Training Compute-Optimal Large Language Models

By Anika Maskara and Simon Park
10/24/2022
Outline

1. Introduction

2. Initial Scaling Law (Kaplan et al., 2020)

3. Modified Scaling Law (Hoffman et al., 2022)

4. Chinchilla (Hoffman et al., 2022)

5. Beyond Scaling Law
Language Models have been Getting Bigger…

[Image Source]  [Image Source]
.....a lot bigger

Megatron Turing NLG (2021)

GPT-3 (2020)

Megatron LM (2021)

T5 (2020)

Turing - NLG (2020)
…..a lot bigger
…..a lot bigger

BERT large (2018)

Megatron Turing NLG (2021)

GPT-3 (2020)

T5 (2020)

Turing - NLG (2020)
…..a lot bigger
Q1: Why do we care about studying scaling law of LLMs?
Common carbon footprint benchmarks

- Roundtrip flight b/w NY and SF (1 passenger): 1,984 lbs
- Human life (avg 1 year): 11,023 lbs
- American life (avg 1 year): 36,156 lbs
- US car including fuel (avg 1 lifetime): 126,000 lbs
- GPT-3: 1,216,950 lbs
- T5: 103,617 lbs

Created with Datawrapper

[Data Source: (Strubell et al., 2019)] [Data Source: (Patterson et al. 2021)]
Big Models Require Big Pockets, and not just at training

Sources estimate that training GPT-3 required at least $4,600,000

That’s a lot, but at least few-shot means the model only has to be trained once?
Big Models Require Big Pockets, and not just at training

Sources estimate that training GPT-3 required at least $4,600,000

That’s a lot, but at least few-shot means the model only has to be trained once?
Yes, but inference is still expensive

One recent estimate pegged the cost of running GPT-3 on a single AWS web server to cost $87,000 a year at minimum
Our assumption

bigger models $\rightarrow$ better performance

This may be true, but is increasing model size the most efficient way of improving performance?
Understanding FLOPs
(floating point operations)

\[ C \sim 6ND \]

- \( C \) = number of FLOPs (computations)
- \( N \) = number of model parameters
- \( D \) = amount of training data
Understanding FLOPs — Forward Pass

Matrix multiplication (e.g., attention QKV projection) requires **2 * size of matrix** (1 for multiplication, 1 for addition)

\[
\begin{bmatrix}
A_{11} & A_{12} & \cdots & A_{1n} \\
A_{21} & A_{22} & \cdots & A_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
A_{m1} & A_{m2} & \cdots & A_{mn}
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{bmatrix} =
\begin{bmatrix}
A_{11}x_1 + A_{12}x_2 + \cdots + A_{1n}x_n \\
A_{21}x_1 + A_{22}x_2 + \cdots + A_{2n}x_n \\
\vdots \\
A_{m1}x_1 + A_{m2}x_2 + \cdots + A_{mn}x_n
\end{bmatrix}
\]
Understanding FLOPs — Forward Pass

N is roughly the sum of size of all matrices

FLOPs for forward pass on a single token is roughly $2N$

FLOPs for forward pass for the entire dataset is roughly $2ND$
Understanding FLOPs — Backward Pass

Backward pass needs to calculate the derivative of loss with respect to each hidden state and for each parameter.

FLOPs for backward pass is roughly twice of forward pass.

FLOPs for backward pass for the entire dataset is roughly 4ND.
Understanding FLOPs

\[ C \sim 6ND \]

If we had a computational budget on C, Increasing model size \( N = \) Decreasing dataset size \( D \)

But we also expect more data \( \rightarrow \) better performance
Key Question

Increase N $\rightarrow$ better performance

Increase D $\rightarrow$ better performance

But we have a **budget** on $C \sim 6ND$
Key Question

To maximize model performance,

how should we allocate C to N and D?
Key Question

To maximize model performance, how should we allocate $C$ to $N$ and $D$?

$$N_{opt}(C), D_{opt}(C) = \arg\min_{N,D \text{ s.t. } \text{FLOPs}(N,D) = C} L(N, D)$$

[Equation Source: (Hoffman et al., 2022)]
Key Question (rephrased)

What is the relationship between loss and $N, D$?

\[
\hat{L}(N, D) \triangleq E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}
\]

[Equation Source: (Hoffman et al., 2022)]
Is Power-Law the best fit?

Based on **empirical observation**

No theoretical background

(Hoffman et al.) also observe concavity in their model at high compute budgets, suggesting the need for a more detailed model

[Figure Source: (Hoffman et al., 2022)]
Outline

1. Introduction

2. Initial Scaling Law (Kaplan et al., 2020)

3. Modified Scaling Law (Hoffman et al., 2022)

4. Chinchilla (Hoffman et al., 2022)

5. Beyond Scaling Law
Scaling Laws for Neural Language Models

Jared Kaplan *
Johns Hopkins University, OpenAI
jaredk@jhu.edu

Sam McCandlish*
OpenAI
sam@openai.com

Tom Henighan
OpenAI
henighan@openai.com

Tom B. Brown
OpenAI
tom@openai.com

Benjamin Chess
OpenAI
bchess@openai.com

Rewon Child
OpenAI
rewon@openai.com

Scott Gray
OpenAI
scott@openai.com

Alec Radford
OpenAI
alec@openai.com

Jeffrey Wu
OpenAI
jeffwu@openai.com

Dario Amodei
OpenAI
damodei@openai.com
Training Details

Model: **Decoder-only Transformer** (N = 0.7K ~ 1.5B params)

Dataset: WebText2 (D = 22B tokens)

Batch Size (B): 0.5M

Step Size (S): 0.25M

Optimizer: Adam (+ Adafactor)

Learning rate: 3000 warmup steps, max LR = 2e-3, **cosine decay to 0**

Loss: **autoregressive cross-entropy loss** over 1024-token context
Main Results

- Performance scales with **model size** (N) and **dataset size** (D)
- If assuming **fixed batch size**, D should increase by **1.7x** when N increases by **2x**
- If assuming **optimal batch size**, D should increase by **1.3x** when N increases by **2x**
Main Results

- Performance scales with \textbf{model size} (N) and \textbf{dataset size} (D).
- If assuming \textbf{fixed batch size},
  \[ D \text{ should increase by } 1.7x \text{ when } N \text{ increases by } 2x \]
- If assuming \textbf{optimal batch size},
  \[ D \text{ should increase by } 1.3x \text{ when } N \text{ increases by } 2x \]

Commonly Cited Result
Outline

2. Initial Scaling Law (Kaplan et al., 2020)
   a. Fixed Batch Size Case
   b. Optimal Batch Size Case
   c. Limitations
Experiment 1 : Change D

Fix $N = 1.5B$

Fix $B = 0.5M$

Vary $D = 21M \sim 22B$ (fixed subsets of WebText2)

**Early stop** whenever loss ceased to decrease
Experiment 2 : Change N

Fix D = 22B
Fix B = 0.5M
Fix S = 0.25M

Vary N = 0.7K ~ 1.5B

Train until convergence
Results of Experiment 1, 2

![Graphs showing the relationship between dataset size and parameters](image)

Figure Source: (Kaplan et al., 2020)
Experiment 3 : Change both D and N

Fix B = 0.5M

Vary N = 0.4M ~ 0.7B

Vary D = 21M ~ 22B

Early stop whenever loss ceased to decrease
Result of Experiment 3

Data Size Bottleneck

Test Loss

Data Size
- 21M
- 43M
- 86M
- 172M
- 344M
- 688M
- 1.4B
- 22.0B

[Figure Source: (Kaplan et al., 2020)]
Conclusion

D should increase by $1.7x$ when N increases by $2x$
Conclusion

\( D \) should increase by 1.7\( x \) when \( N \) increases by 2\( x \)

But what if we have a compute budget?
Outline

2. Initial Scaling Law ([Kaplan et al., 2020](#))
   a. Fixed Batch Size Case
   b. **Optimal Batch Size Case**
   c. Limitations
Compute-Optimal Batch Size

Critical Batch Size dependent on the loss (not $N, D$) (McCandlish et al., 2018)
Compute-Optimal Batch Size

Critical Batch Size dependent on the loss (not N, D) (McCandlish et al., 2018)

(e.g., ~1M at the end of training for the best models in Experiments 1~3)
Compute-Optimal Batch Size

Critical Batch Size dependent on the loss (not N, D) (McCandlish et al., 2018)
(e.g., ~1M at the end of training for the best models in Experiments 1~3)

B << Critical Batch Size: FLOP minimized
Compute-Optimal Batch Size

Critical Batch Size dependent on the loss (not N, D) (McCandlish et al., 2018) (e.g., ~1M at the end of training for the best models in Experiments 1~3)

B << Critical Batch Size: FLOP minimized

B >> Critical Batch Size: Training Time (i.e., step size) minimized
Compute-Optimal Batch Size

Critical Batch Size dependent on the loss (not N, D) (McCandlish et al., 2018)
(e.g., ~1M at the end of training for the best models in Experiments 1~3)

B << Critical Batch Size: FLOP minimized
B >> Critical Batch Size: Training Time (i.e., step size) minimized
B == Critical Batch Size: Trade-off
Compute-Optimal Batch Size

Critical Batch Size vs. Performance

[Figure Source: (Kaplan et al., 2020)]
Revisiting Experiment 3

Assuming we ran Experiment 3 again with $B \ll \text{Critical Batch Size}$,

It is possible to estimate the minimum FLOP ($C_{\text{min}}$) to reach the same loss

$$C_{\text{min}}(C) \equiv \frac{C}{1 + \frac{B}{B_{\text{crit}}(L)}}$$
Revisiting Experiment 3

Assuming we ran Experiment 3 again with $B << \text{Critical Batch Size}$,

It is possible to estimate the \text{minimum} FLOP ($C_{\text{min}}$) to reach the same loss

And the \text{optimal} model size $N$ for the \text{target} $C_{\text{min}}$
Conclusion

\[ N \propto C_{min}^{0.73} \quad D \propto C_{min}^{0.27} \]
Conclusion

\[ N \propto C_{\text{min}}^{0.73} \quad D \propto C_{\text{min}}^{0.27} \]

D should increase by \textbf{1.3x} when N increases by \textbf{2x}
Outline

2. Initial Scaling Law (Kaplan et al., 2020)
   a. Fixed Batch Size Case
   b. Optimal Batch Size Case
   c. Limitations
Limitations

1. Needs to **adjust batch size** during training

2. Results were based on **early stop**, while **learning rate schedule** was calculated for the full 250K steps
Learning Rate Schedule and Early Stop

![Graph showing learning rate schedule and step size. The graph has a downward curve starting from a higher rate at step size 0 and decreasing to a lower rate by step size 300,000.]
Learning Rate Schedule and Early Stop

![Graph showing learning rate schedule and early stop](image-url)
Learning Rate Schedule and Early Stop
Outline

1. Introduction
2. Initial Scaling Law (Kaplan et al., 2020)
3. Modified Scaling Law (Hoffman et al., 2022)
4. Chinchilla (Hoffman et al., 2022)
5. Beyond Scaling Law
Training Compute-Optimal Large Language Models


*Equal contributions
Given a particular FLOPs (Floating Point Operation) budget, how should one trade-off model size and training data?

\[ N_{\text{opt}}(C), D_{\text{opt}}(C) = \arg\min_{N,D \text{ s.t. FLOPs}(N,D) = C} L(N, D) \]

C = number of FLOPs (computations)
N = number of model parameters
D = amount of training data
N, D should scale at **same rate**

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

[Table Source: (Hoffman et al., 2022)]
Q2: How do the conclusions of (Kaplan et al.) and (Hoffman et al.) differ? What caused the differences?
Early Stopping leads to Underperformance

(Figure Source: (Hoffman et al., 2022))
(Kaplan et al.) vs (Hoffman et al.)

(Kaplan et al.)

Learning rate - based on 250K steps
Batch Size - based on $B \leq \text{critical batch size}$

(Hoffman et al.)

Learning rate - based on actual step size
Batch Size - fixed
3. Modified Scaling Law (Hoffman et al., 2022)
   a. Approach 1
   b. Approach 2
   c. Approach 3
   d. Results
Approach 1: Fix N and vary D

For each N, train 4 different models with different D

Interpolate these curves to get a continuous mapping

For each FLOPs, pick the model with the lowest training loss

C = number of FLOPs (computations)
N = number of model parameters
D = amount of training data
Figure Source: (Hoffman et al., 2022)
Approach 1: Fix $N$ and Vary $D$

For each $N$, train 4 different models with different $D$

Interpolate these curves to get a continuous mapping

For each FLOPs, pick the model with the lowest training loss

Fit a power law relationship between $C$ and $N$, $D$
## Results of Approach 1

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

[Table Source: (Hoffman et al., 2022)]
Outline

3. Modified Scaling Law (Hoffman et al., 2022)
   a. Approach 1
   b. Approach 2
   c. Approach 3
   d. Results
Approach 2: IsoFLOP Profiles

For each FLOPs budget $C$, train models of different size $N$.

For each model, choose the appropriate $D$ such that $C \sim 6ND$.

E.g., bigger models are trained on less data to meet FLOPs constraint.

$C = \text{number of FLOPs (computations)}$
$N = \text{number of model parameters}$
$D = \text{amount of training data}$
Approach 2: IsoFLOP Profiles

For each FLOPs budget \( C \), train models of different size \( N \)

For each model, choose the appropriate \( D \) such that \( C \sim 6ND \)

E.g., bigger models are trained on less data to meet FLOPs constraint

Fit a power law relationship between \( C \) and \( N, D \)
Figure Source: (Hoffman et al., 2022)
## Results of Approach 2

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

[Table Source: (Hoffman et al., 2022)]
Outline

3. Modified Scaling Law (Hoffman et al., 2022)
   a. Approach 1
   b. Approach 2
   c. Approach 3
   d. Results
Approach 3: Parametric Loss Function

\[ \hat{L}(N, D) \triangleq E + \frac{A}{N^\alpha} + \frac{B}{D^\beta} \]

1. **E**: loss of ideal generative model (entropy of natural language)
2. **N**: larger model \(\rightarrow\) better performance
3. **D**: larger dataset \(\rightarrow\) better performance
Determining Coefficients

1. Choose initial values of $E, A, B, \alpha, \beta$ from a grid of values
2. Find the **Huber loss** based on the predicted log loss of the model on $(N, D)$ and observed log loss (data from Approach 1, 2)
3. Iteratively, run the L-BFGS algorithm (some variant of Gradient Descent)
# Results of Approach 3

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

[Table Source: (Hoffman et al., 2022)]
Outline

3. Modified Scaling Law (Hoffman et al., 2022)
   a. Approach 1
   b. Approach 2
   c. Approach 3
   d. Results
## Results of Approach 1 ~ 3

![Table Source: (Hoffman et al., 2022)]

<table>
<thead>
<tr>
<th>Approach</th>
<th>Coeff. $a$ where $N_{opt} \propto C^a$</th>
<th>Coeff. $b$ where $D_{opt} \propto C^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Minimum over training curves</td>
<td>0.50 (0.488, 0.502)</td>
<td>0.50 (0.501, 0.512)</td>
</tr>
<tr>
<td>2. IsoFLOP profiles</td>
<td>0.49 (0.462, 0.534)</td>
<td>0.51 (0.483, 0.529)</td>
</tr>
<tr>
<td>3. Parametric modelling of the loss</td>
<td>0.46 (0.454, 0.455)</td>
<td>0.54 (0.542, 0.543)</td>
</tr>
<tr>
<td>Kaplan et al. (2020)</td>
<td>0.73</td>
<td>0.27</td>
</tr>
</tbody>
</table>
Today’s models are **overparameterized** and **undertrained**

[Figure Source: (Hoffman et al., 2022)]
Outline

1. Introduction
2. Initial Scaling Law (Kaplan et al., 2020)
3. Modified Scaling Law (Hoffman et al., 2022)
4. Chinchilla (Hoffman et al., 2022)
5. Beyond Scaling Law
Given Gopher’s compute budget, can we train a more computationally efficient model?
Chinchilla is small(er)

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

[Image Source] [Table Source: (Hoffman et al., 2022)]
Comparison with Gopher

N smaller by 4x, D larger by 4x

Less compute for inference and fine-tuning
But also stronger performance
Performance of Chinchilla
Evaluations Tasks for Chinchilla

- Language Modelling
- MMLU
- Reading Comprehension
- BIG-bench
- Common Sense
- Closed Book QA
- Gender Bias and Toxicity
Evaluations Tasks for Chinchilla

- Language Modelling
- MMLU
- Reading Comprehension
- BIG-bench
- Common Sense
- Closed Book QA
- Gender Bias and Toxicity
Language Modelling

Measure test perplexity (in bits-per-byte) of 20 datasets from the Pile (Gao et al., 2021)

Chinchilla outperforms Gopher on all 20 datasets

Note: because of large training data, there is an increased risk of train/test leak
## Analysis Per Dataset

<table>
<thead>
<tr>
<th>Subset</th>
<th>Chinchilla (70B)</th>
<th>Gopher (280B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>pile_cc</td>
<td>0.667</td>
<td>0.691</td>
</tr>
<tr>
<td>pubmed_abstracts</td>
<td>0.559</td>
<td>0.578</td>
</tr>
<tr>
<td>stackexchange</td>
<td>0.614</td>
<td>0.641</td>
</tr>
<tr>
<td>github</td>
<td>0.337</td>
<td>0.377</td>
</tr>
<tr>
<td>openwebtext2</td>
<td>0.647</td>
<td>0.677</td>
</tr>
<tr>
<td>arxiv</td>
<td>0.627</td>
<td>0.662</td>
</tr>
<tr>
<td>uspto_backgrounds</td>
<td>0.526</td>
<td>0.546</td>
</tr>
<tr>
<td>freelaw</td>
<td>0.476</td>
<td>0.513</td>
</tr>
<tr>
<td>pubmed_central</td>
<td>0.504</td>
<td>0.525</td>
</tr>
<tr>
<td>dm_mathematics</td>
<td>1.111</td>
<td>1.142</td>
</tr>
<tr>
<td>hackernews</td>
<td>0.859</td>
<td>0.890</td>
</tr>
<tr>
<td>nih_exporter</td>
<td>0.572</td>
<td>0.590</td>
</tr>
<tr>
<td>opensubtitles</td>
<td>0.871</td>
<td>0.900</td>
</tr>
<tr>
<td>europarl</td>
<td>0.833</td>
<td>0.938</td>
</tr>
<tr>
<td>books3</td>
<td>0.675</td>
<td>0.712</td>
</tr>
<tr>
<td>philpapers</td>
<td>0.656</td>
<td>0.695</td>
</tr>
<tr>
<td>gutenberg_pg_19</td>
<td>0.548</td>
<td>0.656</td>
</tr>
<tr>
<td>bookcorpus2</td>
<td>0.714</td>
<td>0.741</td>
</tr>
<tr>
<td>ubuntu irc</td>
<td>1.026</td>
<td>1.090</td>
</tr>
</tbody>
</table>

[Table Source: (Hoffman et al., 2022)]
Evaluations Tasks for Chinchilla

- Language Modelling
- MMLU
- Reading Comprehension
- BIG-bench
- Common Sense
- Closed Book QA
- Gender Bias and Toxicity
MMLU — Massive Multitask Language Understanding

Answer exam-like multiple choice questions on 57 subjects (Hendrycks et al., 2020)

E.g., college mathematics, high school physics, professional law
Example Data from MMLU

An observational study in diabetics assesses the role of an increased plasma fibrinogen level on the risk of cardiac events. 130 diabetic patients are followed for 5 years to assess the development of acute coronary syndrome. In the group of 60 patients with a normal baseline plasma fibrinogen level, 20 develop acute coronary syndrome and 40 do not. In the group of 70 patients with a high baseline plasma fibrinogen level, 40 develop acute coronary syndrome and 30 do not. Which of the following is the best estimate of relative risk in patients with a high baseline plasma fibrinogen level compared to patients with a normal baseline plasma fibrinogen level?

(A) (40/30)/(20/40)
(B) (40*40)/(20*30)
(C) (40*70)/(20*60)
(D) (40/70)/(20/60)

Figure 69: A Virology example.

[Figure Source: (Hendrycks et al., 2020)]
Chinchilla Outperforms Gopher on Average

<table>
<thead>
<tr>
<th>Model</th>
<th>5-shot Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>25.0%</td>
</tr>
<tr>
<td>Average human rater</td>
<td>34.5%</td>
</tr>
<tr>
<td>GPT-3 175B</td>
<td>43.9%</td>
</tr>
<tr>
<td>Gopher 280B</td>
<td>60.0%</td>
</tr>
<tr>
<td>Chinchilla 70B</td>
<td>67.6%</td>
</tr>
<tr>
<td>Average human expert performance</td>
<td>89.8%</td>
</tr>
</tbody>
</table>

Table Source: (Hoffman et al., 2022)
Chinchilla outperforms Gopher on 51 tasks

Achieves a similar performance on 2 tasks

Underperforms Gopher on 4 tasks (college mathematics, econometrics, moral scenarios, formal logic)
Analysis Per Task

[Figure Source: (Hoffman et al., 2022)]
Analysis Per Task

Chinchilla achieves >90% accuracy on 4 tasks

High school government and politics, international law, sociology, US foreign policy

First model to achieve 90% accuracy on a particular subject
Evaluations Tasks for Chinchilla

- Language Modelling
- MMLU
- Reading Comprehension
- BIG-bench
- Common Sense
- Closed Book QA
- Gender Bias and Toxicity
Reading Comprehension

Answer a fill-in-the-blank question on a passage

LAMBADA (Paperno et al., 2016): novel excerpt

RACE-M, RACE-H (Lai et al., 2017): middle-, high-school exam questions
Example Data from LAMBADA

*Context:* The battery on Logan’s radio must have been on the way out. So he told himself. There was no other explanation beyond Cygan and the staff at the White House having been overrun. Lizzie opened her eyes with a flutter. They had been on the icy road for an hour without incident.

*Target sentence:* Jack was happy to do all of the ______.

*Target word:* driving

[Figure Source: (Paperno et al., 2016)]
Example Data from RACE-M, RACE-H

Evidence: “The park is open from 8 am to 5 pm.”

Question: The park is open for __ hours a day.

Options: A. eight   B. nine   C. ten   D. eleven
## Chinchilla Outperforms Gopher

<table>
<thead>
<tr>
<th></th>
<th>70B</th>
<th>280B</th>
<th>175B</th>
<th>530B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chinchilla</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAMBADA Zero-Shot</td>
<td><strong>77.4</strong></td>
<td>74.5</td>
<td>76.2</td>
<td>76.6</td>
</tr>
<tr>
<td>RACE-m Few-Shot</td>
<td><strong>86.8</strong></td>
<td>75.1</td>
<td>58.1</td>
<td>-</td>
</tr>
<tr>
<td>RACE-h Few-Shot</td>
<td><strong>82.3</strong></td>
<td>71.6</td>
<td>46.8</td>
<td>47.9</td>
</tr>
<tr>
<td><strong>Gopher</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAMBADA Zero-Shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RACE-m Few-Shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RACE-h Few-Shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GPT-3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAMBADA Zero-Shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RACE-m Few-Shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RACE-h Few-Shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MT-NLG 530B</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAMBADA Zero-Shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RACE-m Few-Shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RACE-h Few-Shot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Table Source: (Hoffman et al., 2022)]
Evaluations Tasks for Chinchilla

- Language Modelling
- MMLU
- Reading Comprehension
- BIG-bench
- Common Sense
- Closed Book QA
- Gender Bias and Toxicity
BIG-bench

Collection of ‘difficult’ tasks for current models (Srivastava et al., 2022)

Currently has 204 tasks and is growing with Github pull requests

(Hoffman et al., 2022) used 62 tasks
Example Data from BIG-bench

Which of the following sentences makes more sense?
choice: It started raining because the driver turned the wipers on.
choice: The driver turned the wipers on because it started raining.

[Figure Source: (Srivastava et al., 2022)]
Analysis Per Task

Chinchilla **outperforms** Gopher on **58 tasks**

Underperforms Gopher on **4 tasks**
Analysis Per Task

[Figure Source: (Hoffman et al., 2022)]
Evaluations Tasks for Chinchilla

- Language Modelling
- MMLU
- Reading Comprehension
- BIG-bench
- Common Sense
- Closed Book QA
- Gender Bias and Toxicity
Common Sense

Answer various common sense questions

E.g., reasoning about the physical world, pronoun resolution, emotion inferrance
Example Data from PIQA

To separate egg whites from the yolk using a water bottle, you should...

a. **Squeeze** the water bottle and press it against the yolk. **Release**, which creates suction and lifts the yolk.

b. **Place** the water bottle and press it against the yolk. **Keep pushing**, which creates suction and lifts the yolk.

[Figure Source: (Bisk et al., 2019)]
Example Data from SIQA

**Reasoning about Emotional Reactions**

In the school play, Robin played a hero in the struggle to the death with the angry villain.

**Q** How would others feel afterwards?

**A** (a) sorry for the villain (b) hopeful that Robin will succeed ✓ (c) like Robin should lose

[Figure Source: (Sap et al., 2019)]
# Chinchilla Outperforms Gopher

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

[Table Source: (Hoffman et al., 2022)]
Evaluations Tasks for Chinchilla

- Language Modelling
- MMLU
- Reading Comprehension
- BIG-bench
- Common Sense
- Closed Book QA
- Gender Bias and Toxicity
Closed Book QA

Answer short-answer questions without external sources

**Question:** what color was John Wilkes Booth’s hair

**Wikipedia Page:** John Wilkes Booth

**Long answer:** Some critics called Booth “the handsomest man in America” and a “natural genius”, and noted his having an “astonishing memory”; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair, and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a “muscular, perfect man” with “curling hair, like a Corinthian capital”.

**Short answer:** jet-black

[Source: (Kwiatkowski et al., 2019)]
## Chinchilla Outperforms Gopher

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

*Table Source: (Hoffman et al., 2022)*
Outline

1. Introduction

2. Initial Scaling Law (Kaplan et al., 2020)

3. Modified Scaling Law (Hoffman et al., 2022)

4. Chinchilla (Hoffman et al., 2022)

5. Beyond Scaling Law
Generalization of the Scaling Law

Other architecture — (Kaplan et al., 2020) tests the scaling law on LSTM and Universal Transformers (encoder-decoder model)

Other dataset — (Hoffman et al., 2022) tests the scaling law on different datasets (e.g., C4, Github)

Other domain — (Henighan et al., 2020) test the scaling law on different domains (e.g., image, video)
Generalization to LSTM

[Figure Source: (Kaplan et al., 2020)]
Generalization to Universal Transformers

[Figure Source: (Kaplan et al., 2020)]
Generalization to C4 and Github code

[Figure Source: (Hoffman et al., 2022)]
Generalization to C4 and Github code

<table>
<thead>
<tr>
<th>Approach</th>
<th>Coef. $a$ where $N_{opt} \propto C^a$</th>
<th>Coef. $b$ where $D_{opt} \propto C^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>GitHub</td>
<td>0.53</td>
<td>0.47</td>
</tr>
<tr>
<td>Kaplan et al. (2020)</td>
<td>0.73</td>
<td>0.27</td>
</tr>
</tbody>
</table>

[Figure Source: (Hoffman et al., 2022)]
Generalization to Image, Video, etc.

[Figure Source: (Henighan et al., 2020)]
Is Power-Law the best fit?

(Hoffman et al.) observe concavity in their model at high compute budgets.

The importance of dataset might increase for high compute budgets.

[Figure Source: (Hoffman et al., 2022)]
Scaling Law For Fine-Tuning (Tay et al., 2021)

Downstream performance after fine-tuning does not scale with model size.

Downstream performance does scale with depth, but not necessarily with dimension.
Downstream Performance Does Not Depend on N

[Figure Source: (Tay et al., 2021)]
Train Large, Then Compress (Li et al., 2020)

- Common Practice:
  - Train Small Model
  - Stop Training When Converged
  - Lightly Compress

- Optimal:
  - Train Large Model
  - Stop Training Early
  - Heavily Compress

[Figure Source: (Li et al., 2020)]
Deeper and Wider Models Converge in Fewer Steps

[Figure Source: Slides by (Li et al., 2020)]
Data Pruning ([Sorcher et al., 2022](#))

Develop a metric to measure the **quality of data**

**Prune the data** to include only high quality data

**Importance of dataset size decreases significantly**
The More Data We Prune, The Less Data Matters

[Figure Source: (Sorcher et al., 2022)]
Q3: (a) Do you think we can extend this study of LLMs to other types such as encoder-decoder models? Can you make your guess of the scaling law?

(b) These studies simply consider # of tokens as a proxy for training corpus. Do you think it is possible to take the quality/redundancy of the training data into account?