i>Chinchilla</i>	<i>Gopher</i>	GPT-3	MT-NLG 530B
LAMBADA Zero-Shot	<b>77.4</b>	74.5	76.2	76.6
RACE-m Few-Shot	<b>86.8</b>	75.1	58.1	-
RACE-h Few-Shot	<b>82.3</b>	71.6	46.8	47.9

Table 7 | **Reading comprehension.** On RACE-h and RACE-m (Lai et al., 2017), *Chinchilla* considerably improves performance over *Gopher*. Note that GPT-3 and MT-NLG 530B use a different prompt format than we do on RACE-h/m, so results are not comparable to *Gopher* and *Chinchilla*. On LAMBADA (Paperno et al., 2016), *Chinchilla* outperforms both *Gopher* and MT-NLG 530B.

mathematical\_induction and logical\_args. Full accuracy results for *Chinchilla* can be found in Table A7.

#### 4.2.5. Common sense

We evaluate *Chinchilla* on various common sense benchmarks: PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), Winogrande (Sakaguchi et al., 2020), HellaSwag (Zellers et al., 2019), and BoolQ (Clark et al., 2019). We find that *Chinchilla* outperforms both *Gopher* and GPT-3 on all tasks and outperforms MT-NLG 530B on all but one task—see Table 8.

On TruthfulQA (Lin et al., 2021), *Chinchilla* reaches 43.6%, 58.5%, and 66.7% accuracy with 0-shot, 5-shot, and 10-shot respectively. In comparison, *Gopher* achieved only 29.5% 0-shot and 43.7% 10-shot accuracy. In stark contrast with the findings of Lin et al. (2021), the large improvements (14.1% in 0-shot accuracy) achieved by *Chinchilla* suggest that better modelling of the pre-training data alone can lead to substantial improvements on this benchmark.

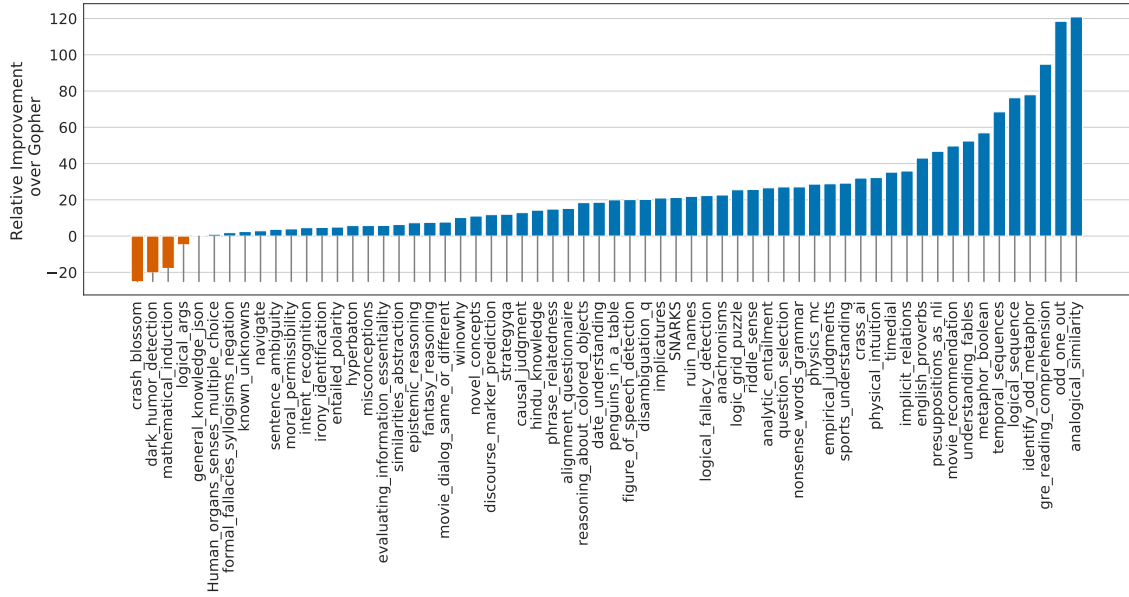


Figure 7 | **BIG-bench results compared to Gopher** Chinchilla out performs Gopher on all but four BIG-bench tasks considered. Full results are in Table A7.

#### 4.2.6. Closed-book question answering

Results on closed-book question answering benchmarks are reported in Table 9. On the Natural Questions dataset (Kwiatkowski et al., 2019), Chinchilla achieves new closed-book SOTA accuracies: 31.5% 5-shot and 35.5% 64-shot, compared to 21% and 28% respectively, for Gopher. On TriviaQA (Joshi et al., 2017) we show results for both the filtered (previously used in retrieval and open-book work) and unfiltered set (previously used in large language model evaluations). In both cases, Chinchilla substantially out performs Gopher. On the filtered version, Chinchilla lags behind the open book SOTA (Izacard and Grave, 2020) by only 7.9%. On the unfiltered set, Chinchilla outperforms GPT-3—see Table 9.

#### 4.2.7. Gender bias and toxicity

Large Language Models carry potential risks such as outputting offensive language, propagating social biases, and leaking private information (Bender et al., 2021; Weidinger et al., 2021). We expect Chinchilla to carry risks similar to Gopher because Chinchilla is trained on the same data,

	Chinchilla	Gopher	GPT-3	MT-NLG 530B	Supervised SOTA
HellaSWAG	<b>80.8%</b>	79.2%	78.9%	80.2%	93.9%
PIQA	81.8%	81.8%	81.0%	<b>82.0%</b>	90.1%
Winogrande	<b>74.9%</b>	70.1%	70.2%	73.0%	91.3%
SIQA	<b>51.3%</b>	50.6%	-	-	83.2%
BoolQ	<b>83.7%</b>	79.3%	60.5%	78.2%	91.4%

Table 8 | **Zero-shot comparison on Common Sense benchmarks.** We show a comparison between Chinchilla, Gopher, and MT-NLG 530B on various Common Sense benchmarks. We see that Chinchilla matches or outperforms Gopher and GPT-3 on all tasks. On all but one Chinchilla outperforms the much larger MT-NLG 530B model.

	Method	<i>Chinchilla</i>	<i>Gopher</i>	GPT-3	SOTA (open book)
Natural Questions (dev)	0-shot	16.6%	10.1%	14.6%	54.4%
	5-shot	31.5%	24.5%	-	
	64-shot	35.5%	28.2%	29.9%	
TriviaQA (unfiltered, test)	0-shot	67.0%	52.8%	64.3 %	-
	5-shot	73.2%	63.6%	-	
	64-shot	72.3%	61.3%	71.2%	
TriviaQA (filtered, dev)	0-shot	55.4%	43.5%	-	72.5%
	5-shot	64.1%	57.0%	-	
	64-shot	64.6%	57.2%	-	

Table 9 | **Closed-book question answering.** For Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017), *Chinchilla* outperforms *Gopher* in all cases. On Natural Questions, *Chinchilla* outperforms GPT-3. On TriviaQA we show results on two different evaluation sets to allow for comparison to GPT-3 and to open book SOTA (FiD + Distillation (Izacard and Grave, 2020)).

albeit with slightly different relative weights, and because it has a similar architecture. Here, we examine gender bias (particularly gender and occupation bias) and generation of toxic language. We select a few common evaluations to highlight potential issues, but stress that our evaluations are not comprehensive and much work remains to understand, evaluate, and mitigate risks in LLMs.

**Gender bias.** As discussed in Rae et al. (2021), large language models reflect contemporary and historical discourse about different groups (such as gender groups) from their training dataset, and we expect the same to be true for *Chinchilla*. Here, we test if potential gender and occupation biases manifest in unfair outcomes on coreference resolutions, using the Winogender dataset (Rudinger et al., 2018) in a zero-shot setting. Winogender tests whether a model can correctly determine if a pronoun refers to different occupation words. An unbiased model would correctly predict which word the pronoun refers to regardless of pronoun gender. We follow the same setup as in Rae et al. (2021) (described further in Section H.3).

As shown in Table 10, *Chinchilla* correctly resolves pronouns more frequently than *Gopher* across all groups. Interestingly, the performance increase is considerably smaller for male pronouns (increase of 3.2%) than for female or neutral pronouns (increases of 8.3% and 9.2% respectively). We also consider *gotcha* examples, in which the correct pronoun resolution contradicts gender stereotypes (determined by labor statistics). Again, we see that *Chinchilla* resolves pronouns more accurately than *Gopher*. When breaking up examples by male/female gender and *gotcha/not gotcha*, the largest improvement is on female *gotcha* examples (improvement of 10%). Thus, though *Chinchilla* uniformly overcomes gender stereotypes for more coreference examples than *Gopher*, the rate of improvement is higher for some pronouns than others, suggesting that the improvements conferred by using a more compute-optimal model can be uneven.

**Sample toxicity.** Language models are capable of generating toxic language—including insults, hate speech, profanities and threats (Gehman et al., 2020; Rae et al., 2021). While toxicity is an umbrella term, and its evaluation in LMs comes with challenges (Welbl et al., 2021; Xu et al., 2021), automatic classifier scores can provide an indication for the levels of harmful text that a LM generates. Rae et al. (2021) found that improving language modelling loss by increasing the number of model parameters has only a negligible effect on toxic text generation (unprompted); here we analyze



	<i>Chinchilla</i>	<i>Gopher</i>		<i>Chinchilla</i>	<i>Gopher</i>
All	78.3%	71.4%	Male <i>gotcha</i>	62.5%	59.2%
Male	71.2%	68.0%	Male <i>not gotcha</i>	80.0%	76.7%
Female	79.6%	71.3%	Female <i>gotcha</i>	76.7%	66.7%
Neutral	84.2%	75.0%	Female <i>not gotcha</i>	82.5%	75.8%

Table 10 | **Winogender results.** **Left:** *Chinchilla* consistently resolves pronouns better than *Gopher*. **Right:** *Chinchilla* performs better on examples which contradict gender stereotypes (*gotcha* examples). However, difference in performance across groups suggests *Chinchilla* exhibits bias.

whether the same holds true for a lower LM loss achieved via more compute-optimal training. Similar to the protocol of [Rae et al. \(2021\)](#), we generate 25,000 unprompted samples from *Chinchilla*, and compare their *PerspectiveAPI* toxicity score distribution to that of *Gopher*-generated samples. Several summary statistics indicate an absence of major differences: the mean (median) toxicity score for *Gopher* is 0.081 (0.064), compared to 0.087 (0.066) for *Chinchilla*, and the 95<sup>th</sup> percentile scores are 0.230 for *Gopher*, compared to 0.238 for *Chinchilla*. That is, the large majority of generated samples are classified as non-toxic, and the difference between the models is negligible. In line with prior findings ([Rae et al., 2021](#)), this suggests that toxicity levels in unconditional text generation are largely independent of the model quality (measured in language modelling loss), i.e. that better models of the training dataset are not necessarily more toxic.

## 5. Discussion & Conclusion

The trend so far in large language model training has been to increase the model size, often without increasing the number of training tokens. The largest dense transformer, MT-NLG 530B, is now over 3× larger than GPT-3’s 170 billion parameters from just two years ago. However, this model, as well as the majority of existing large models, have all been trained for a comparable number of tokens—around 300 billion. While the desire to train these mega-models has led to substantial engineering innovation, we hypothesize that the race to train larger and larger models is resulting in models that are substantially underperforming compared to what could be achieved with the same compute budget.

We propose three predictive approaches towards optimally setting model size and training duration, based on the outcome of over 400 training runs. All three approaches predict that *Gopher* is substantially over-sized and estimate that for the same compute budget a smaller model trained on more data will perform better. We directly test this hypothesis by training *Chinchilla*, a 70B parameter model, and show that it outperforms *Gopher* and even larger models on nearly every measured evaluation task.

Whilst our method allows us to make predictions on how to scale large models when given additional compute, there are several limitations. Due to the cost of training large models, we only have two comparable training runs at large scale (*Chinchilla* and *Gopher*), and we do not have additional tests at intermediate scales. Furthermore, we assume that the efficient computational frontier can be described by a power-law relationship between the compute budget, model size, and number of training tokens. However, we observe some concavity in  $\log(N_{opt})$  at high compute budgets (see [Appendix E](#)). This suggests that we may still be overestimating the optimal size of large models. Finally, the training runs for our analysis have all been trained on less than an epoch of data; future work may consider the multiple epoch regime. Despite these limitations, the comparison of *Chinchilla* to *Gopher* validates our performance predictions, that have thus enabled training a better (and more

lightweight) model at the same compute budget.

Though there has been significant recent work allowing larger and larger models to be trained, our analysis suggests an increased focus on dataset scaling is needed. Speculatively, we expect that scaling to larger and larger datasets is only beneficial when the data is high-quality. This calls for responsibly collecting larger datasets with a high focus on dataset quality. Larger datasets will require extra care to ensure train-test set overlap is properly accounted for, both in the language modelling loss but also with downstream tasks. Finally, training for trillions of tokens introduces many ethical and privacy concerns. Large datasets scraped from the web will contain toxic language, biases, and private information. With even larger datasets being used, the quantity (if not the frequency) of such information increases, which makes dataset introspection all the more important. *Chinchilla* does suffer from bias and toxicity but interestingly it seems less affected than *Gopher*. Better understanding how performance of large language models and toxicity interact is an important future research question.

While we have applied our methodology towards the training of auto-regressive language models, we expect that there is a similar trade-off between model size and the amount of data in other modalities. As training large models is very expensive, choosing the optimal model size and training steps beforehand is essential. The methods we propose are easy to reproduce in new settings.

## 6. Acknowledgements

We’d like to thank Jean-baptiste Alayrac, Kareem Ayoub, Chris Dyer, Nando de Freitas, Demis Hassabis, Geoffrey Irving, Koray Kavukcuoglu, Nate Kushman and Angeliki Lazaridou for useful comments on the manuscript. We’d like to thank Andy Brock, Irina Higgins, Michela Paganini, Francis Song, and other colleagues at DeepMind for helpful discussions. We are also very grateful to the JAX and XLA team for their support and assistance.

## References

- M. Artetxe, S. Bhosale, N. Goyal, T. Mihaylov, M. Ott, S. Shleifer, X. V. Lin, J. Du, S. Iyer, R. Pasunuru, G. Anantharaman, X. Li, S. Chen, H. Akin, M. Baines, L. Martin, X. Zhou, P. S. Koura, B. O’Horo, J. Wang, L. Zettlemoyer, M. Diab, Z. Kozareva, and V. Stoyanov. Efficient Large Scale Language Modeling with Mixtures of Experts. [arXiv:2112.10684](https://arxiv.org/abs/2112.10684), 2021.
- E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.
- BIG-bench collaboration. Beyond the imitation game: Measuring and extrapolating the capabilities of language models. *In preparation*, 2021. URL <https://github.com/google/BIG-bench/>.
- Y. Bisk, R. Zellers, J. Gao, Y. Choi, et al. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439, 2020.
- S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. van den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. de Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. W. Rae, E. Elsen, and L. Sifre. Improving language models by retrieving from trillions of tokens. [arXiv 2112.04426](https://arxiv.org/abs/2112.04426), 2021.