MAI 5301 - Presentation

Scaling Laws for Neural Language Models
&
Training Compute-Optimal Large Language Models

- How model size, dataset size, and compute impact the performance of large language models.
- Explores empirical scaling laws that predict test loss and generalization.

- Kaplan et al., 2020 (arXiv:2001.08361):
    - Established fundamental power-law relationships between model size, dataset size, and compute.
    - Provides a predictive framework for training large transformers efficiently.
    -
- Henighan et al., 2022 (arXiv:2203.15556):
    - Extends scaling insights to long-context transformers, showing how performance improves with context length.
    - Highlights practical considerations for training very large models.

- Goal:
    - Understand how and why scaling works.
    - Inform decisions on model design, dataset requirements, and compute allocation.

Scaling Laws for Neural Language Models

Introduction

Core Motivation

- Language modeling is a central benchmark for modern deep learning
- Empirical success of large models lacked a quantitative understanding of scaling
- Prior progress relied on trial and error rather than predictable laws

Primary Goal

- Identify empirical scaling laws governing language model performance
- Understand how loss scales with:
    - Model size (parameters)
    - Dataset size (tokens)
    - Training compute (FLOPs)

Introduction

Key Claims Introduced in this paper

- Model performance improves smoothly with scale
- Loss follows simple power laws
- Model architecture details matter less than scale (depth and width of network)
- Enables compute-optimal training strategies

Summary of Scaling Laws

Language Model tests follow predictable power laws and it applies when performance is limited by:

- Model size
- Data Size
- Optimally Allocated Compute

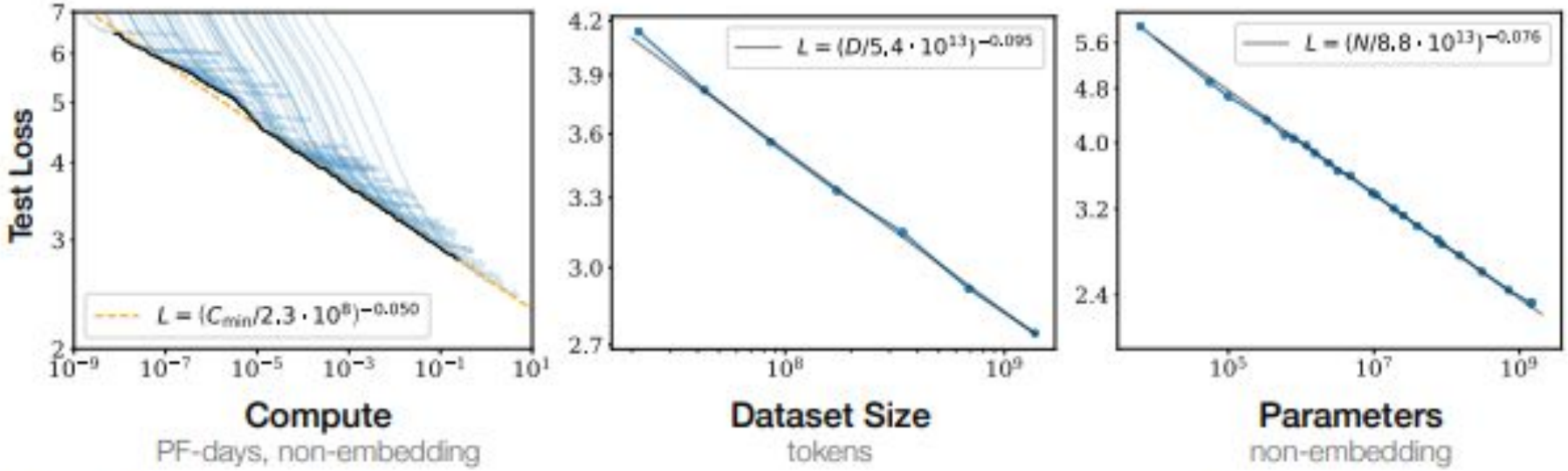


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

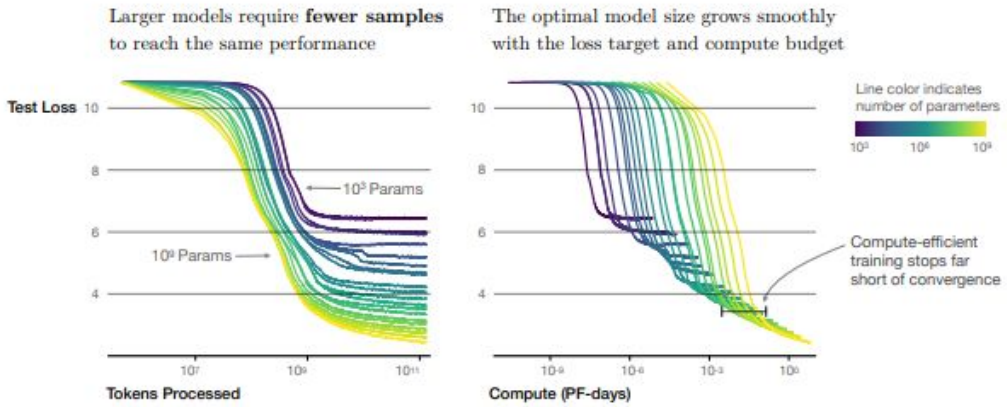


Figure 2 We show a series of language model training runs, with models ranging in size from 10^3 to 10^9 parameters (excluding embeddings).

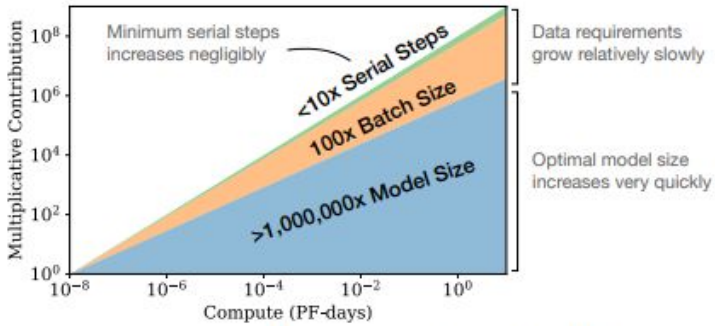


Figure 3 As more compute becomes available, we can choose how much to allocate towards training larger models, using larger batches, and training for more steps. We illustrate this for a billion-fold increase in compute. For optimally compute-efficient training, most of the increase should go towards increased model size. A relatively small increase in data is needed to avoid reuse. Of the increase in data, most can be used to increase parallelism through larger batch sizes, with only a very small increase in serial training time required.

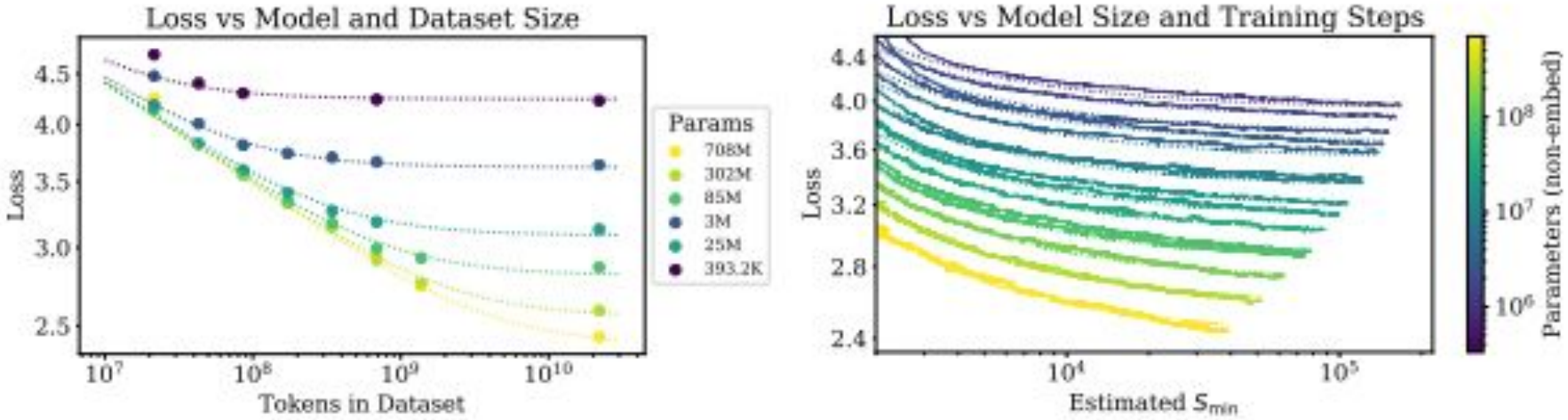


Figure 4 **Left**: The early-stopped test loss $L(N, D)$ varies predictably with the dataset size D and model size N according to Equation (1.5). **Right**: After an initial transient period, learning curves for all model sizes N can be fit with Equation (1.6), which is parameterized in terms of $S_{\min}$, the number of steps when training at large batch size (details in Section 5.1).

Background and Methods

Model Class

- Autoregressive Transformer language models
- Trained to predict next token given prior context (WebText2 dataset, byte-pair encoding, Adam optimizer/Adafactor)~ (Convergence were largely independent of learning rate schedule ~ 3000 step linear warmup followed by a cosine decay to 0)
- Loss measured as cross-entropy on held-out validation data

Background and Methods

Key Scaling Variables

- Model size (N): number of non-embedding parameters (number of layers, dimensions: residual stream, feed forward layer and attention output)
- Dataset size (D): total number of training tokens
- Compute (C): total training FLOPs (≈ N × D)

Background and Methods

Experimental Design

- Large grid of experiments across orders of magnitude in Model size and Data size
- Multiple training runs to isolate limiting factors
- Evaluate both final loss and training dynamics

Purpose of Methodology

- Empirically map performance trends
- Identify regimes where one resource becomes the bottleneck

Empirical Results and Basic Power Laws

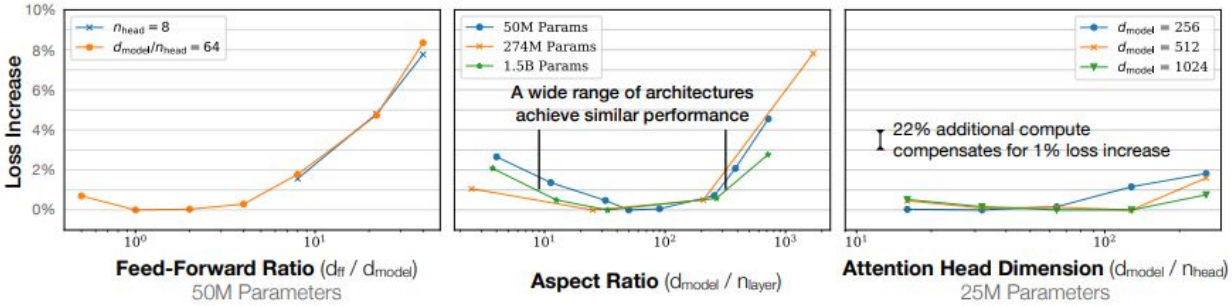


Figure 5 Performance depends very mildly on model shape when the total number of non-embedding parameters N is held fixed. The loss varies only a few percent over a wide range of shapes. Small differences in parameter counts are compensated for by using the fit to $L(N)$ as a baseline. Aspect ratio in particular can vary by a factor of 40 while only slightly impacting performance; an $(n_{\text{layer}}, d_{\text{model}}) = (6, 4288)$ reaches a loss within 3% of the $(48, 1600)$ model used in [RWC+19].

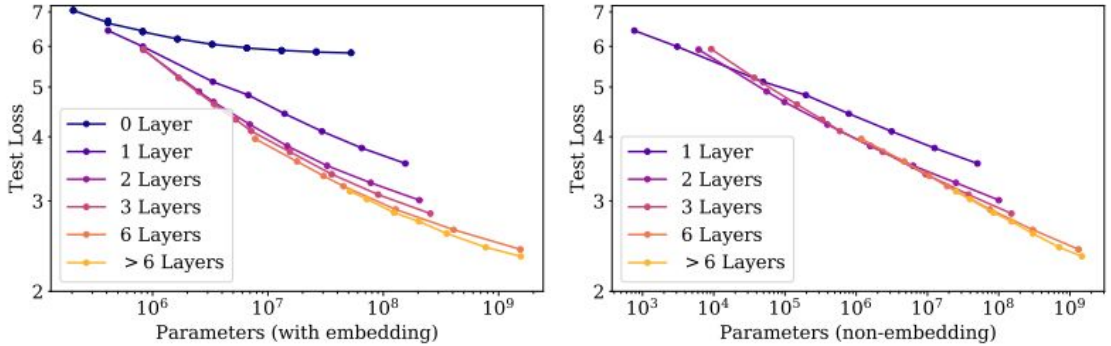


Figure 6 **Left:** When we include embedding parameters, performance appears to depend strongly on the number of layers in addition to the number of parameters. **Right:** When we exclude embedding parameters, the performance of models with different depths converge to a single trend. Only models with fewer than 2 layers or with extreme depth-to-width ratios deviate significantly from the trend.

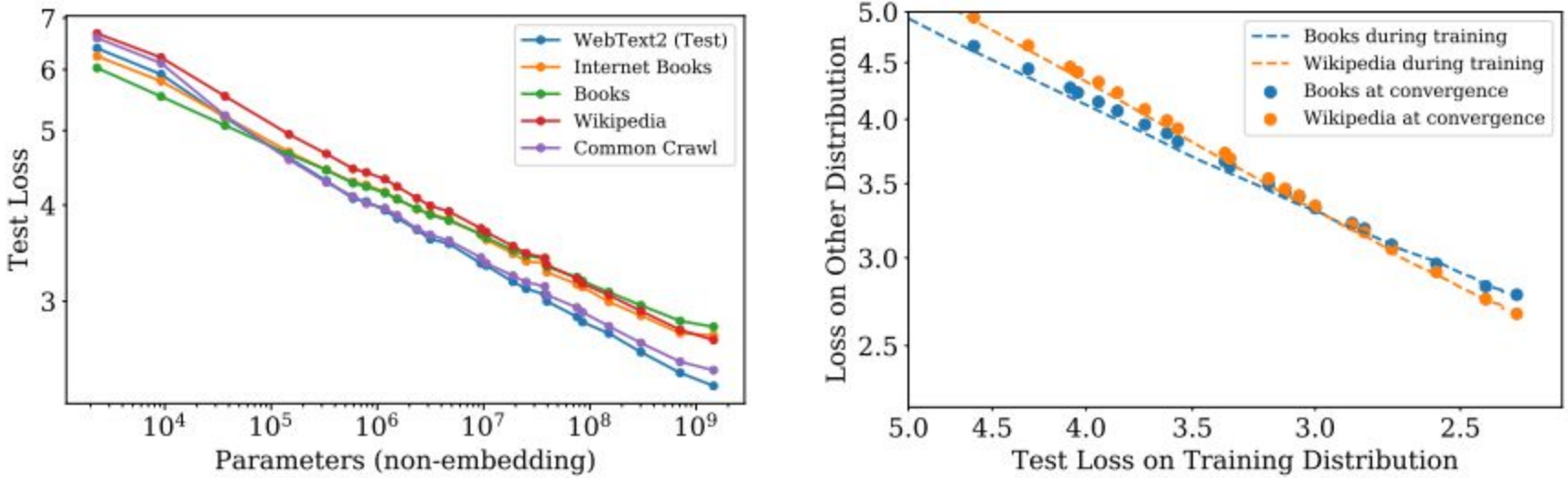


Figure 8 **Left:** Generalization performance to other data distributions improves smoothly with model size, with only a small and very slowly growing offset from the WebText2 training distribution. **Right:** Generalization performance depends only on training distribution performance, and not on the phase of training. We compare generalization of converged models (points) to that of a single large model (dashed curves) as it trains.

Empirical Results and Basic Power Laws

Central Empirical Finding

- Validation loss follows power-law scaling with respect to Model size, Dataset size, and Compute

Loss vs Model Size

- Larger models consistently yield lower loss
- Diminishing returns but no sharp saturation
- Performance weakly dependent on architectural details
- Capacity, not architecture, dominates performance
- Scaling parameters is reliably beneficial

Empirical Results and Basic Power Laws

Loss vs Dataset Size

- More data improves generalization
- Gains diminish smoothly as data grows
- Data becomes ineffective if model is too small

Loss vs Compute

- Increasing compute yields predictable loss reductions
- No evidence of abrupt diminishing returns in studied range

Charting the Infinite Data Limit and Overfitting

Infinite Data Regime

- With sufficiently large datasets:
    - Models converge to a loss floor determined by model size
- Data alone cannot overcome limited model capacity

Overfitting Regime

- When model size grows faster than dataset size:
    - Validation loss worsens due to memorization
- Overfitting scales with:
  Model size ^ 0.74 / Dataset size

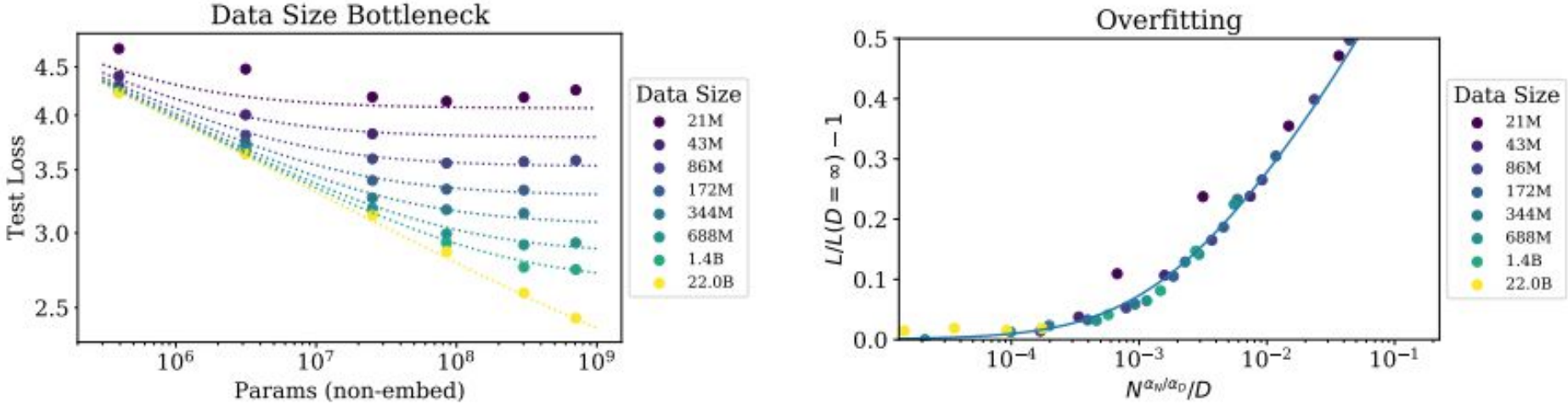


Figure 9 The early-stopped test loss $L(N, D)$ depends predictably on the dataset size D and model size N according to Equation (1.5). **Left**: For large D, performance is a straight power law in N. For a smaller fixed D, performance stops improving as N increases and the model begins to overfit. (The reverse is also true, see Figure 4.) **Right**: The extent of overfitting depends predominantly on the ratio $N^{\frac{\alpha_N}{\alpha_D}}/D$, as predicted in equation (4.3). The line is our fit to that equation.

Charting the Infinite Data Limit and Overfitting

Key Implications

- Larger models require proportionally less additional data
- Model and dataset must be scaled together
- Overfitting behavior is predictable and quantifiable

Scaling Laws with Model Size and Training Time

Training Dynamics

- Loss decreases as a power law with training steps
- Training curves align across scales when rescaled
- Early training behavior predicts long-term outcomes

Sample Efficiency

- Larger models:
    - Learn faster per token
    - Achieve lower loss with fewer training samples

Scaling Laws with Model Size and Training Time

Batch Size Scaling

- Optimal batch size increases with scale
- Gradient noise scale grows with training efficiency
- Very large batch sizes become optimal for large models
- Bcrit roughly doubles every time loss decreases by ~13%

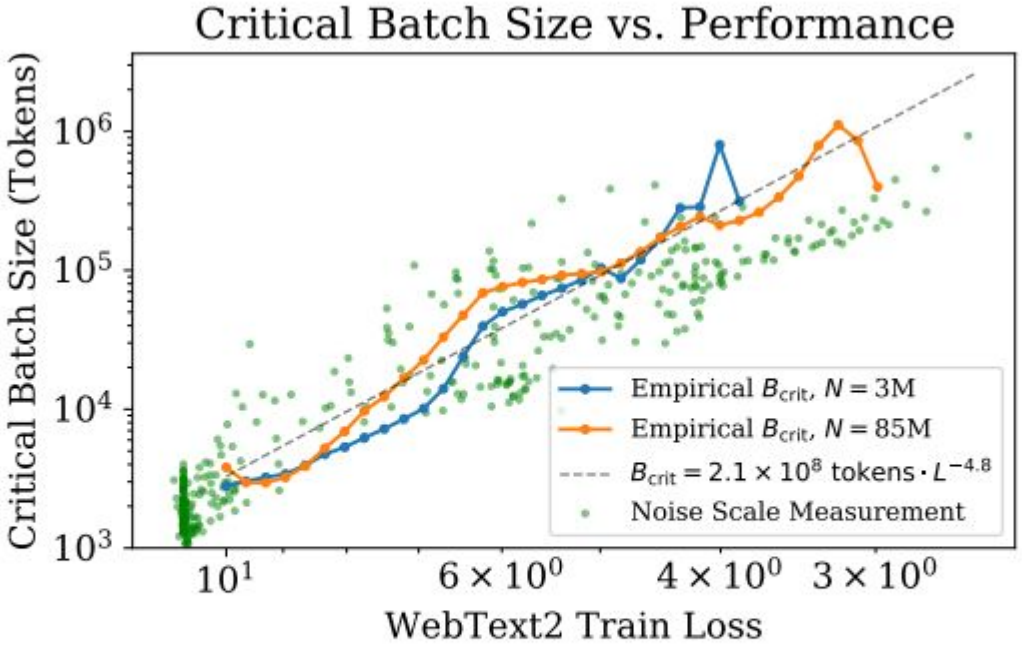


Figure 10 The critical batch size B_{crit} follows a power law in the loss as performance increase, and does not depend directly on the model size. We find that the critical batch size approximately doubles for every 13% decrease in loss. B_{crit} is measured empirically from the data shown in Figure 18, but it is also roughly predicted by the gradient noise scale, as in [MKAT18].

Optimal Allocation of the Compute Budget

Problem Statement

- Given a fixed compute budget Compute:
    - How should resources be split between model size, data, and training duration?

Key Result

- Compute-optimal strategy:
    - Train very large models
    - Use relatively modest datasets
    - Stop training early

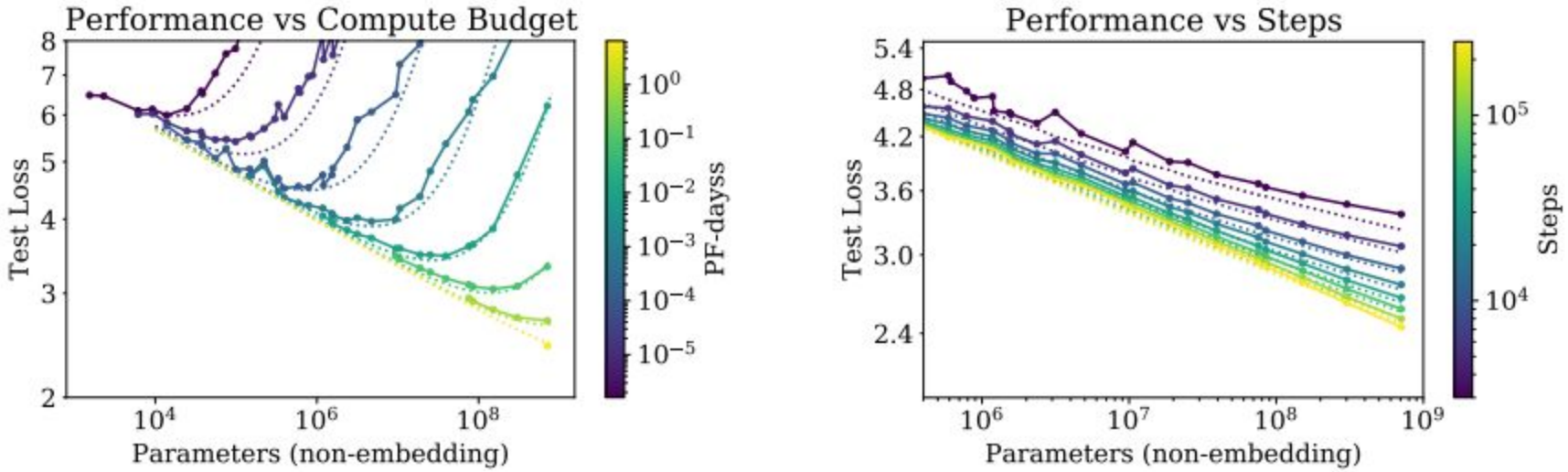


Figure 11 When we hold either total compute or number of training steps fixed, performance follows $L(N, S)$ from Equation (5.6). Each value of compute budget has an associated optimal model size that maximizes performance. Mediocre fits at small S are unsurprising, as the power-law equation for the learning curves breaks down very early in training.

Optimal Allocation of the Compute Budget

Why This Works

- Larger models extract more information per token
- Training to full convergence is compute-inefficient
- Marginal gains diminish faster with longer training than with larger models

Practical Impact

- Contradicts prior practice of training smaller models to convergence
- Directly informs large-scale model training strategies

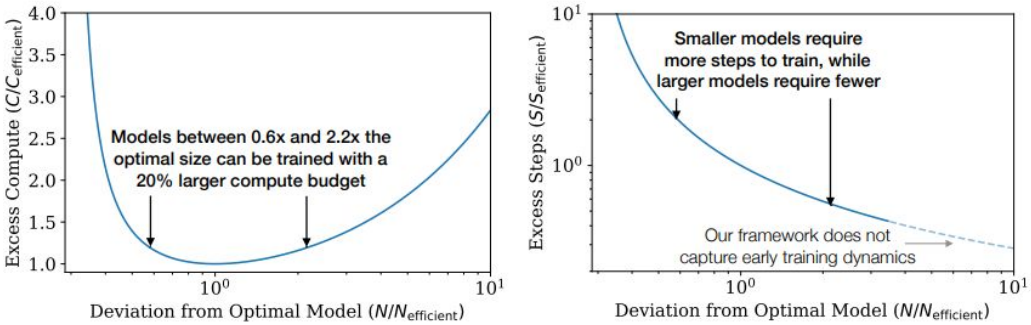


Figure 12 **Left:** Given a fixed compute budget, a particular model size is optimal, though somewhat larger or smaller models can be trained with minimal additional compute. **Right:** Models larger than the compute-efficient size require fewer steps to train, allowing for potentially faster training if sufficient additional parallelism is possible. Note that this equation should not be trusted for very large models, as it is only valid in the power-law region of the learning curve, after initial transient effects.

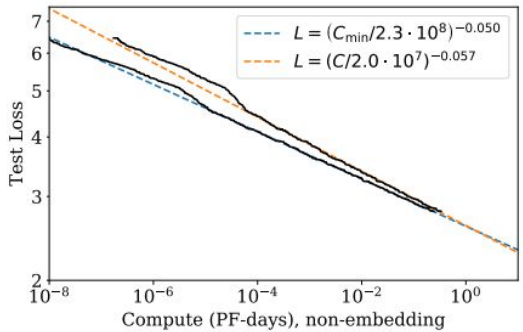


Figure 13 When adjusting performance to simulate training far below the critical batch size, we find a somewhat altered power law for $L(C_{\text{min}})$ when compared with the fully empirical results. The conspicuous lump at 10^{-5} PF-days marks the transition from 1-layer to 2-layer networks; we exclude 1-layer networks in the power-law fits. It is the $L(C_{\text{min}})$ trend that we expect to provide a reliable extrapolation for larger compute.

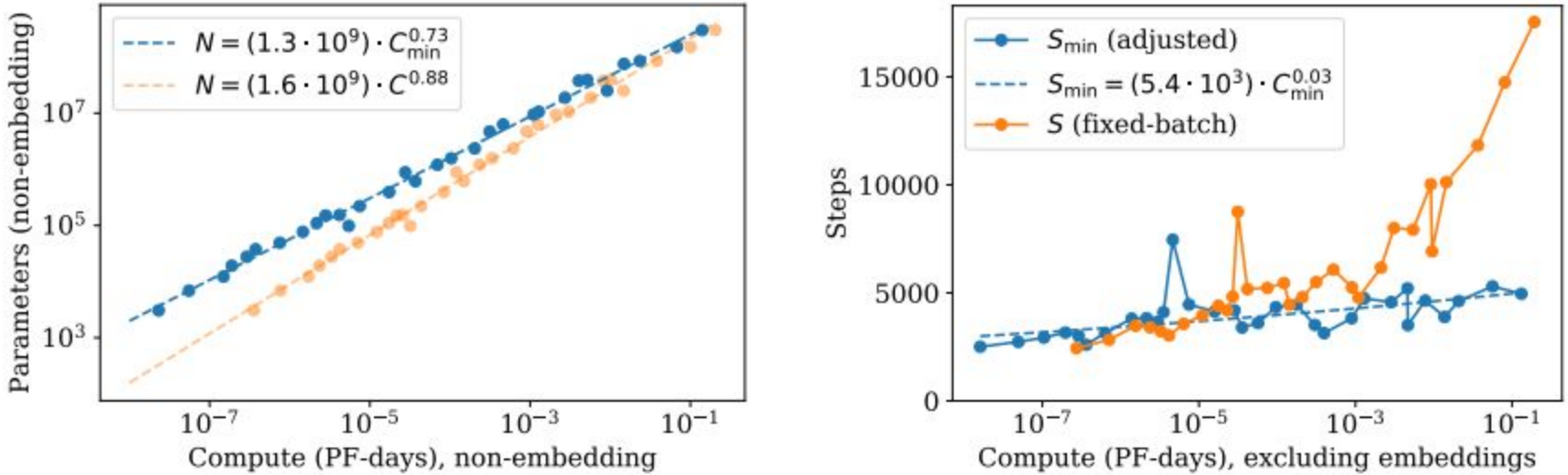


Figure 14 **Left:** Each value of the compute budget C_{min} has an associated optimal model size N. Optimal model size grows very rapidly with C_{min}, increasing by 5x for each 10x increase in compute. The number of data examples processed makes up the remainder of the increase, growing relatively modestly by only 2x. **Right:** The batch-adjusted number of optimization steps also grows very slowly, if at all, meaning that most of the growth in data examples processed can be used for increased batch sizes.

Related Work

Context

- Power-law behavior previously observed in:
    - Vision models
    - Optimization dynamics
    - Generalization studies

Contribution of This Paper

- First comprehensive, large-scale empirical study for language models
- Demonstrates consistency across many orders of magnitude
- Establishes scaling laws as a unifying framework

Discussion

- Language model performance scales predictably with resources
- Scaling laws enable:
    - Performance forecasting
    - Compute-efficient planning
- Larger models are more efficient learners

Broader Impact

- Explains why continued scaling of LMs yields improvements
- Provides foundation for modern large-model training approaches

Discussion

Limitations

- Results specific to autoregressive Transformers
- Ultimate limits of scaling not reached
- Data quality effects not deeply explored

Discussion

Final Takeaway

Core Message

- Scale dominates architecture
- Loss follows power laws
- Compute should favor large models + early stopping
- Scaling laws transform model training from guesswork into optimization

Q&A

Training Compute-Optimal Large Language Models

Introduction

Goal of the paper: To determine how to optimally allocate training compute between:

- Model size (number of parameters)
- Training data (number of tokens)

Motivation of the paper:

- Compute is finite and expensive
- Prior large language models scaled parameters aggressively but did not proportionally increase training tokens resulting in underperforming/undertrained models relative to their scale and wasted compute.

Introduction

Core question of the paper:

- For a fixed compute budget, what combination of Model size and Training data minimizes loss?

Introduction

Language model performance improves with:

- Increased model size
- Increased training data
- Increased training compute

Previous large models:

- Scaled parameter count rapidly
- Trained on relatively fixed token counts e.g (Gopher, GPT-3, Jurassic-1) ~ 300B Tokens

Introduction

Authors' key observation:

- Many models are undertrained relative to their size

Hypothesis:

- Training larger models on too few tokens leads to inefficient compute usage

Contribution:

- Empirically identify the compute-optimal tradeoff between model size and training data
- Validate findings by training a new model (Chinchilla)

Related Work

Prior work established:

- Power-law scaling of loss with model size and data
- Predictable performance improvement with more compute

Limitations of prior approaches:

- Did not explicitly optimize under a fixed compute constraint
- Did not vary training duration sufficiently across model size

Related Work

Key gap:

- Lack of empirical analysis on how long models should be trained at each size

This paper’s advancement:

- Explicitly studies loss as a function of both parameters and tokens
- Focuses on compute-constrained optimization

Problem Formulation

Define key quantities:

- N: number of model parameters
- D: number of training tokens
- C: total training compute
- L(N, D): validation loss

Training compute approximation:

- $C \propto N \times D$

Problem Formulation

Optimization objective:

- Minimize L(N, D) subject to fixed C

Core trade off:

- Increasing N requires reducing D (and vice versa) under fixed compute

Experimental Methodology

Trained a large number of transformer models:

- Wide range of model sizes (70M to 10B)
- Wide range of training token counts

Each model evaluated at multiple training points - Result:

- Dense empirical measurements of loss across (N, D) space

Purpose:

- Identify which combinations of parameters and tokens are compute-optimal

Approach 1

- Fix model size, vary number of training tokens

What is varied

- Fixed model sizes (70M → 10B parameters)
- Multiple training horizons per model (different token counts)

What is measured

- Training loss throughout the entire training run
- Loss interpolated as a function of FLOPs

Approach 1

- Train each model with 4 different learning-rate schedules
- Smooth and interpolate loss curves
- For each FLOP value, find the lowest loss achieved by any model
- Extract the envelope of minimal loss per FLOP
- Fit power laws for optimal model size and tokens

Findings

- Equal scaling of model size and data is compute-optimal

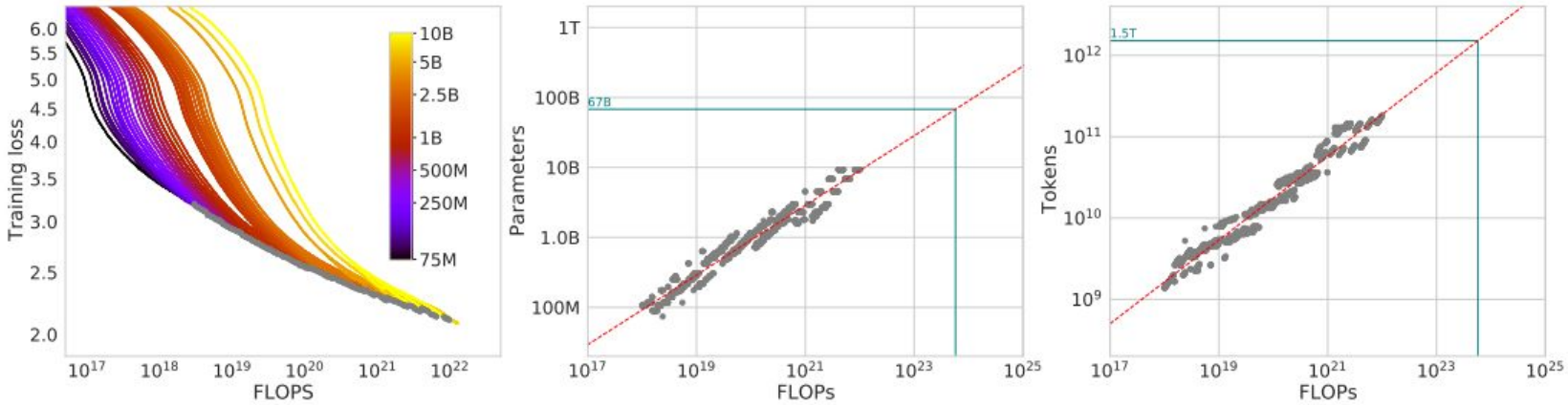


Figure 2 | **Training curve envelope.** On the **left** we show all of our different runs. We launched a range of model sizes going from 70M to 10B, each for four different cosine cycle lengths. From these curves, we extracted the envelope of minimal loss per FLOP, and we used these points to estimate 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* (5.76×10^{23}).

Approach 2

Fix compute budget, vary model size

What is varied

- Model size (up to 16B parameters)
- Training tokens automatically adjusted to keep FLOPs constant

Key Question:

For a fixed compute budget, which model size gives the lowest final loss?

Approach 2

- Choose several fixed FLOP budgets
- For each budget, train many models of different sizes
- Adjust token count so all runs use identical FLOPs
- Plot final loss vs model size (IsoFLOP curves)
- Fit a parabola to find the loss minimum

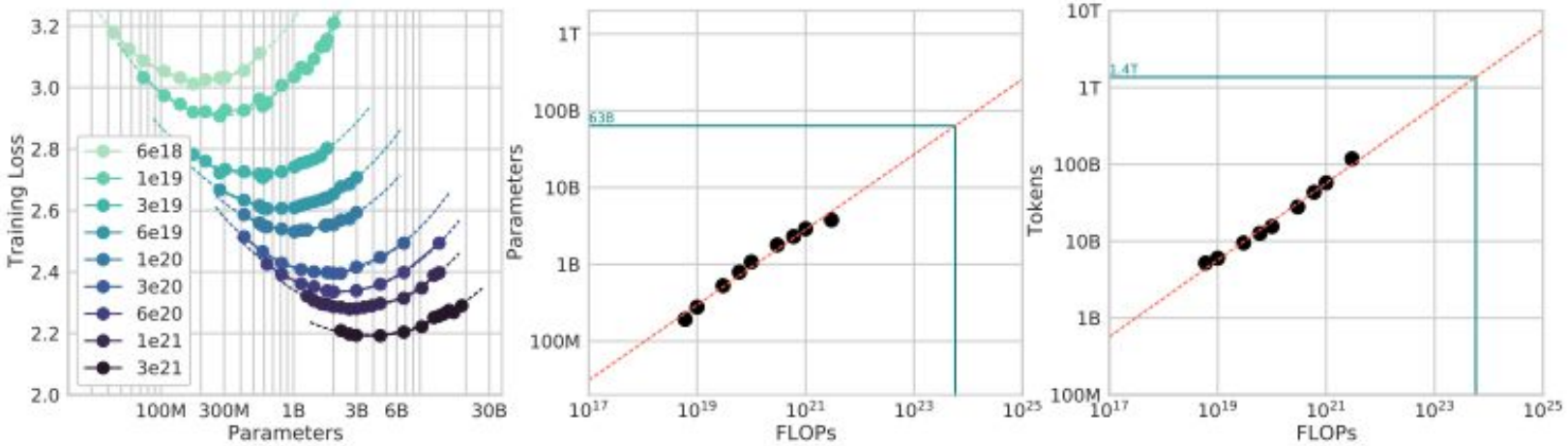


Figure 3 | **IsoFLOP curves.** For various model sizes, we choose the number of training tokens such that the final FLOPs is a constant. The cosine cycle length is set to match the target FLOP count. We find a clear valley in loss, meaning that for a given FLOP budget there is an optimal model to train (**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*.

Approach 3

Fit a global mathematical model for loss as a function of:

- Model size
- Training tokens

Loss Decomposition

- Irreducible data entropy
- Error due to finite model capacity
- Error due to limited training data

Approach 3

Loss is modeled as:

- A constant term (ideal data entropy)
- A term decreasing with model size
- A term decreasing with training tokens

All model runs are used simultaneously

Robust fitting using:

- Log-loss
- Huber loss (to reduce sensitivity to outliers)

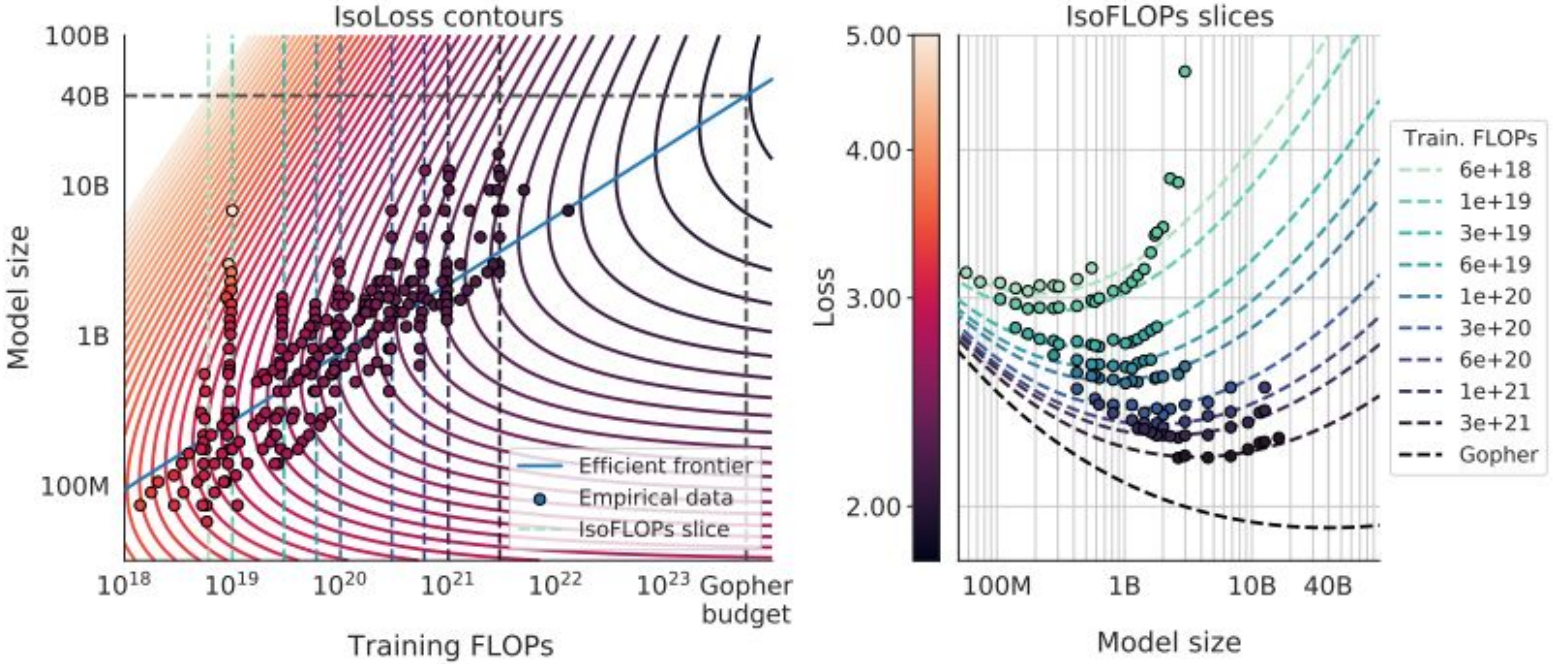


Figure 4 | **Parametric fit.** We fit a parametric modelling of the loss $\hat{L}(N, D)$ and display contour (**left**) and isoFLOP slices (**right**). For each isoFLOP slice, we include a corresponding dashed line in the left plot. In the left plot, we show the efficient frontier in blue, which is a 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.

Key Result

All three estimation methods agree:

- Compute-optimal training requires increasing training tokens proportionally with model size

Main empirical finding:

- Larger models trained on too few tokens perform worse than smaller models trained longer

Conclusion:

- Previous large models were systematically undertrained

Chinchilla (Motivation)

Purpose:

- Empirically validate compute-optimal scaling predictions

Approach:

- Use same compute budget as a previously trained large model
- Allocate compute according to optimal N–D tradeoff

Key design choice:

- Smaller model
- Significantly more training tokens

Chinchilla (Training)

Chinchilla characteristics:

- Fewer parameters than prior large models
- Trained on substantially more tokens
- Total training compute approximately unchanged

Training procedure:

- Same architecture class
- Standard language modeling objective

Goal:

- Test whether compute-optimal allocation improves performance

Chinchilla (Results)

Chinchilla achieves:

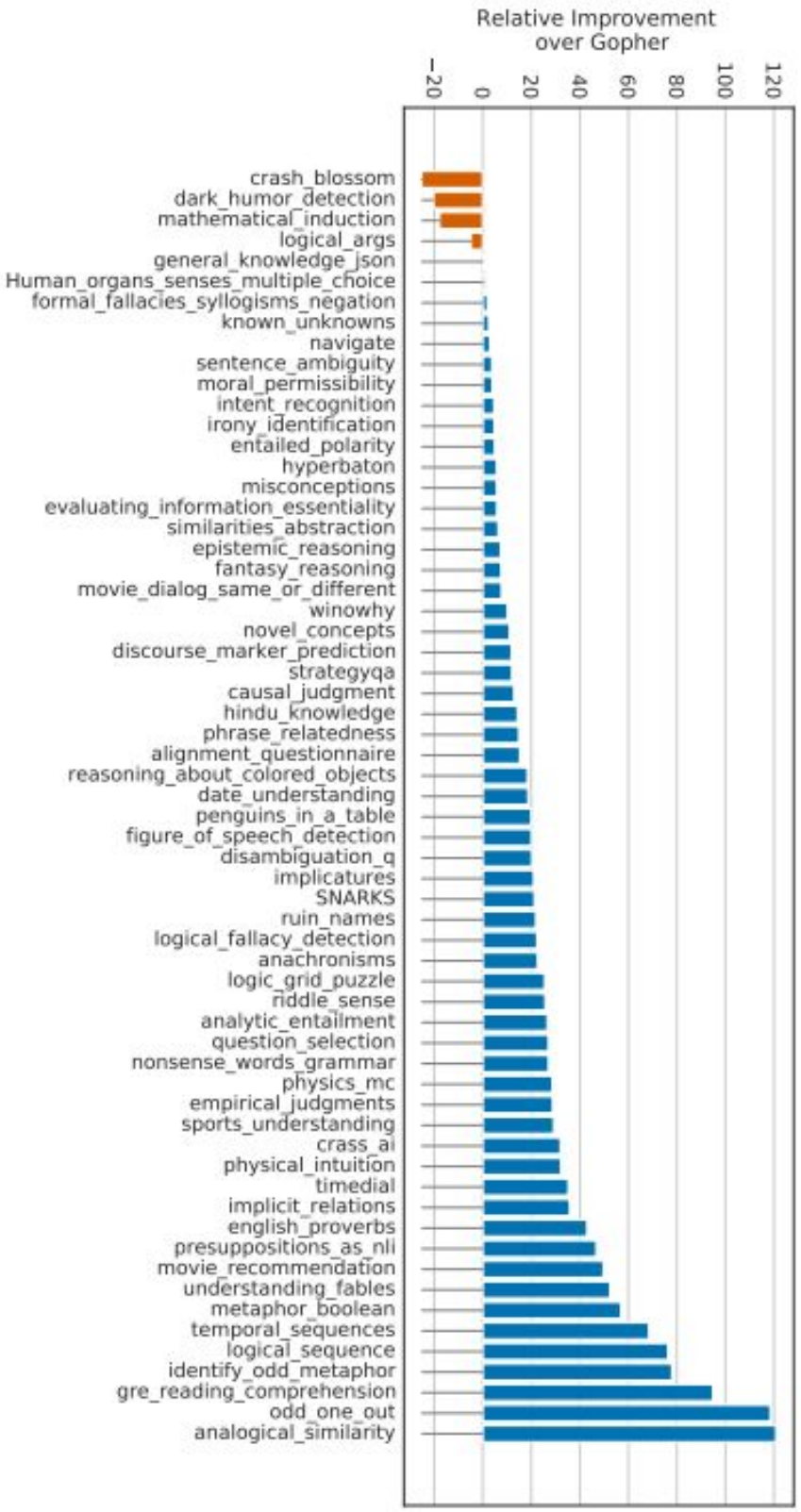
- Lower validation loss
- Better performance on downstream tasks

Outperforms:

- Larger models trained with similar compute

Empirical confirmation:

- Optimal compute allocation matters more than parameter count alone


Relative Improvement
over Gopher
−20
0
20
40
60
80
100
120
crash_blossom
dark_humor_detection
mathematical_induction
logical_args
general_knowledge_json
Human_organs_senses_multiple_choice
formal_fallacies_syllogisms_negation
known_unknowns
navigate
sentence_ambiguity
moral_permissibility
intent_recognition
irony_identification
entailed_polarity
hyperbaton
misconceptions
evaluating_information_essentiality
similarities_abstraction
epistemic_reasoning
fantasy_reasoning
movie_dialog_same_or_different
winowhy
novel_concepts
discourse_marker_prediction
strategyqa
causal_judgment
hindu_knowledge
phrase_relatedness
alignment_questionnaire
reasoning_about_colored_objects
date_understanding
penguins_in_a_table
figure_of_speech_detection
disambiguation_q
implicatures
SNARKS
ruin_names
logical_fallacy_detection
anachronisms
logic_grid_puzzle
riddle_sense
analytic_entailment
question_selection
nonsense_words_grammar
physics_mc
empirical_judgments
sports_understanding
crass_ai
physical_intuition
timedial
implicit_relations
english_proverbs
presuppositions_as_nli
movie_recommendation
understanding_fables
metaphor_boolean
temporal_sequences
logical_sequence
identify_odd_metaphor
gre_reading_comprehension
odd_one_out
analogical_similarity

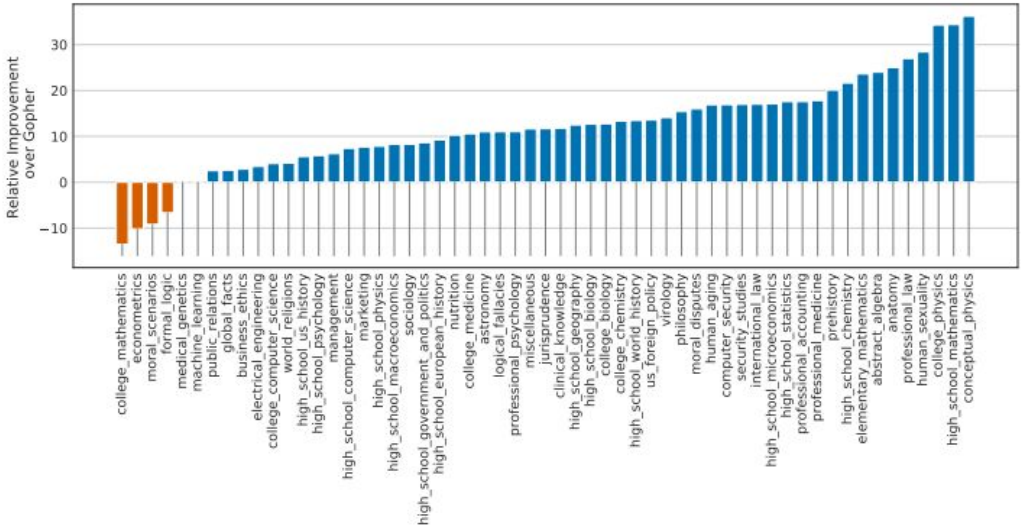


Figure 6 | **MMLU results compared to *Gopher*** We find that *Chinchilla* outperforms *Gopher* by 7.6% on average (see Table 6) in addition to performing better on 51/57 individual tasks, the same on 2/57, and worse on only 4/57 tasks.

	Chinchilla	*Gopher*	GPT-3	MT-NLG 530B
LAMBADA Zero-Shot	**77.4**	74.5	76.2	76.6
RACE-m Few-Shot	**86.8**	75.1	58.1	-
RACE-h Few-Shot	**82.3**	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.

Discussion

Main observations:

- Increasing parameters without increasing data is inefficient
- Training duration is critical for performance

Implication:

- Model size alone is not a reliable indicator of quality

Compute-optimal training leads to:

- Better performance
- More efficient use of resources

Conclusions

Key conclusions:

- Most large language models are undertrained
- Optimal scaling requires:
  - More data
  - Longer training
  - Balanced growth of parameters and tokens

Conclusions

Chinchilla demonstrates:

- Practical benefits of compute-optimal scaling

Overall message:

- Efficient compute usage is essential for future LLM development

Final Takeaway

- Training compute ≈ model size × training tokens
- Larger models need more data to reach their potential
- Compute-optimal scaling improves:
    - Performance
    - Efficiency
    - Practical usability
- Shift in perspective:
    - From “bigger models” → “better-trained models”

Final Takeaway

Scaling Laws

- Explores how loss improves with model size (N), dataset size (D), or compute (C) individually.
- Finds predictable power-law trends.
- Focus: *One factor at a time.*

Compute-Optimal Frontier

- Determines the best model + dataset combination for a fixed compute budget.
- Uses a parametric fit to define the efficient frontier.
- Focus: *Optimal allocation of compute for best performance.*

Final Takeaway

- Scaling Laws: “Bigger model, more data, or more compute improves loss.”

- Compute-Optimal Frontier : “Given limited compute, what mix of model + data gives the best result?”
- Both papers show that model performance improves in predictable ways with size, data, and compute.
- Provide formulas and empirical laws to plan efficient experiments.

Q&A