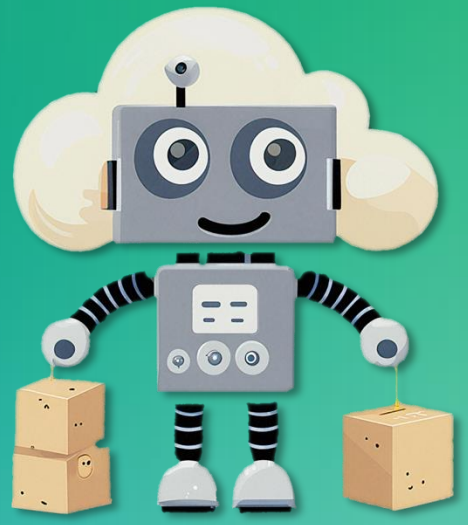
RPTU

Beyond Bigger

The Evolution of Scaling Laws in Large Language Models

Mariano Kamp

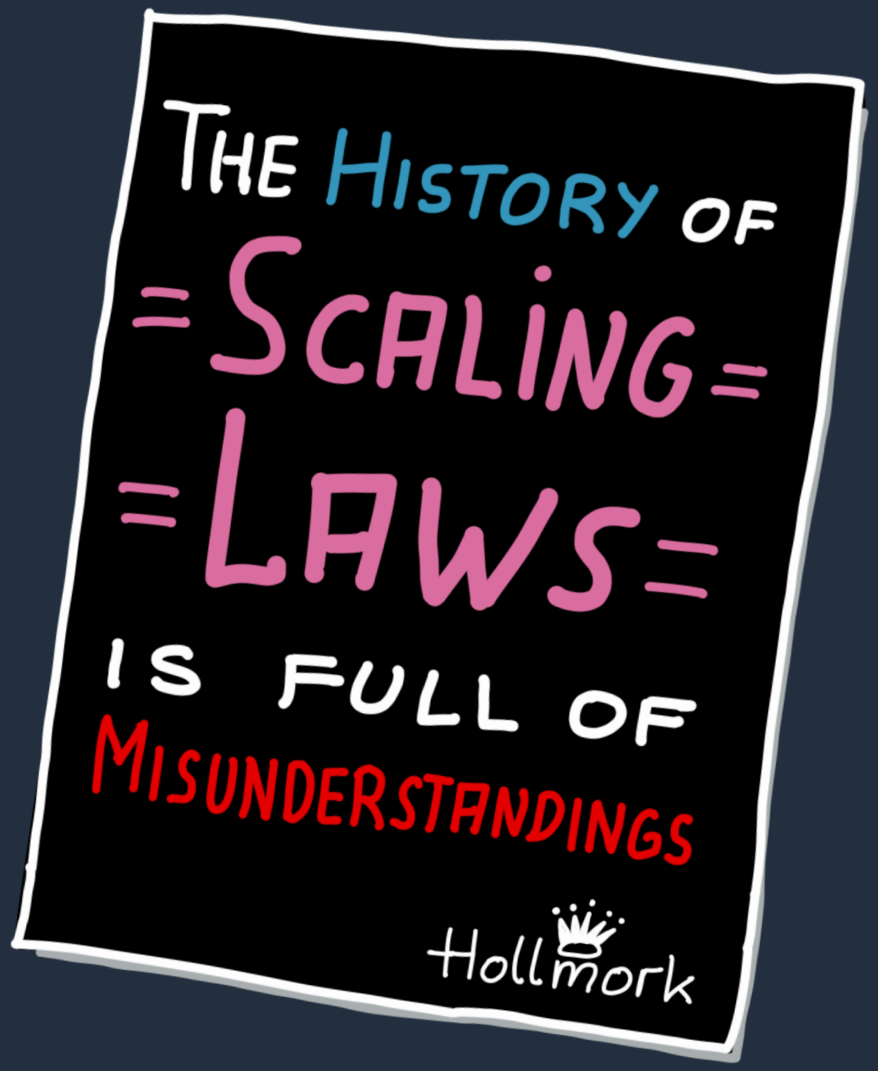
Principal Solutions Architect
Amazon Web Services



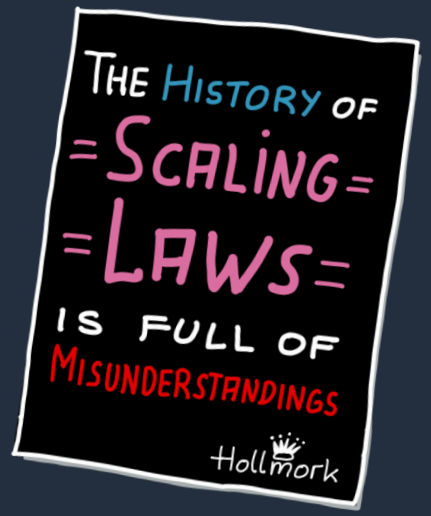
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

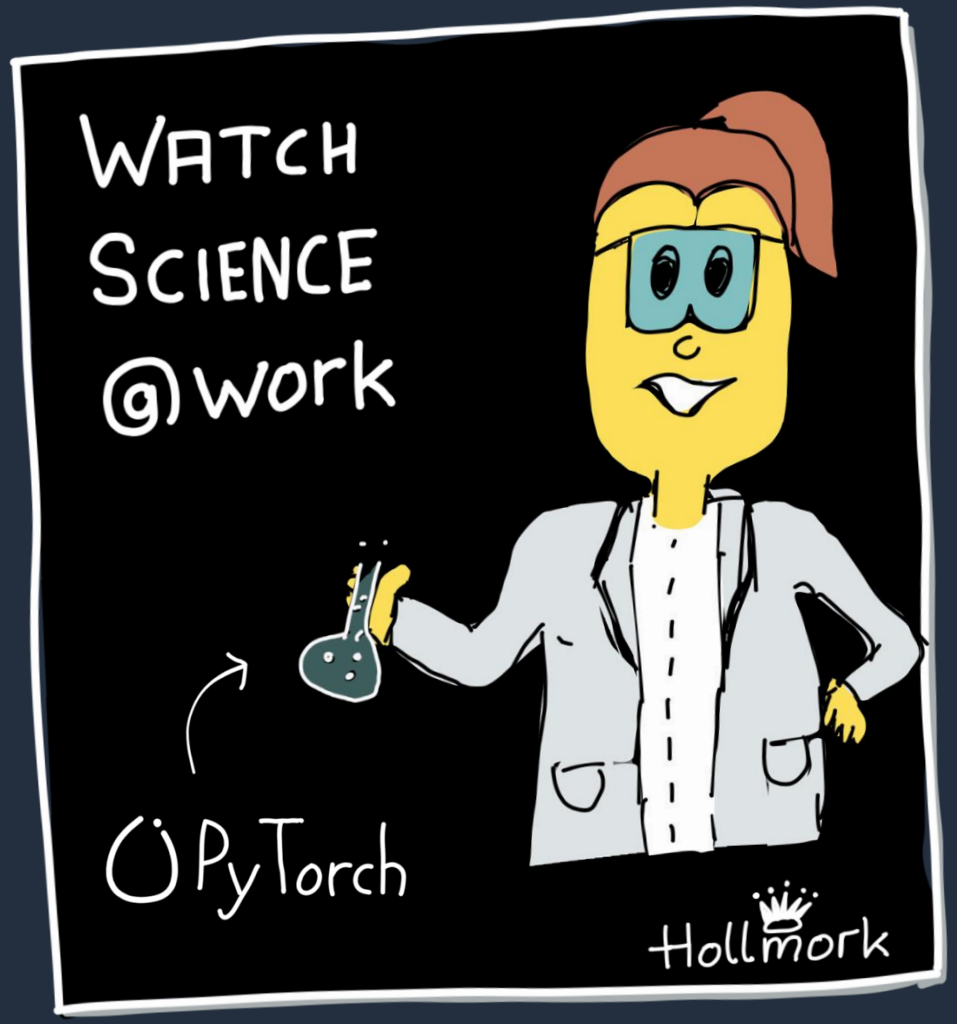
Scaling Laws

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

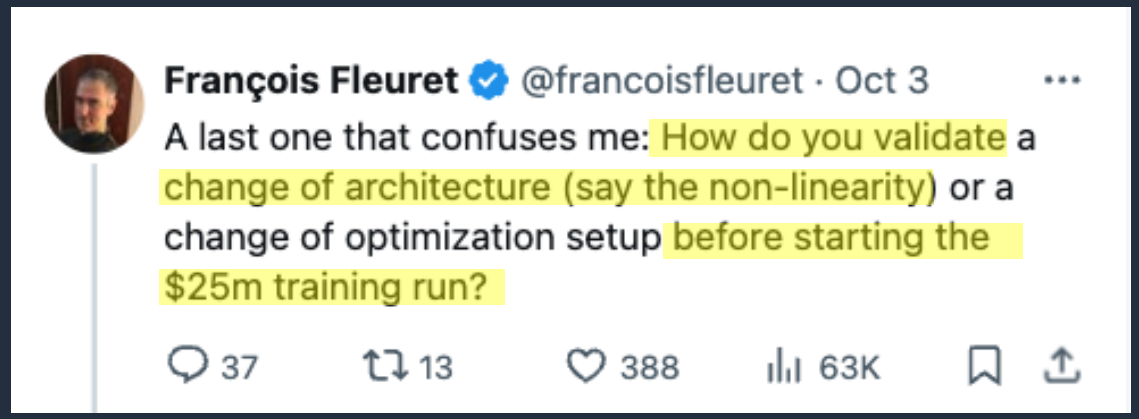


© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



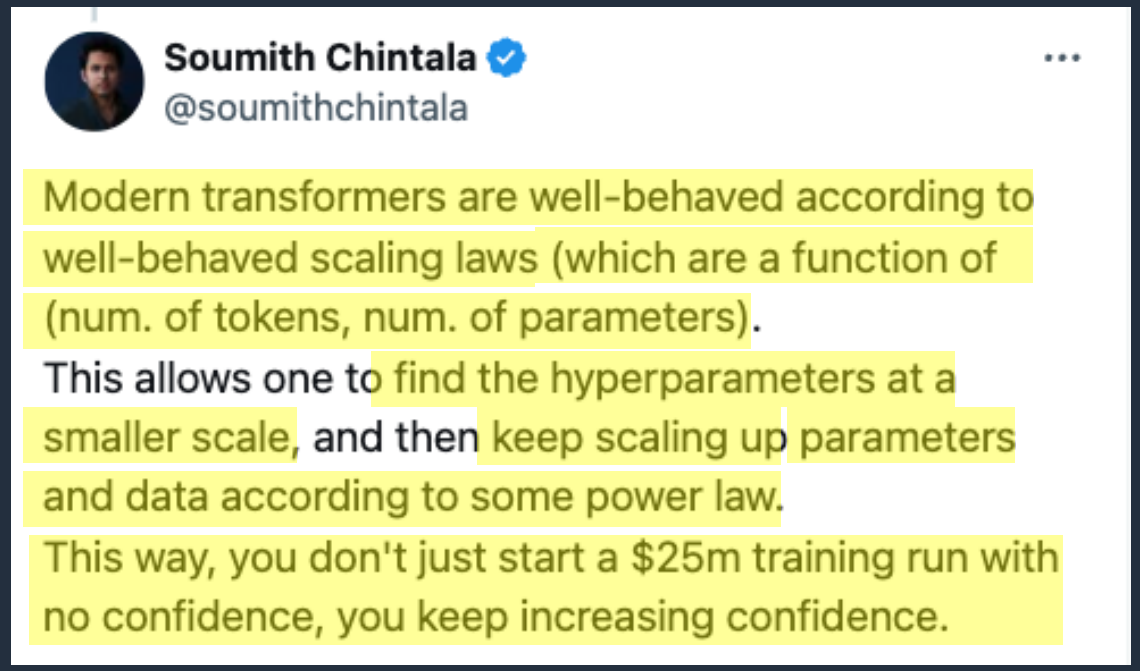


© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Soumith Chintala
@soumithchintala

Modern transformers are well-behaved according to well-behaved scaling laws (which are a function of (num. of tokens, num. of parameters).
This allows one to find the hyperparameters at a smaller scale, and then keep scaling up parameters and data according to some power law.
This way, you don't just start a $25m training run with no confidence, you keep increasing confidence.

But sometimes things do not work out, either because

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Soumith Chintala
@soumithchintala

Modern transformers are well-behaved according to well-behaved scaling laws (which are a function of (num. of tokens, num. of parameters).
This allows one to find the hyperparameters at a smaller scale, and then keep scaling up parameters and data according to some power law.
This way, you don't just start a $25m training run with no confidence, you keep increasing confidence.

But sometimes things do not work out, either because you were in a hurry and didn't scale up carefully -- or because the scaling law broke at some scale.
(like the rumored OpenAI model, or the rumored Gemini Ultra model)

10:00 PM · Oct 3, 2024 · **12.1K** Views

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

DETOUR

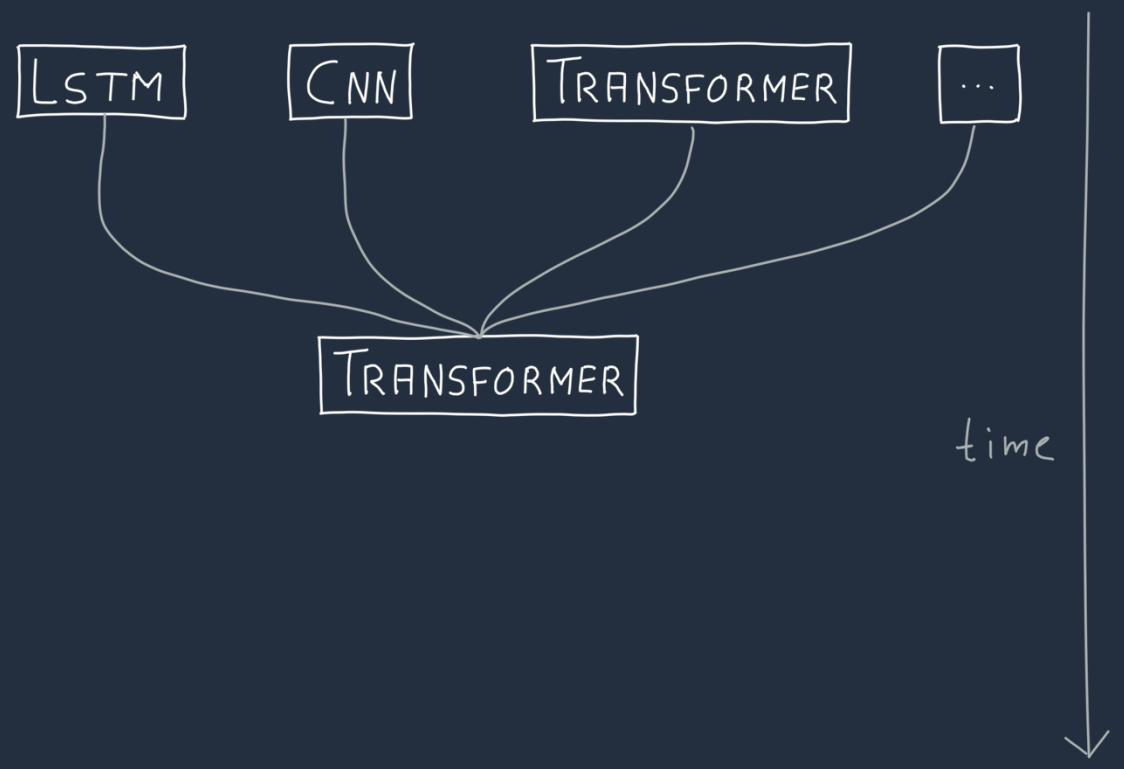
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ONE NN-ARCHITECTURE TO RULE THEM ALL?

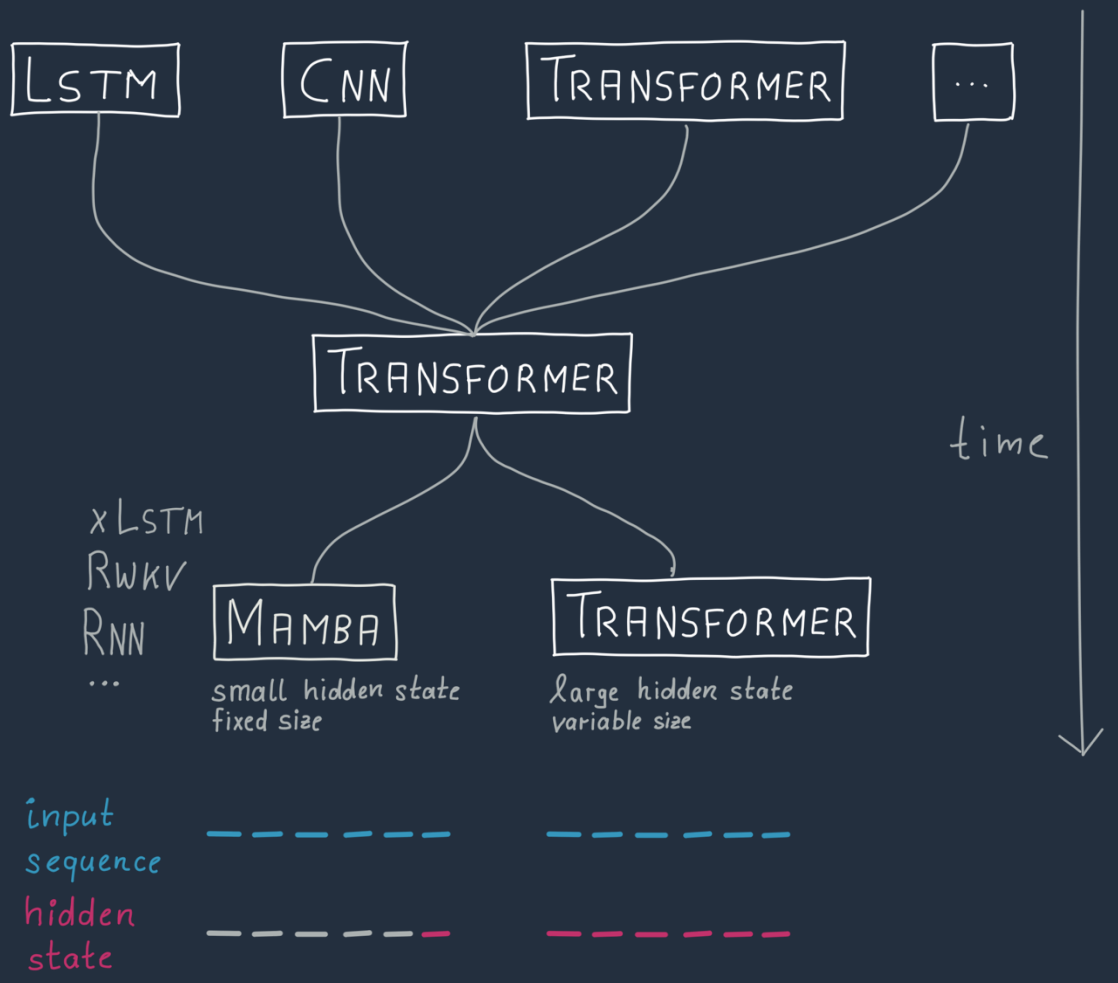
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

LSTM CNN TRANSFORMER ...

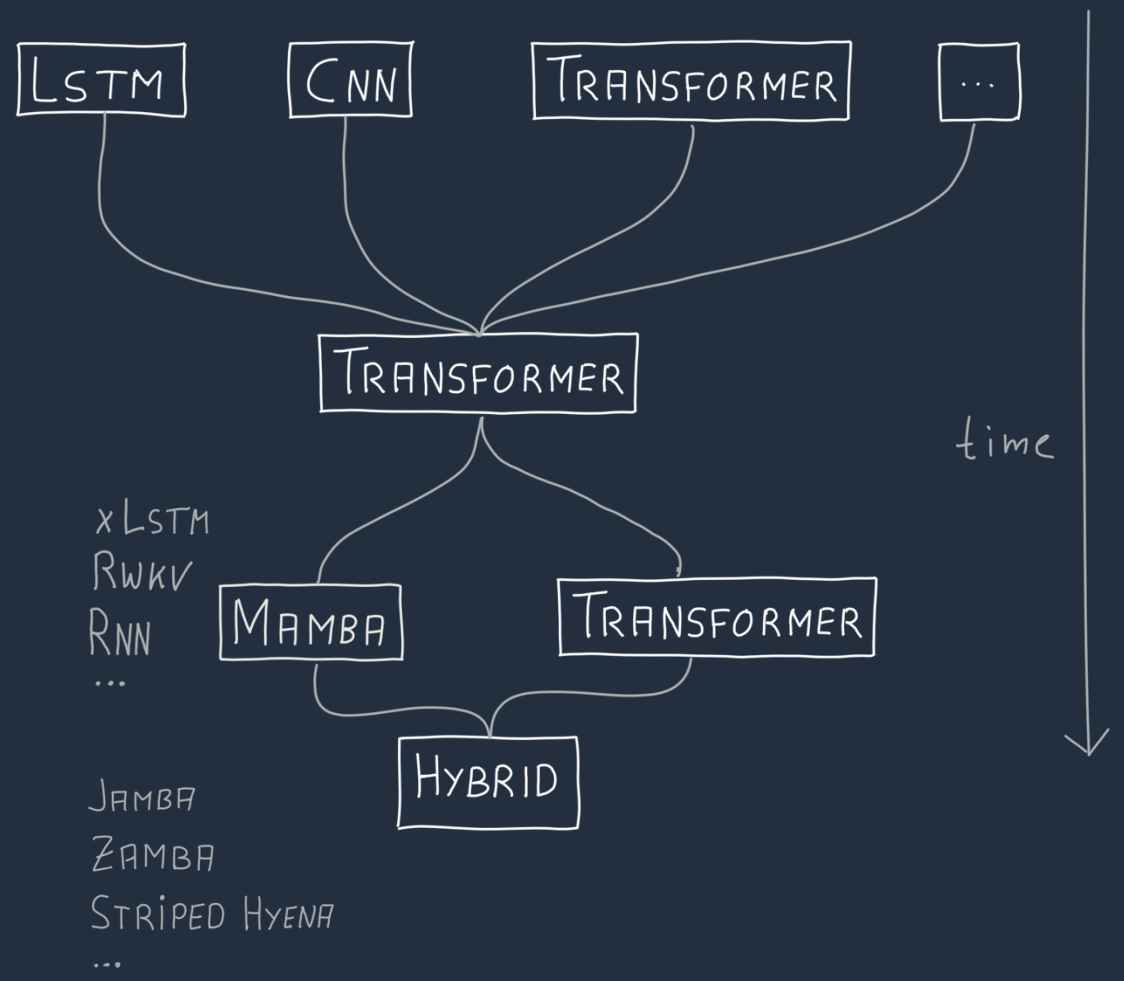
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Hybrid models begin to gain traction

Models that combine attention and other mechanisms are able to maintain or even improve accuracy, while reducing computational costs and memory footprint.

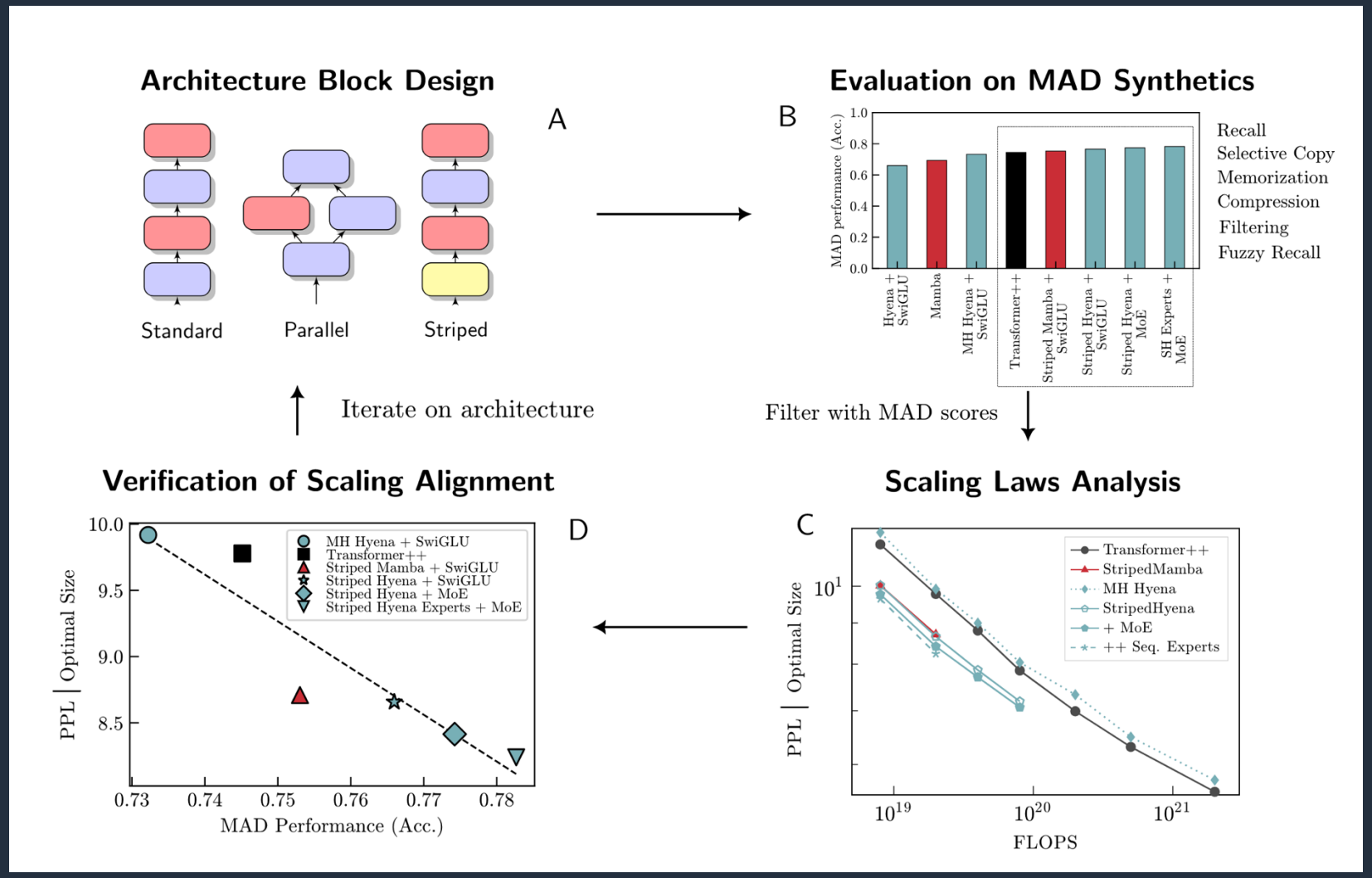
- Selective state-space models like Mamba, designed last year to handle long sequences more efficiently, can to some extent compete with transformers, but lag on tasks that require copying or in-context learning. That said, Falcon's Mamba 7B shows impressive benchmark performance versus similar-sized transformer models.
- Hybrid models appear to be a more promising direction. Combined with self-attention and MLP layers, the AI21's Mamba-Transformer hybrid model outperforms the 8B Transformer across knowledge and reasoning benchmarks, while being up to 8x faster generating tokens in inference.
- In a nostalgia trip, there are early signs of a comeback for recurrent neural networks, which had fallen out of fashion due to training and scaling difficulties.
- Griffin, trained by Google DeepMind, mixes linear recurrences and local attention, holding its own against Llama-2 while being trained on 6x fewer tokens.

	Transformer	Mamba	Jamba
Highest Quality Output	●		●
High Throughput		●	●
Low Memory Footprint		●	●

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Mechanistic Design and Scaling of Hybrid Architectures

Poli, Thomas, Nguyen et al., 2024

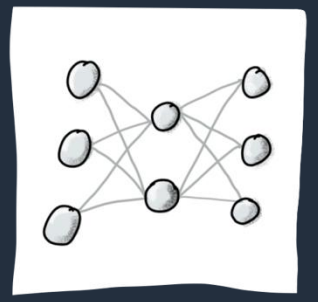


© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

WHAT TO INVEST IN?

COMPUTE
FLOPS
COST($)

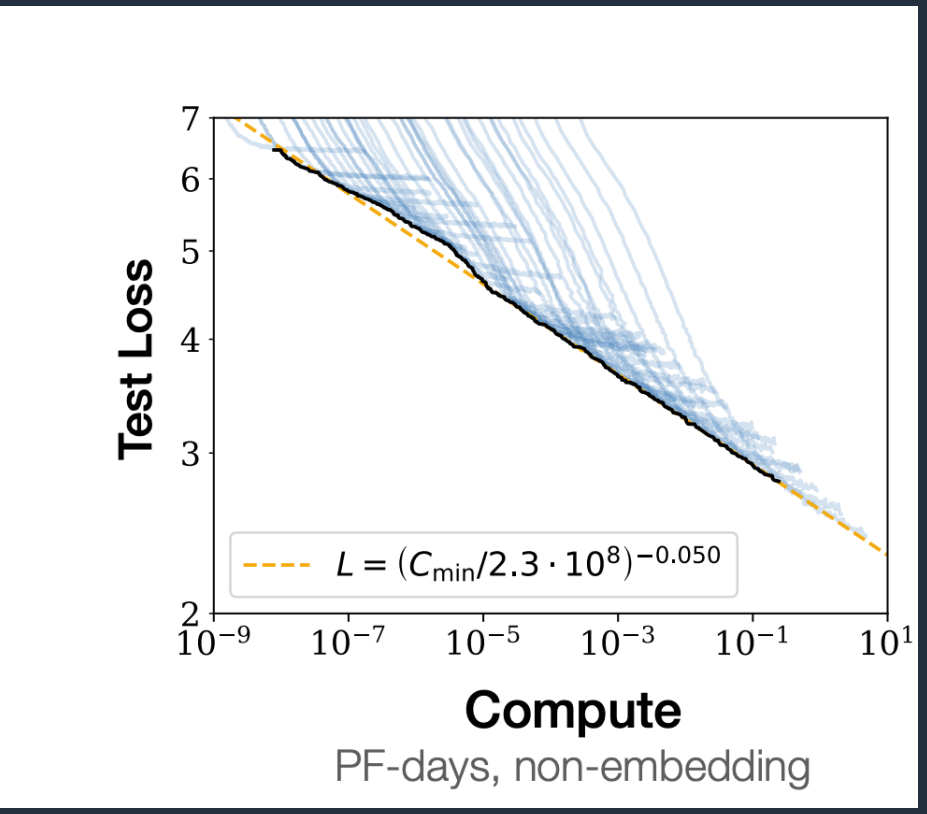
DATA
TOKENS



MODEL
PARAMETERS

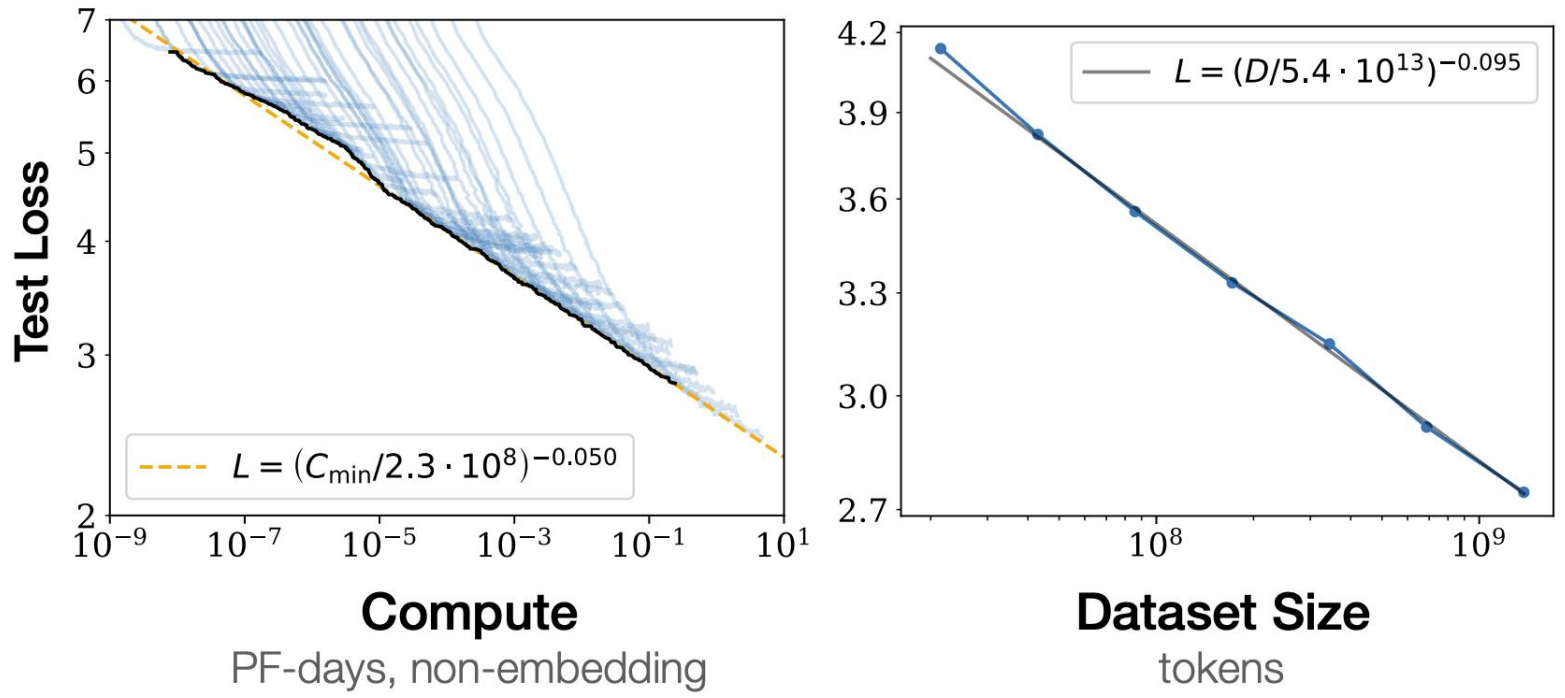
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scaling Laws for Neural Language Models
Kaplan et al., 2020



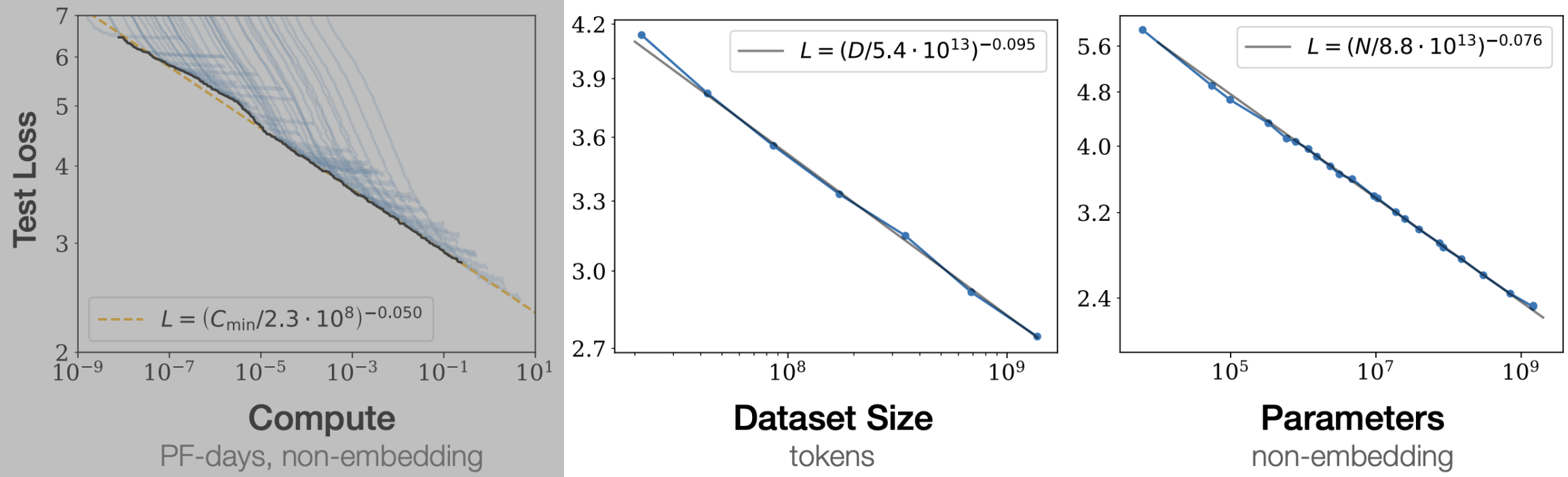
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scaling Laws for Neural Language Models
Kaplan et al., 2020



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

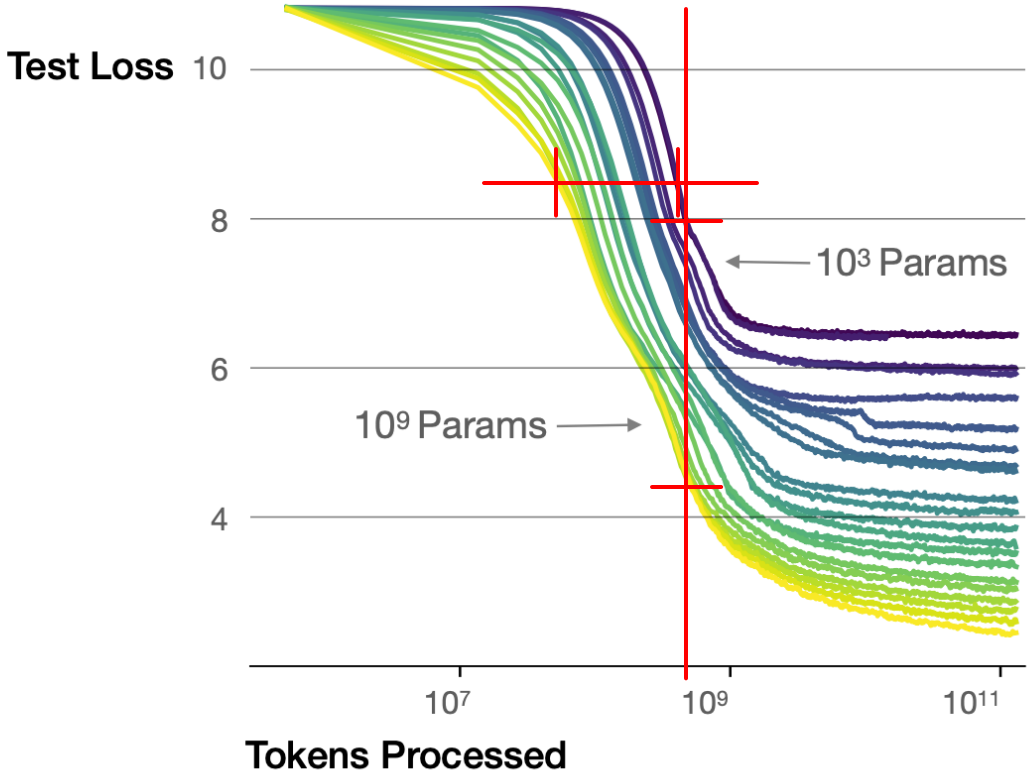
Scaling Laws for Neural Language Models
Kaplan et al., 2020



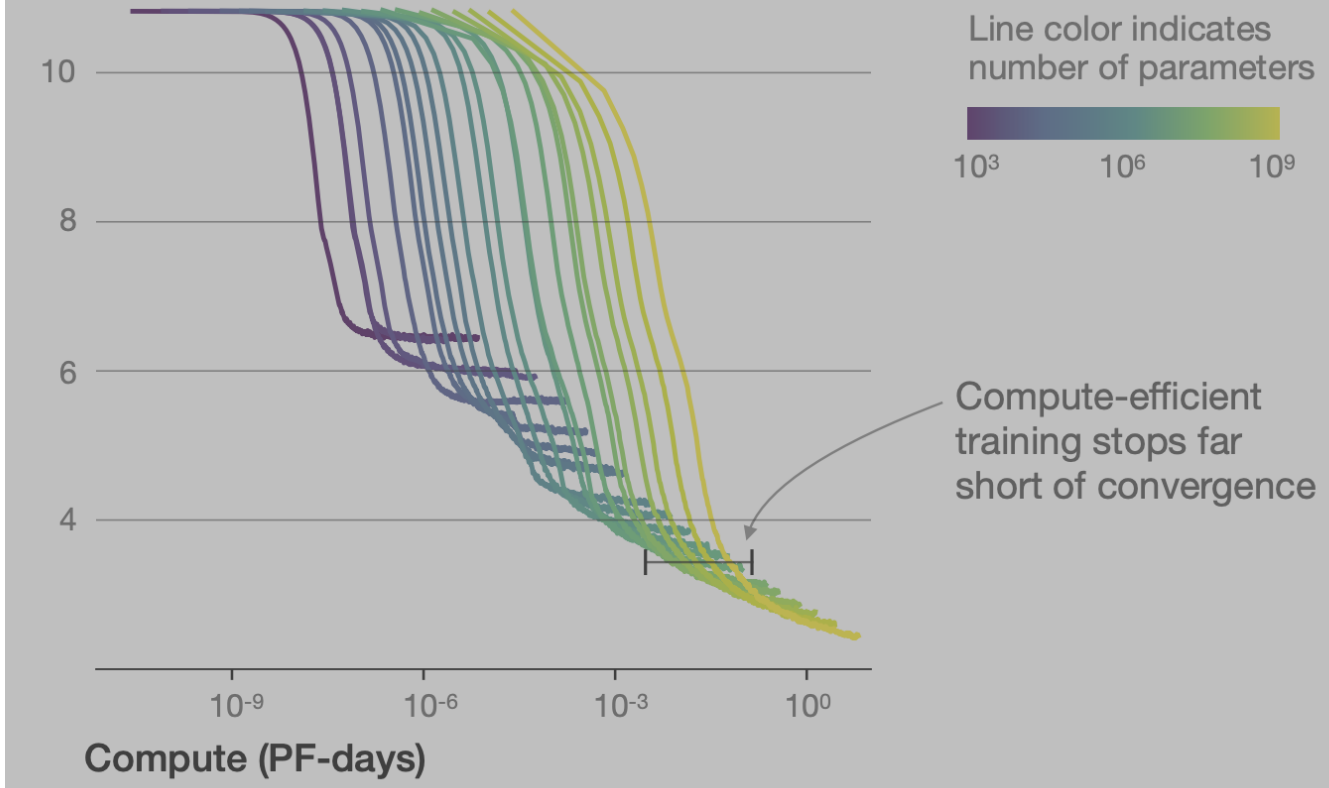
© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scaling Laws for Neural Language Models
Kaplan et al., 2020

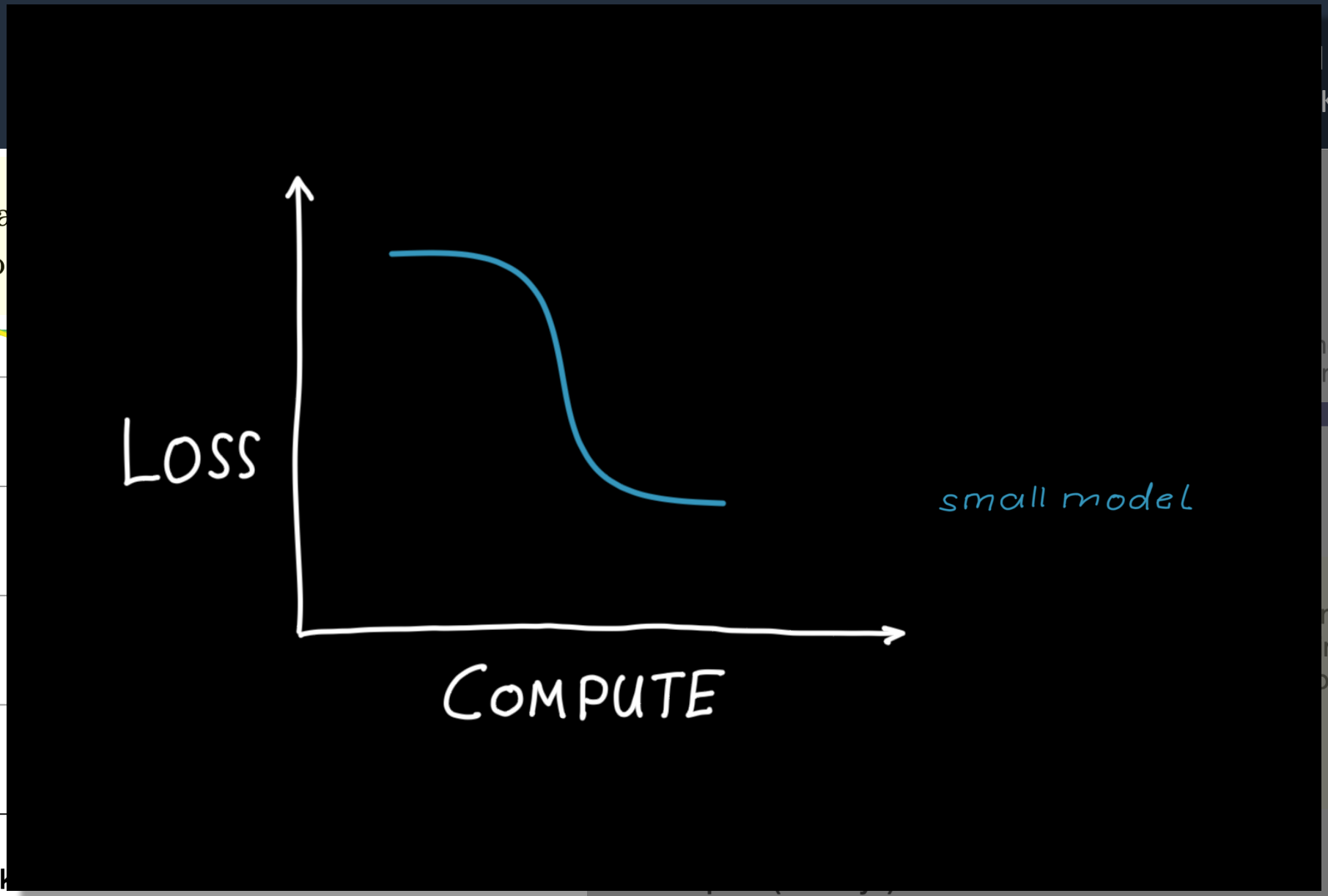
Larger models require **fewer samples** to reach the same performance



The optimal model size grows smoothly with the loss target and compute budget



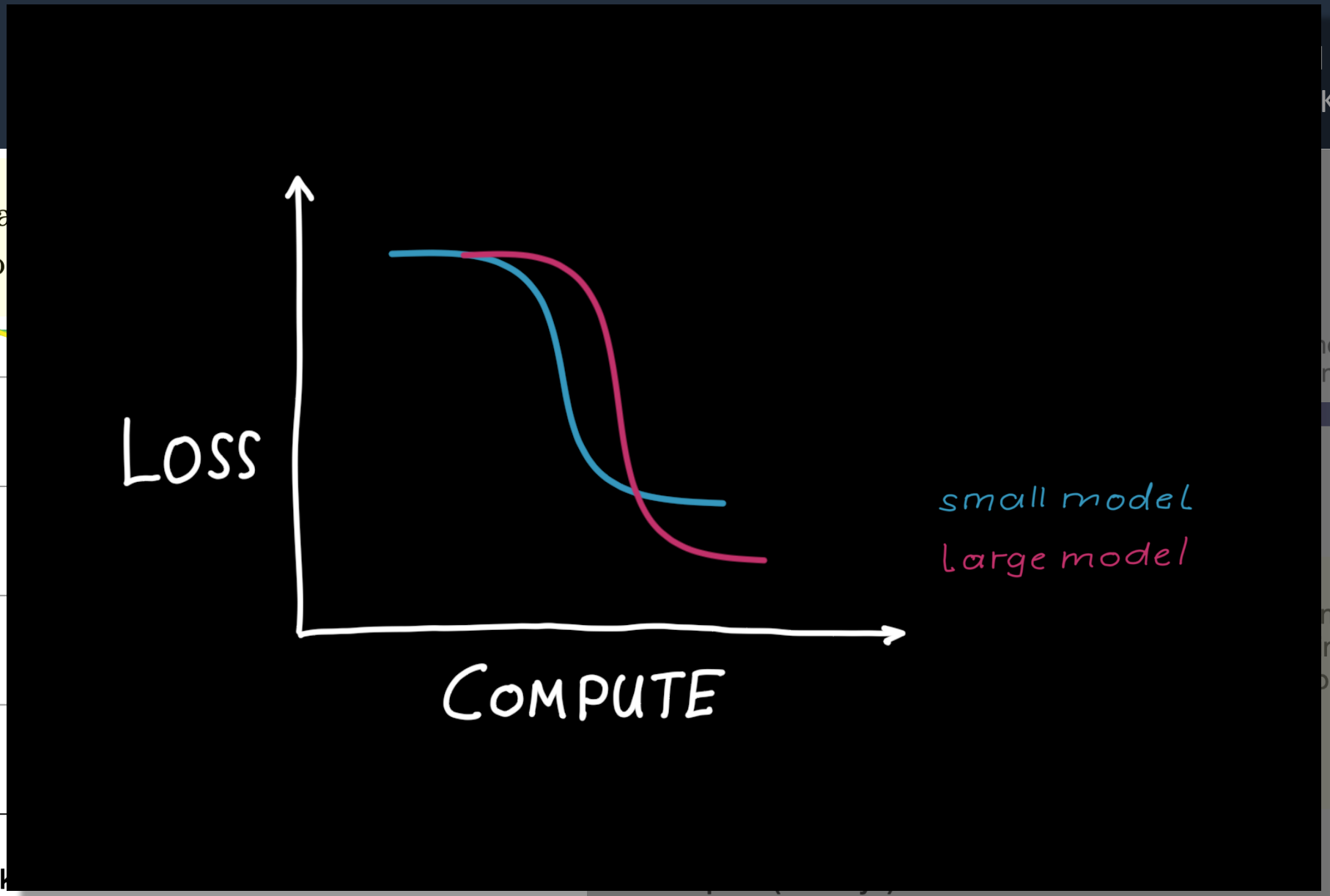
Language Models
Kaplan et al., 2020



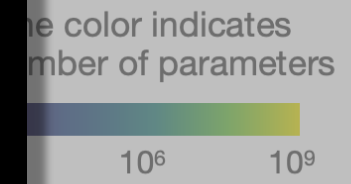
e color indicates
mber of parameters

10^6 10^9

mpute-efficient
ning stops far
ort of convergence

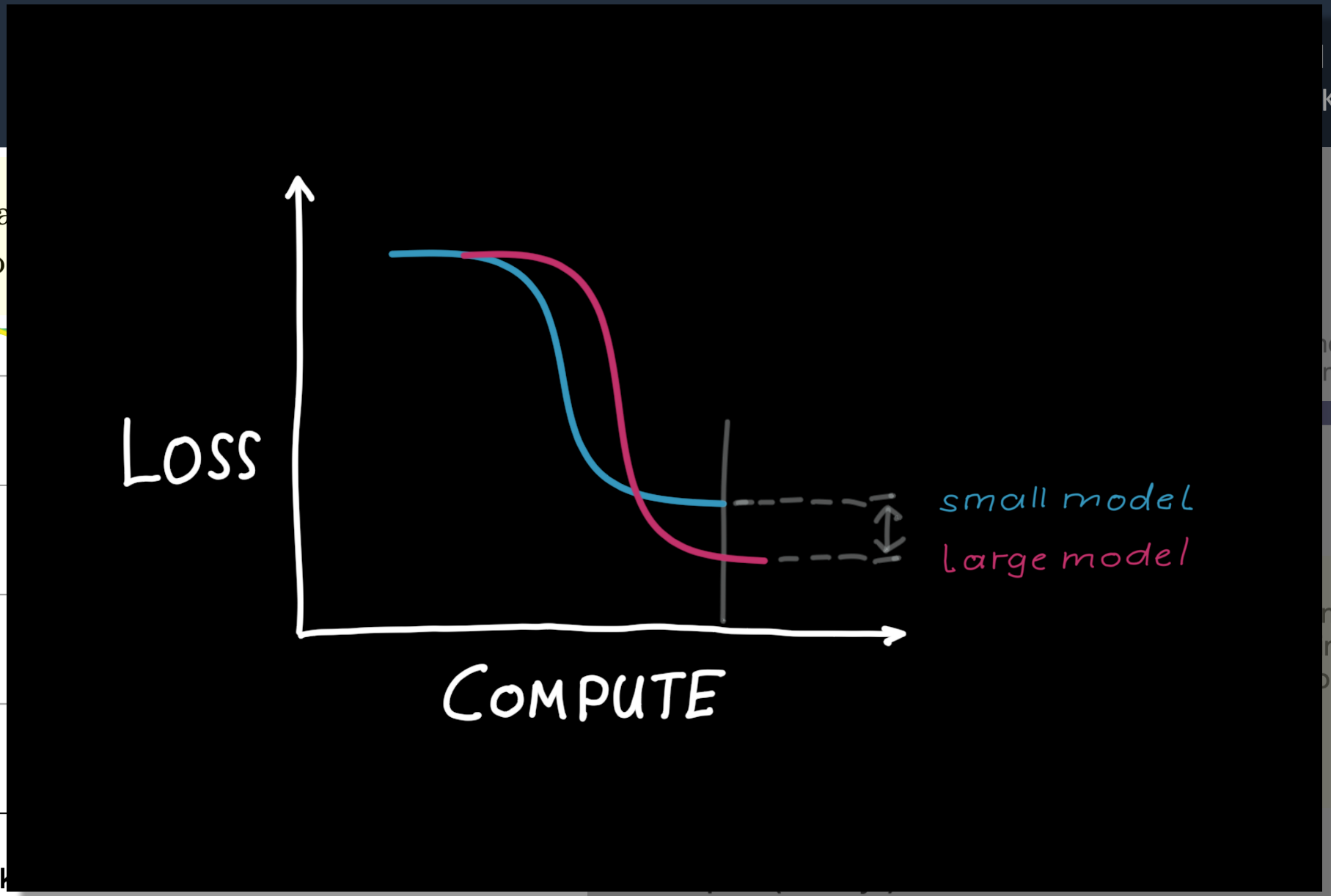


Language Models
Kaplan et al., 2020



e color indicates
mber of parameters

mpute-efficient
ning stops far
ort of convergence



Language Models
Kaplan et al., 2020

Test Loss 10
8
6
4

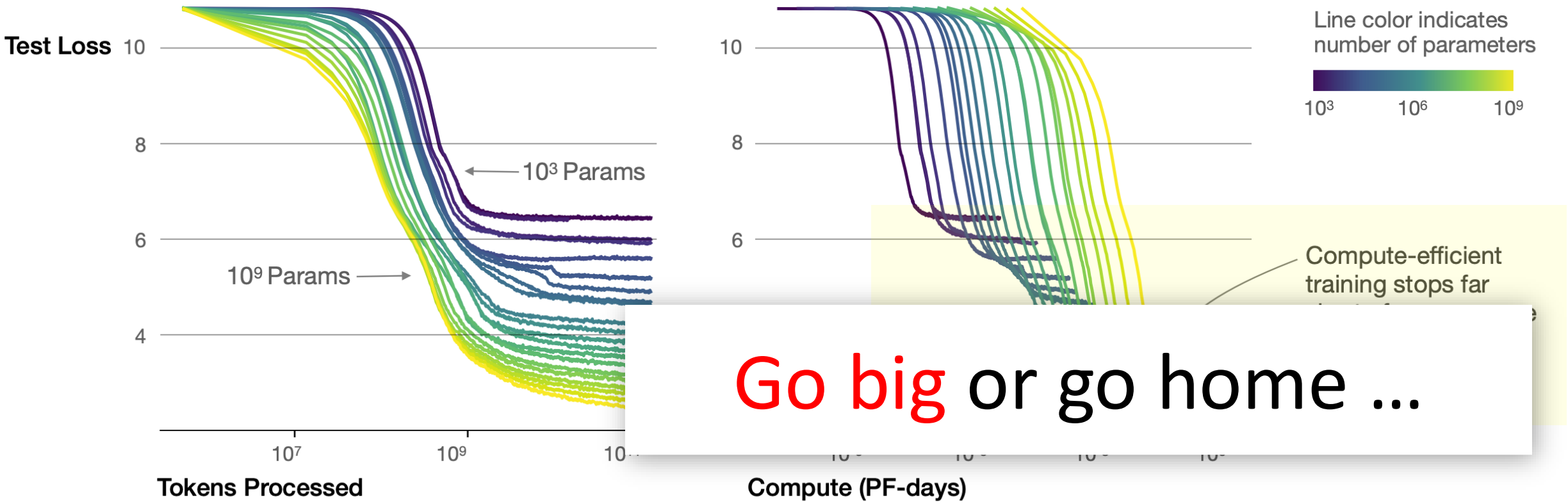
e color indicates
mber of parameters
10^6 10^9

mpute-efficient
ning stops far
ort of convergence

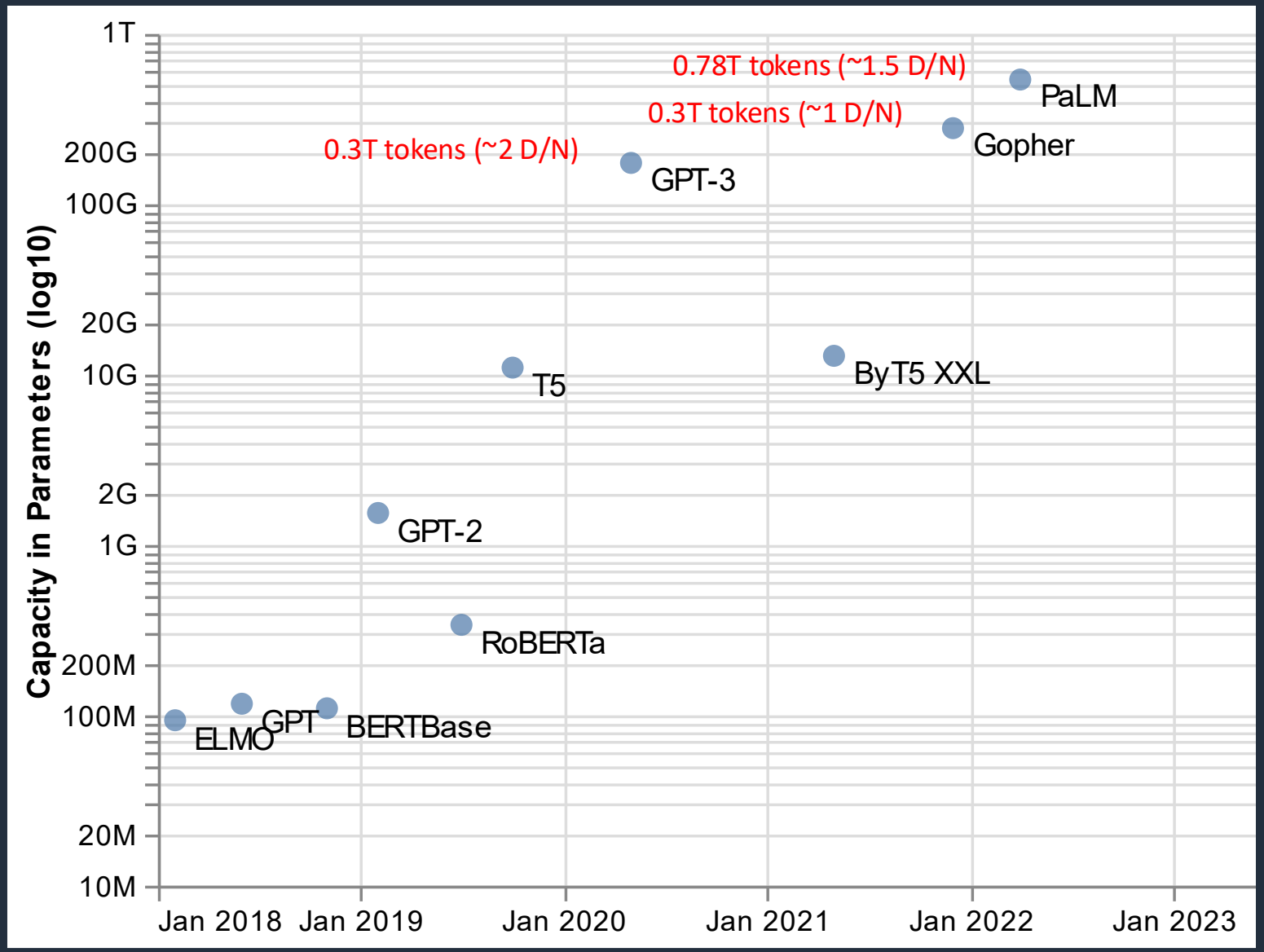
Scaling Laws for Neural Language Models
Kaplan et al., 2020

Larger models require **fewer samples** to reach the same performance

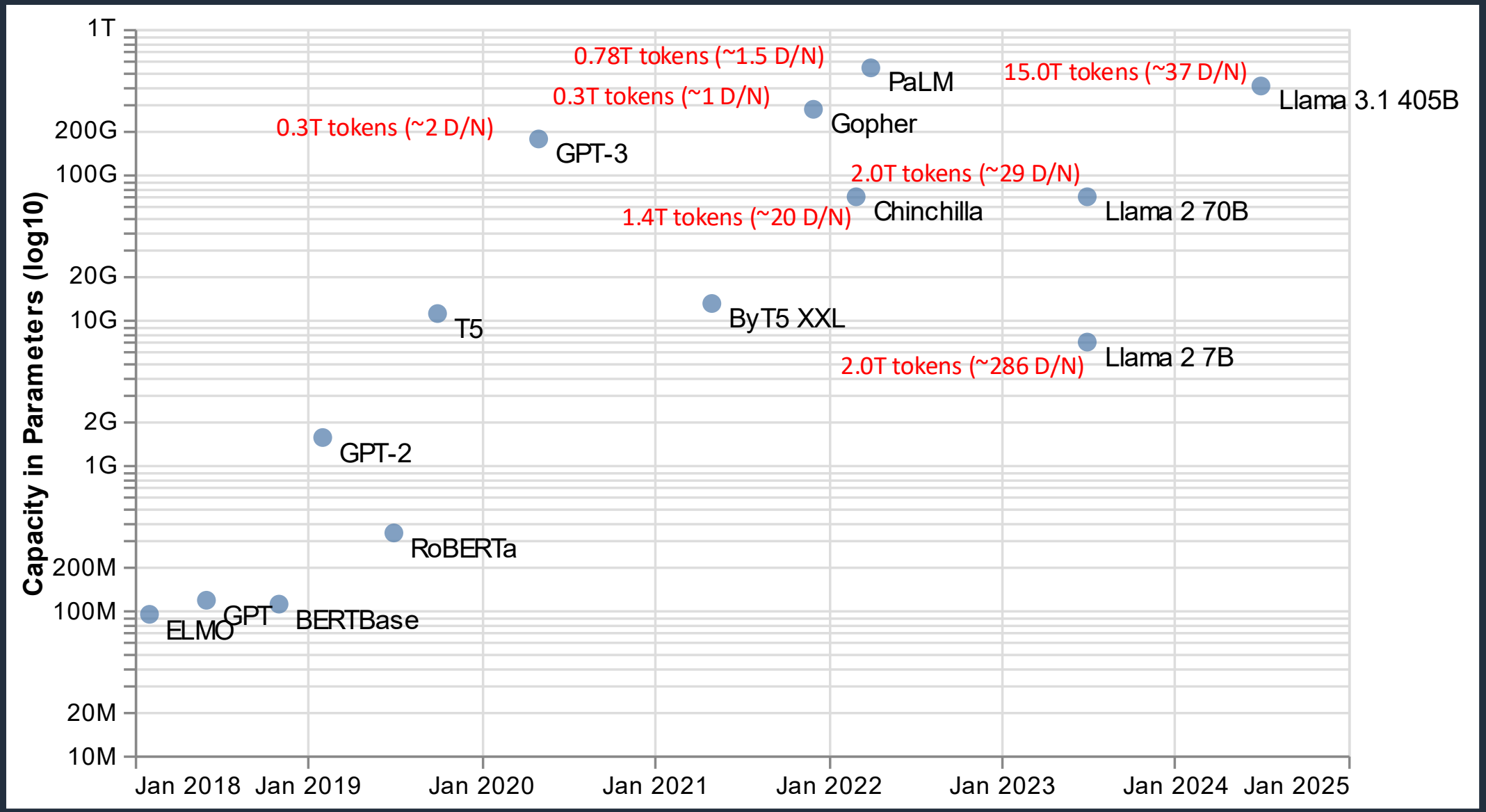
The optimal model size grows smoothly with the loss target and compute budget



Go big or go home …

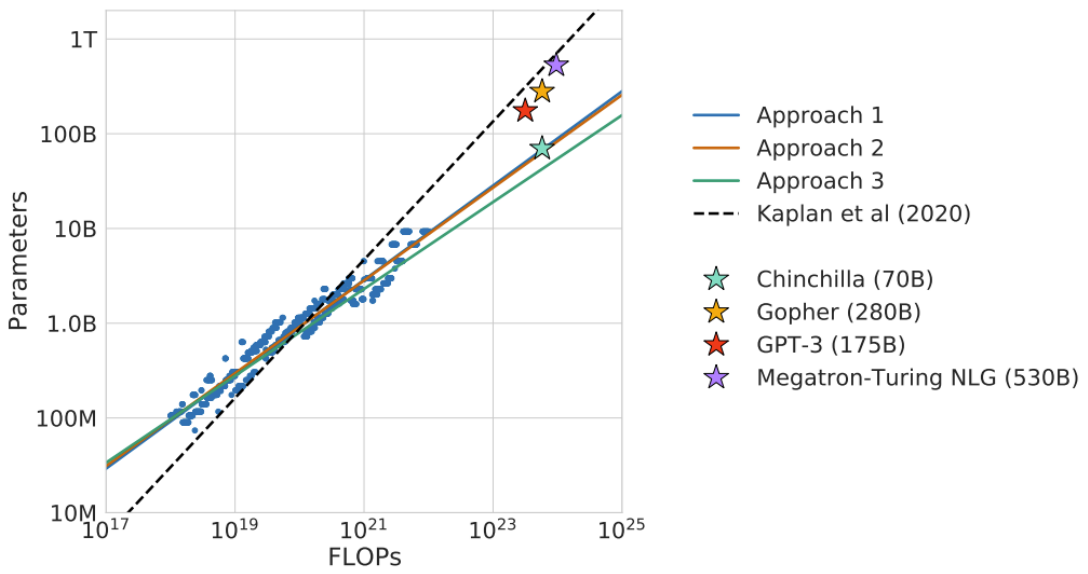


© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Training Compute-Optimal Large Language Models
Hoffmann et al., 2022

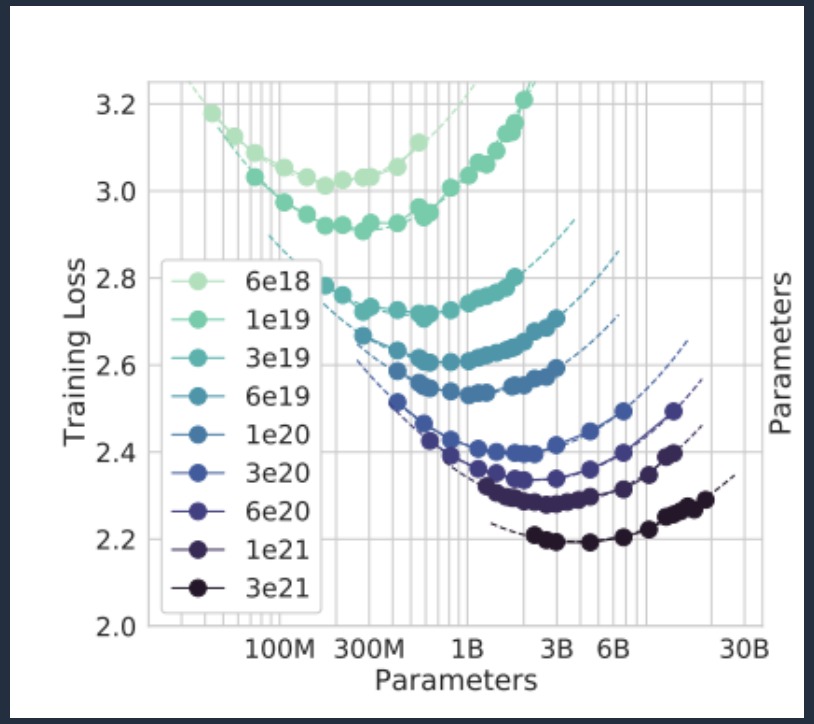


Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270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

GPT-3 5-shot	43.9%
Gopher 5-shot	60.0%
***Chinchilla* 5-shot**	**67.6%**

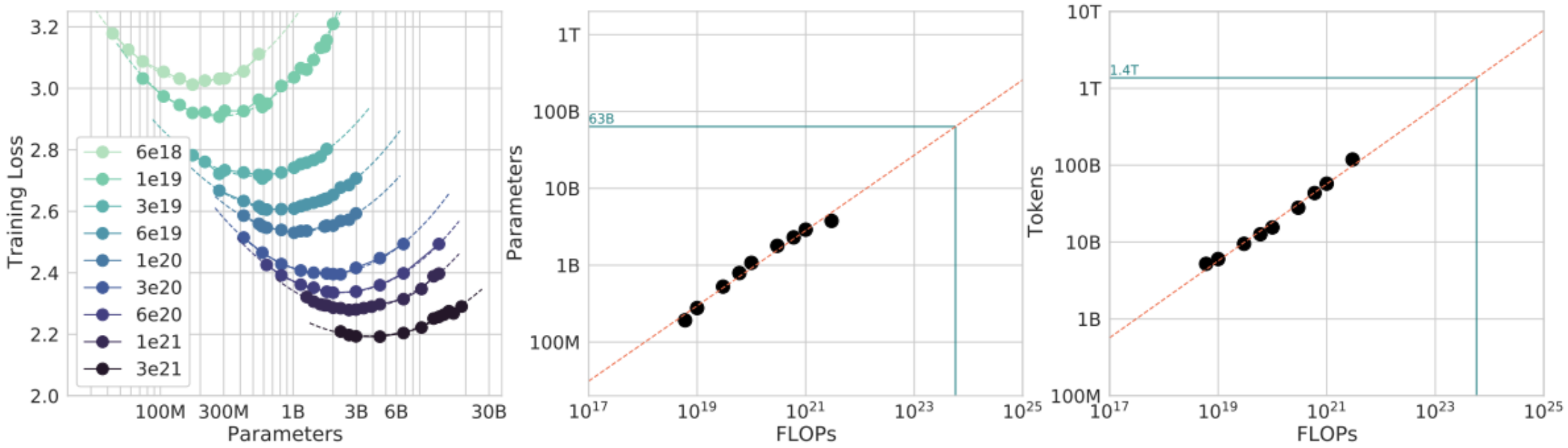
Approach	Coeff. a where $N_{opt} \propto C^a$	Coeff. b where $D_{opt} \propto C^b$
1. Minimum over training curves	0.50 (0.488, 0.502)	0.50 (0.501, 0.512)
2. IsoFLOP profiles	0.49 (0.462, 0.534)	0.51 (0.483, 0.529)
3. Parametric modelling of the loss	0.46 (0.454, 0.455)	0.54 (0.542, 0.543)
Kaplan et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

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



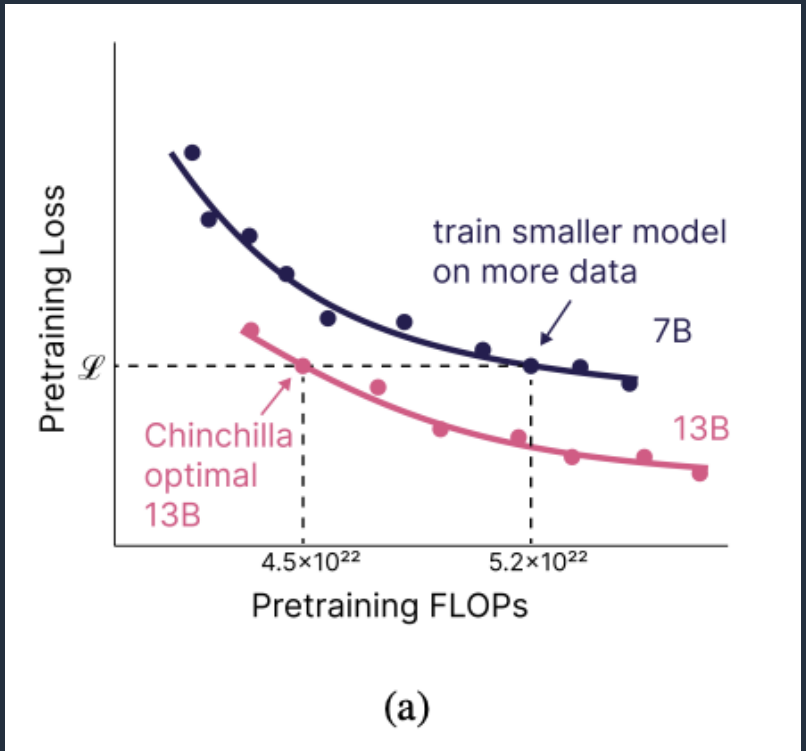
Training Compute-Optimal Large Language Models
Hoffmann et al., 2022

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



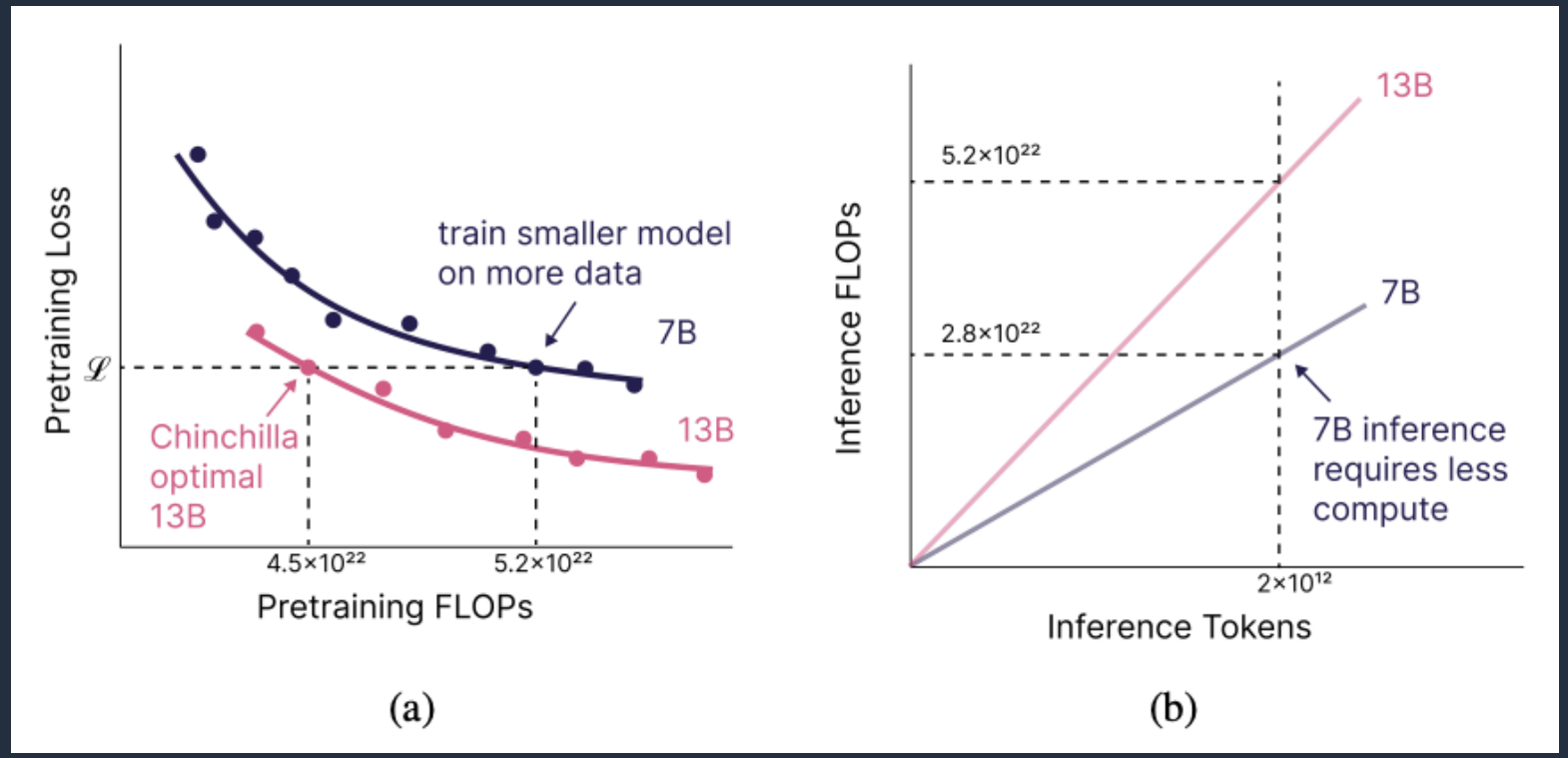
Training Compute-Optimal Large Language Models
Hoffmann et al., 2022

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



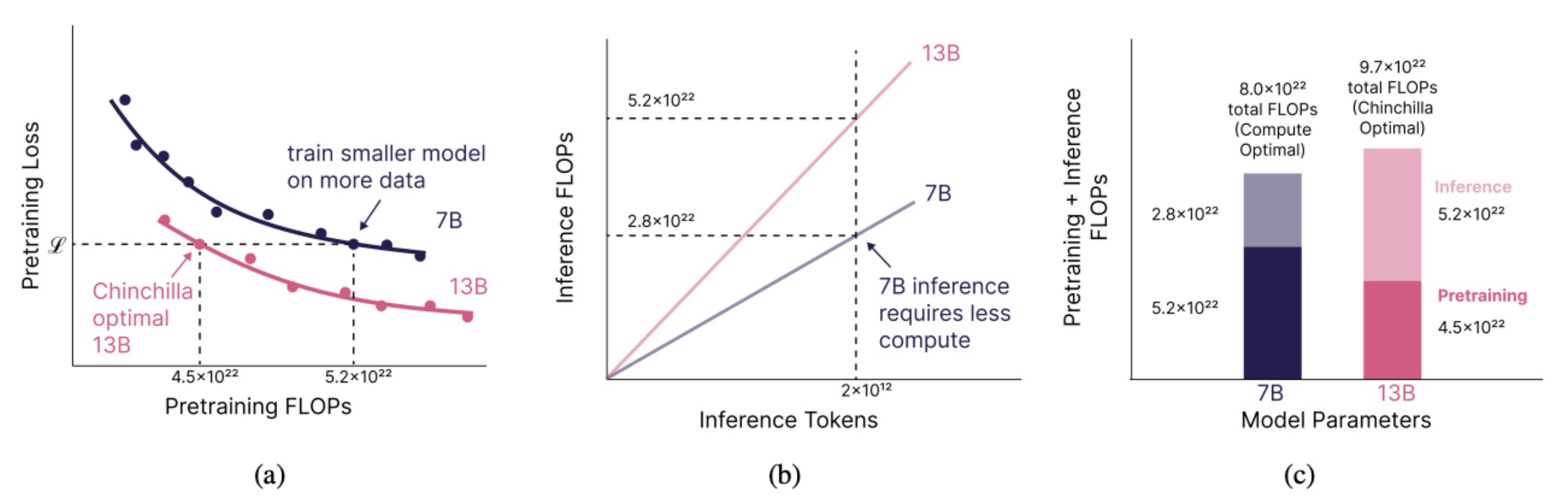
Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws
Sardana et al., 2023

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws
Sardana et al., 2023

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws
Sardana et al., 2023

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws
Sardana et al., 2023

$$N^*(\ell, D_{\text{inp}}, D_{\text{out}}), D^*_{\text{tr}}(\ell, D_{\text{inp}}, D_{\text{out}}) =$$

$$\underset{N, D_{\text{tr}} \mid L(N, D_{\text{tr}}) = \ell}{\arg\min} \frac{6 N D_{\text{tr}} C_{\text{tr}}}{\text{MFU}_{\text{tr}}} + 2 N C_{\text{inf}} \left[\frac{D_{\text{inp}}}{\text{MFU}_{\text{inp}}} + \frac{D_{\text{out}}}{\text{MFU}_{\text{out}}} \right]$$

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws
Sardana et al., 2023

$$N^*(\ell, D_{\text{inp}}, D_{\text{out}}), D^*_{\text{tr}}(\ell, D_{\text{inp}}, D_{\text{out}}) = \underset{N, D_{\text{tr}} \mid L(N, D_{\text{tr}}) = \ell}{\arg\min} \frac{6 N D_{\text{tr}} C_{\text{tr}}}{\text{MFU}_{\text{tr}}} + 2 N C_{\text{inf}} \left[\frac{D_{\text{inp}}}{\text{MFU}_{\text{inp}}} + \frac{D_{\text{out}}}{\text{MFU}_{\text{out}}} \right]$$

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws
Sardana et al., 2023

$$N^*(\ell, D_{\text{inp}}, D_{\text{out}}), D^*_{\text{tr}}(\ell, D_{\text{inp}}, D_{\text{out}}) =$$

$$\underset{N, D_{\text{tr}} \mid L(N, D_{\text{tr}}) = \ell}{\arg\min} \frac{6 N D_{\text{tr}} C_{\text{tr}}}{\text{MFU}_{\text{tr}}} + 2 N C_{\text{inf}} \left[\frac{D_{\text{inp}}}{\text{MFU}_{\text{inp}}} + \frac{D_{\text{out}}}{\text{MFU}_{\text{out}}} \right]$$

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws
Sardana et al., 2023

$$N^*(\ell, D_{\text{inp}}, D_{\text{out}}), D^*_{\text{tr}}(\ell, D_{\text{inp}}, D_{\text{out}}) =$$

$$\underset{N, D_{\text{tr}} \mid L(N, D_{\text{tr}}) = \ell}{\arg\min} \frac{6 N D_{\text{tr}} C_{\text{tr}}}{\text{MFU}_{\text{tr}}} + 2 N C_{\text{inf}} \left[\frac{D_{\text{inp}}}{\text{MFU}_{\text{inp}}} + \frac{D_{\text{out}}}{\text{MFU}_{\text{out}}} \right]$$

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws
Sardana et al., 2023

$$N^*(\ell, D_{\text{inp}}, D_{\text{out}}), D^*_{\text{tr}}(\ell, D_{\text{inp}}, D_{\text{out}}) =$$

$$\underset{N, D_{\text{tr}} \mid L(N, D_{\text{tr}}) = \ell}{\arg\min} \frac{6 N D_{\text{tr}} C_{\text{tr}}}{\text{MFU}_{\text{tr}}} + 2 N C_{\text{inf}} \left[\frac{D_{\text{inp}}}{\text{MFU}_{\text{inp}}} + \frac{D_{\text{out}}}{\text{MFU}_{\text{out}}} \right]$$

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws
Sardana et al., 2023

$$N^*(\ell, D_{\text{inp}}, D_{\text{out}}), D^*_{\text{tr}}(\ell, D_{\text{inp}}, D_{\text{out}}) =$$

$$\underset{N, D_{\text{tr}} \mid L(N, D_{\text{tr}}) = \ell}{\arg\min} \frac{6 N D_{\text{tr}} C_{\text{tr}}}{\text{MFU}_{\text{tr}}} + 2 N C_{\text{inf}} \left[\frac{D_{\text{inp}}}{\text{MFU}_{\text{inp}}} + \frac{D_{\text{out}}}{\text{MFU}_{\text{out}}} \right]$$

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws
Sardana et al., 2023

$$N^*(\ell, D_{\text{inp}}, D_{\text{out}}), D^*_{\text{tr}}(\ell, D_{\text{inp}}, D_{\text{out}}) =$$

$$\underset{N, D_{\text{tr}} \mid L(N, D_{\text{tr}}) = \ell}{\arg\min} \frac{6 N D_{\text{tr}} C_{\text{tr}}}{\text{MFU}_{\text{tr}}} + 2 N C_{\text{inf}} \left[\frac{D_{\text{inp}}}{\text{MFU}_{\text{inp}}} + \frac{D_{\text{out}}}{\text{MFU}_{\text{out}}} \right]$$

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws
Sardana et al., 2023

$$N^*(\ell, D_{\text{inp}}, D_{\text{out}}), D^*_{\text{tr}}(\ell, D_{\text{inp}}, D_{\text{out}}) =$$

$$\underset{N, D_{\text{tr}} \mid L(N, D_{\text{tr}}) = \ell}{\arg\min} \frac{6 N D_{\text{tr}} C_{\text{tr}}}{\text{MFU}_{\text{tr}}} + 2 N C_{\text{inf}} \left[\frac{D_{\text{inp}}}{\text{MFU}_{\text{inp}}} + \frac{D_{\text{out}}}{\text{MFU}_{\text{out}}} \right]$$

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

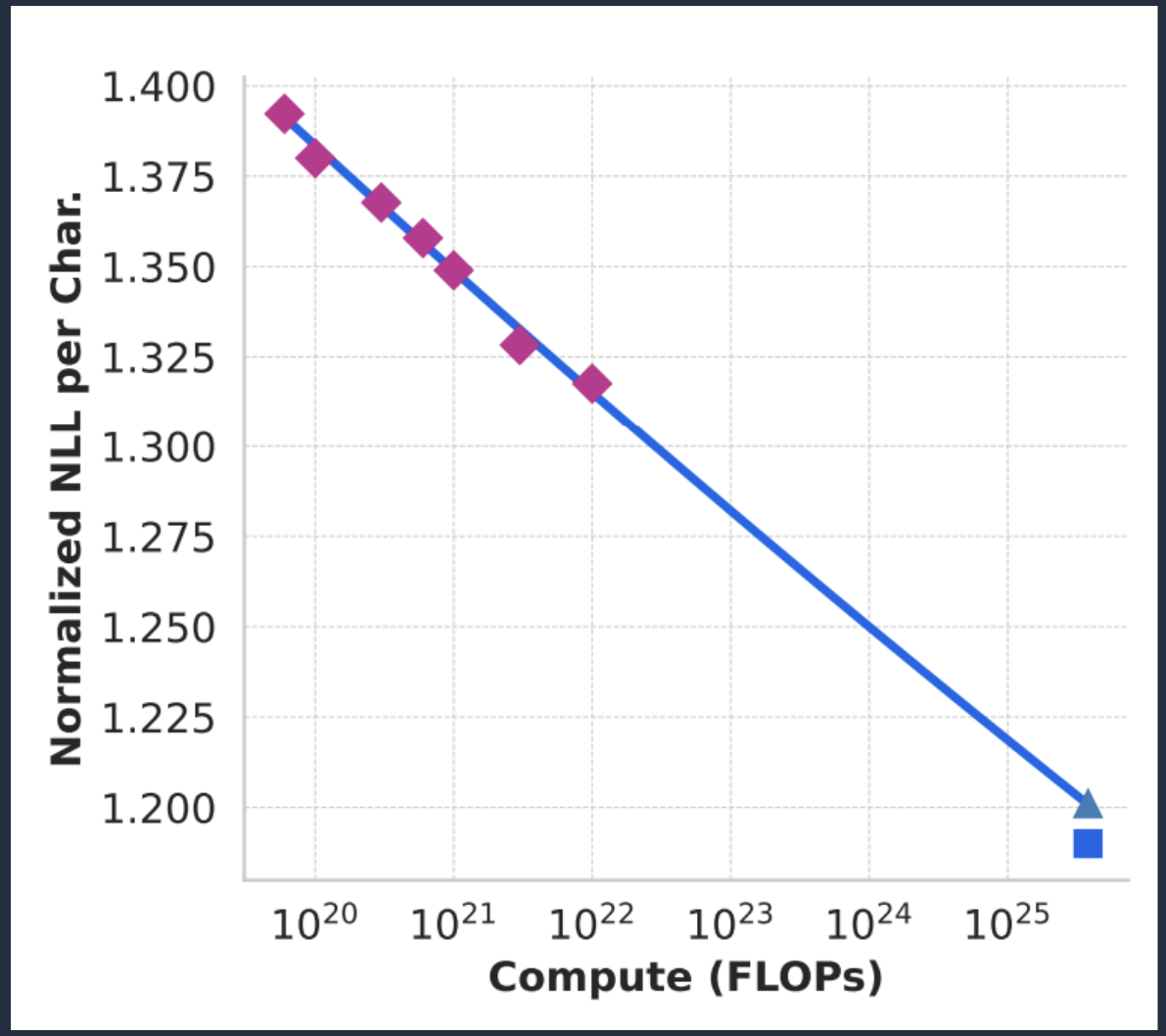
Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws
Sardana et al., 2023

$$N^*(\ell, D_{\text{inp}}, D_{\text{out}}), D^*_{\text{tr}}(\ell, D_{\text{inp}}, D_{\text{out}}) =$$

$$\underset{N, D_{\text{tr}} \mid L(N, D_{\text{tr}}) = \ell}{\arg\min} \frac{6 N D_{\text{tr}} C_{\text{tr}}}{\text{MFU}_{\text{tr}}} + 2 N C_{\text{inf}} \left[\frac{D_{\text{inp}}}{\text{MFU}_{\text{inp}}} + \frac{D_{\text{out}}}{\text{MFU}_{\text{out}}} \right]$$

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The Llama 3 Herd of Models
Dubey et al., 2024

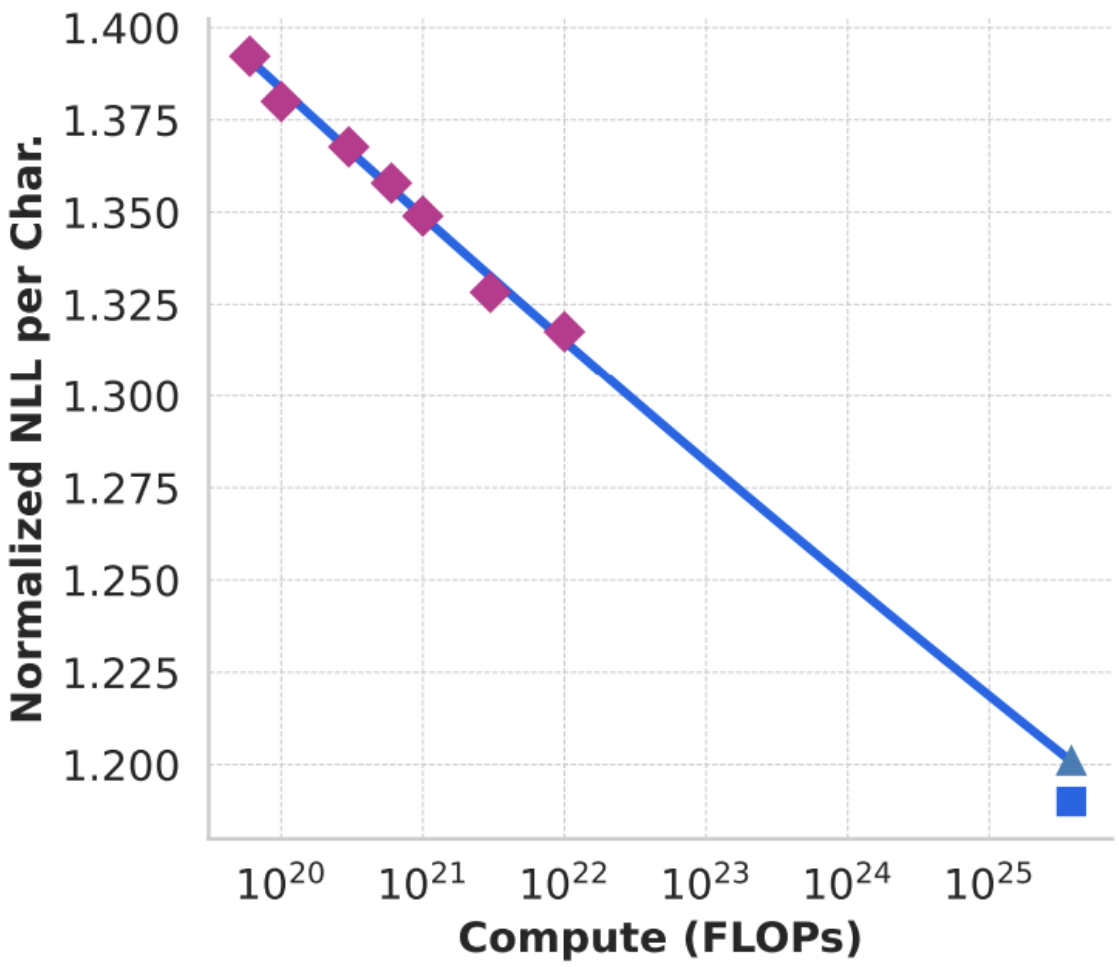


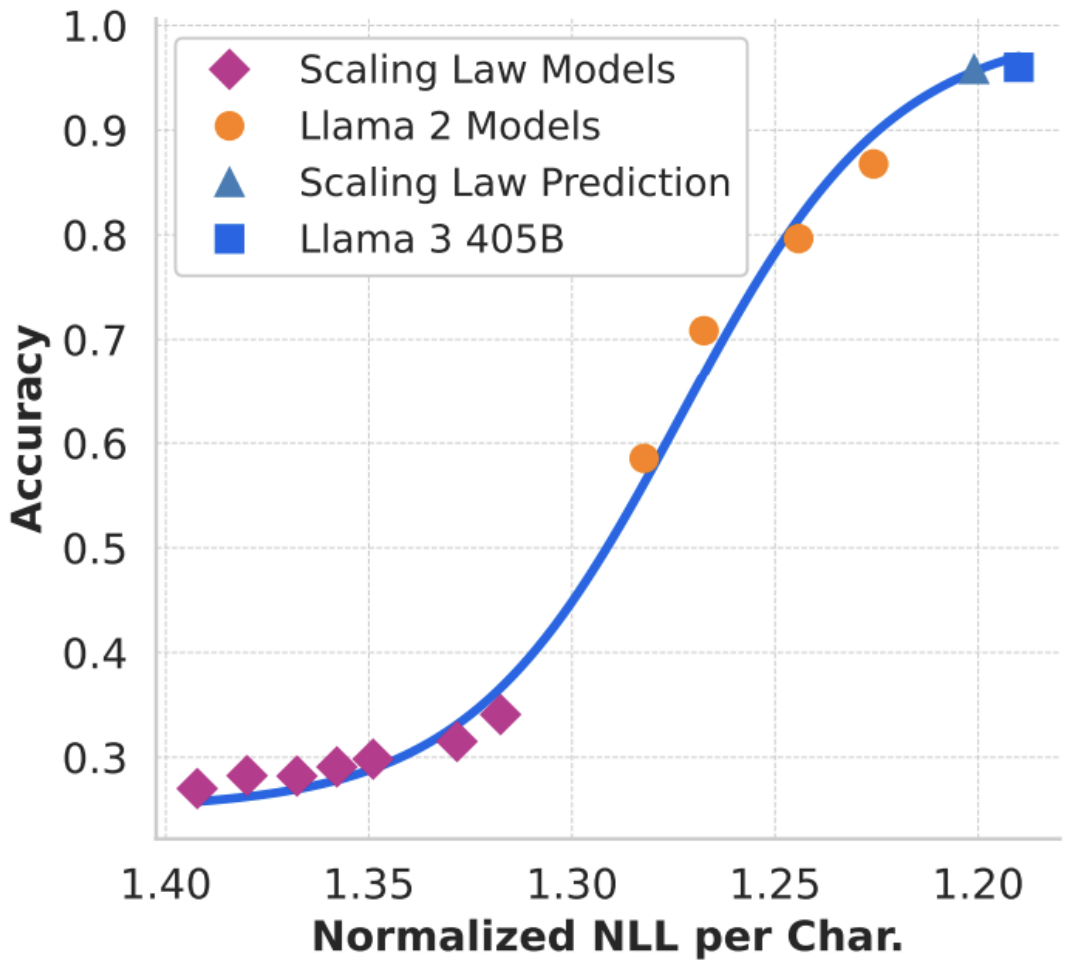
Scaling Law Models

Llama 3 405B

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The Llama 3 Herd of Models
Dubey et al., 2024





© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

TAKE AWAYS

IMPRACTICAL

UNDERTRAINED → CHINCHILLA-OPTIMAL → OVERTRAINED

FEW SMALL MODELS → CORRECT MODEL
↓
SCALING LAWS → SCALING IMPACT — 1000 X
FINAL PERFORMANCE — INCL. DOWNSTREAM

FAIL FORWARD

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Q&A

Mariano Kamp

mariano-kamp

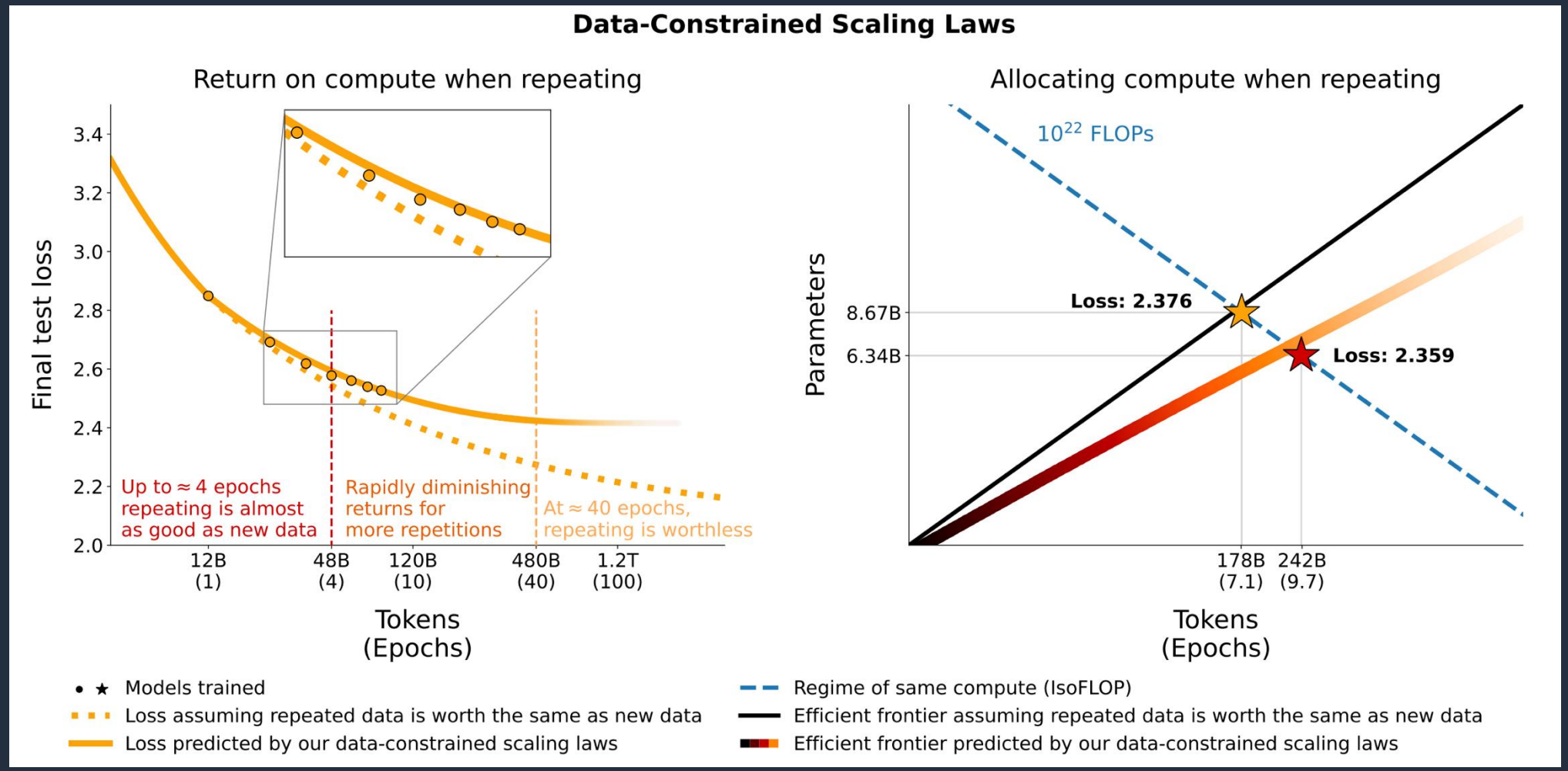
mkamp@amazon.de

Please complete the session survey.

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Backup

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



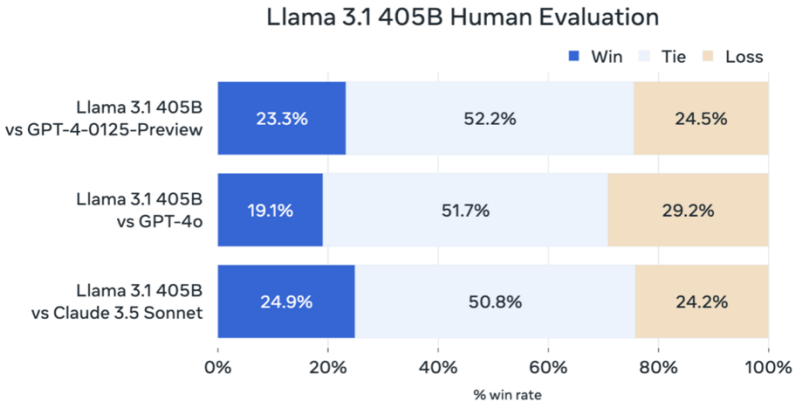
Scaling Data-Constrained Language Models
Müninghoff, 2023

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Llama 3 closes the gap between open and closed models

In April, Meta dropped the Llama 3 family, 3.1 in July, and 3.2 in September. Llama 3.1 405B, their largest to-date, is able to hold its own against GPT-4o and Claude 3.5 Sonnet across reasoning, math, multilingual, and long-context tasks. This marks the first time an open model has closed the gap with the proprietary frontier.

- Meta stuck to the same decoder-only transformer architecture that it's used since Llama 1, with minor adaptations, namely more transformer layers and attention heads.
- Meta used an incredible 15T tokens to train the family. While this blew through the "Chinchilla-optimal" amount of training compute, they found that both the 8B and 70B models improved log-linearly up to 15T.
- Llama 3.1 405B was trained over 16,000 H100 GPUs, the first Llama model trained at this scale.
- Meta followed up with Llama 3.2 in September, which incorporated 11B and 90B VLMs (Llama's multimodal debut). The former was competitive with Claude 3 Haiku, the latter with GPT-4o-mini. The company also released 1B and 3B text-only models, designed to operate on-device.
- Llama-based models have now racked up over 440M downloads on Hugging Face.



stateof.ai 2024

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

chinchilla's wild implications

https://www.alignmentforum.org/posts/6Fpvch8RR29qLEWNH/chinchilla-s-wild-implications

$$L(N, D) = \underbrace{\frac{A}{N^{\alpha}}}_{\text{finite model}} + \underbrace{\frac{B}{D^{\beta}}}_{\text{finite data}} + \underbrace{E}_{\text{irreducible}}$$

Here's the same thing, with the constants fitted to DeepMind's experiments on the MassiveText dataset[3].

$$L(N, D) = \underbrace{\frac{406.4}{N^{0.34}}}_{\text{finite model}} + \underbrace{\frac{410.7}{D^{0.28}}}_{\text{finite data}} + \underbrace{1.69}_{\text{irreducible}}$$

Gopher is a model with 280B parameters, trained on 300B tokens of data. What happens if we plug in those numbers?

$$L(280 \cdot 10^9,\ 300 \cdot 10^9) = \underbrace{0.052}_{\text{finite model}} + \underbrace{0.251}_{\text{finite data}} + \underbrace{1.69}_{\text{irreducible}} = 1.993$$

Chinchilla is a model with the same training compute cost as Gopher, allocated more evenly between the two terms in the equation.

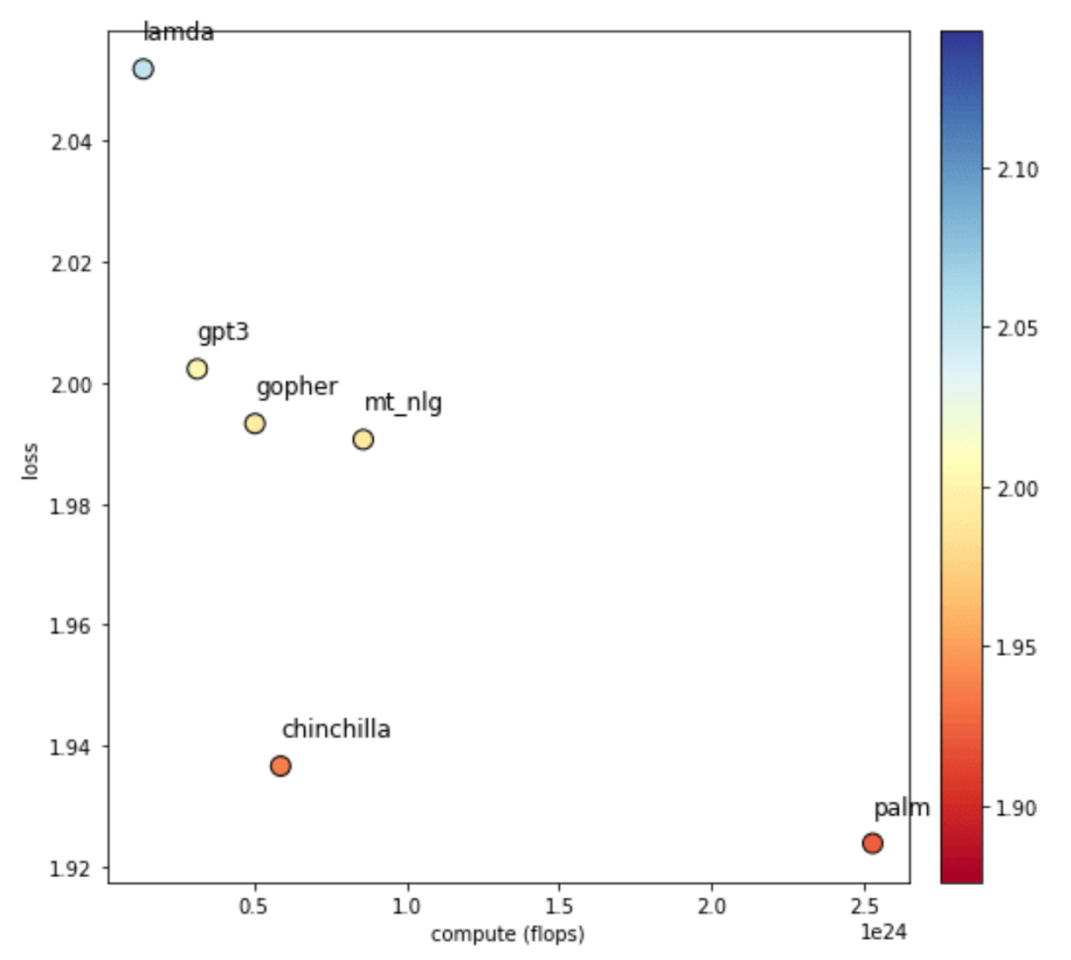
It's 70B params, trained on 1.4T tokens of data. Let's plug that in:

$$L(70 \cdot 10^9,\ 1400 \cdot 10^9) = \underbrace{0.083}_{\text{finite model}} + \underbrace{0.163}_{\text{finite data}} + \underbrace{1.69}_{\text{irreducible}} = 1.936$$

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

chinchilla's wild implications

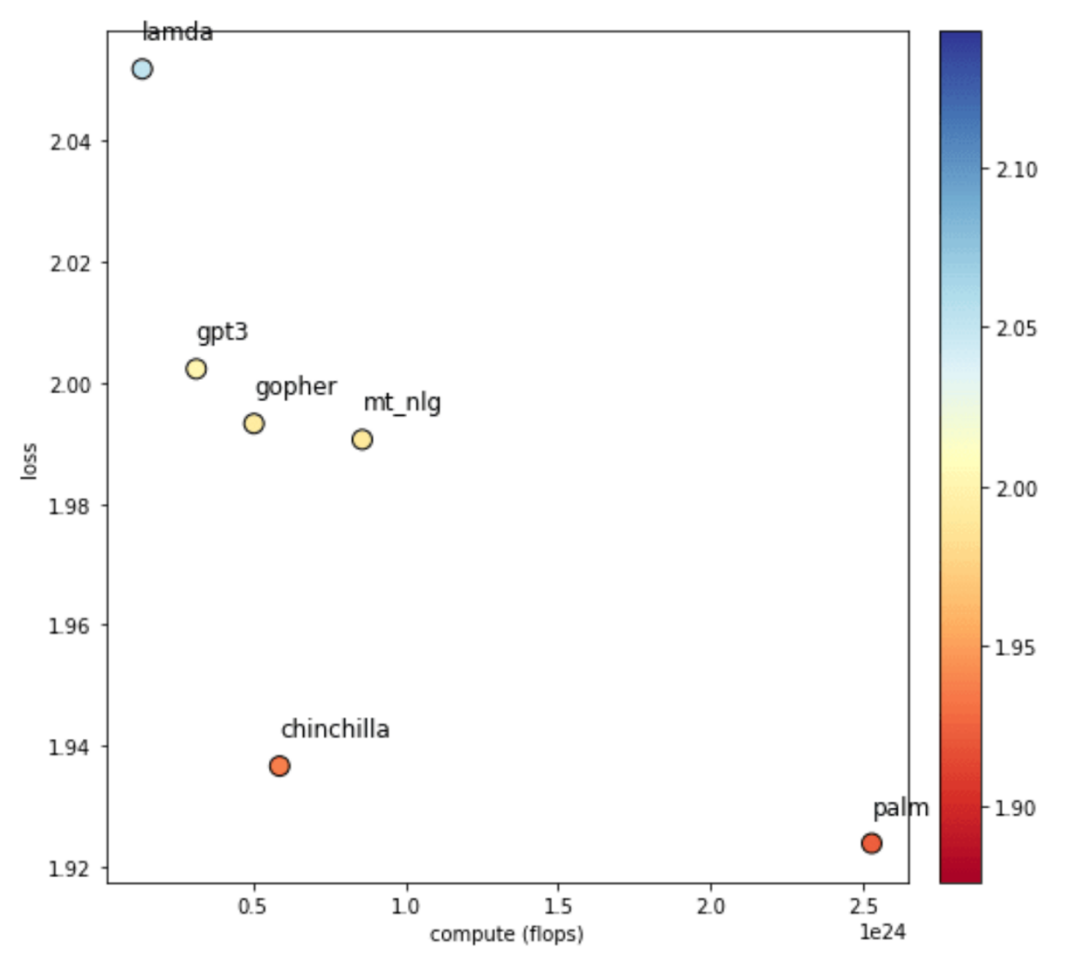
https://www.alignmentforum.org/posts/6Fpvch8RR29qLEWNH/chinchilla-s-wild-implications

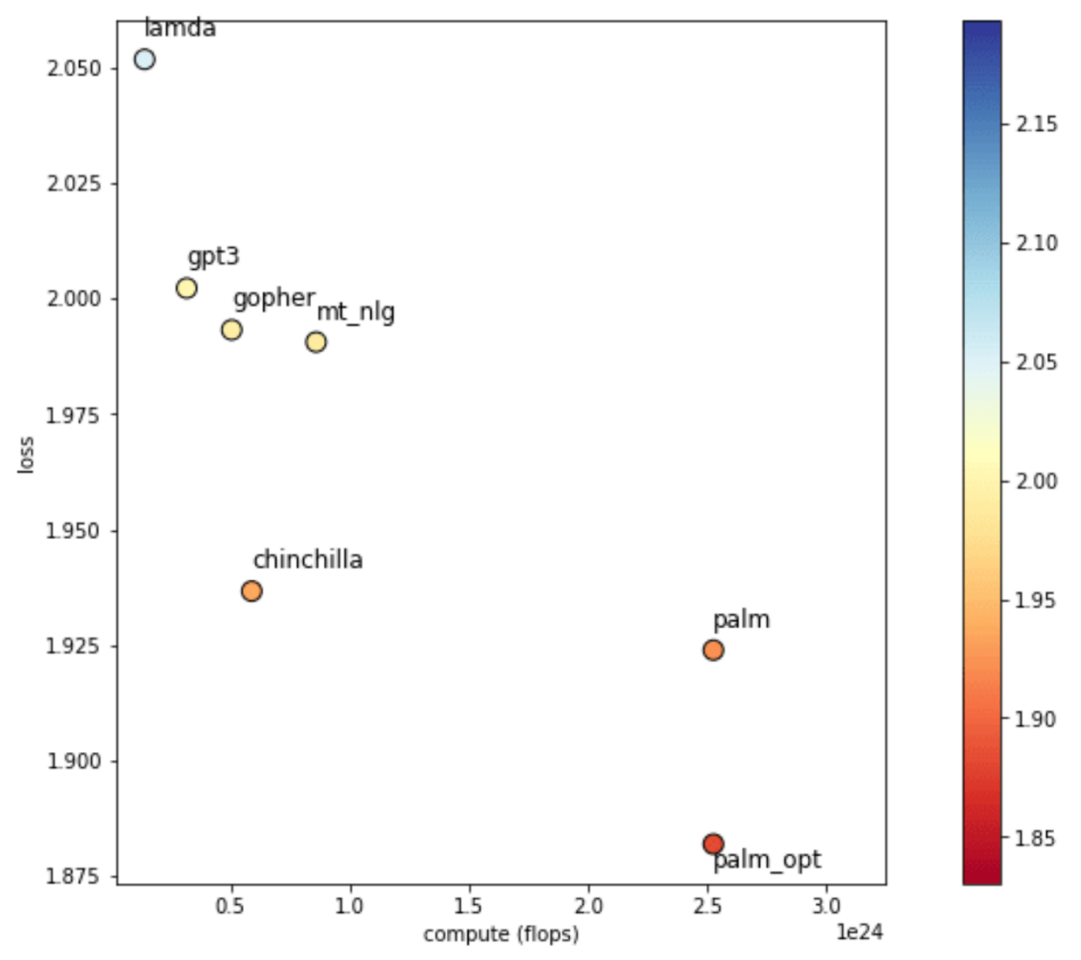


© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

chinchilla's wild implications

https://www.alignmentforum.org/posts/6Fpvch8RR29qLEWNH/chinchilla-s-wild-implications

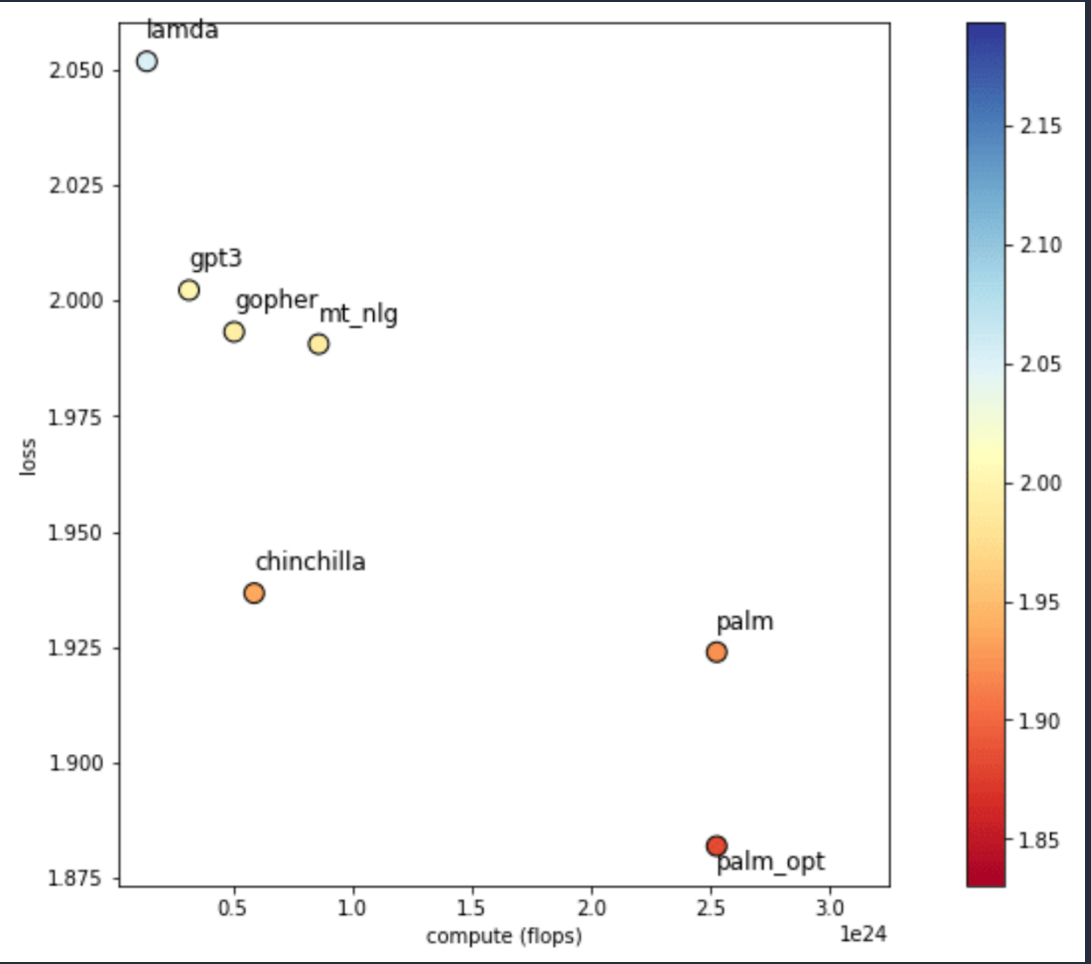


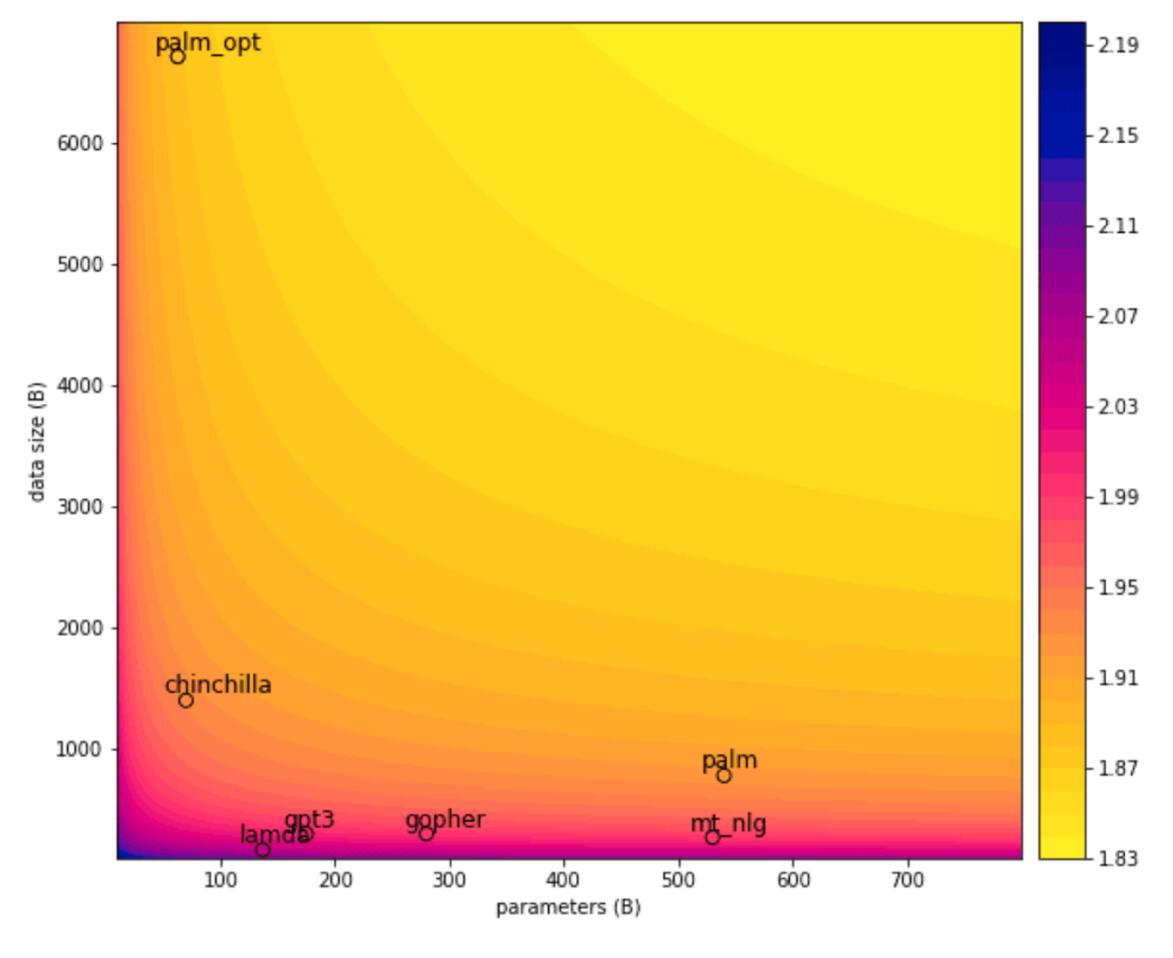


© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

chinchilla's wild implications

https://www.alignmentforum.org/posts/6Fpvch8RR29qLEWNH/chinchilla-s-wild-implications

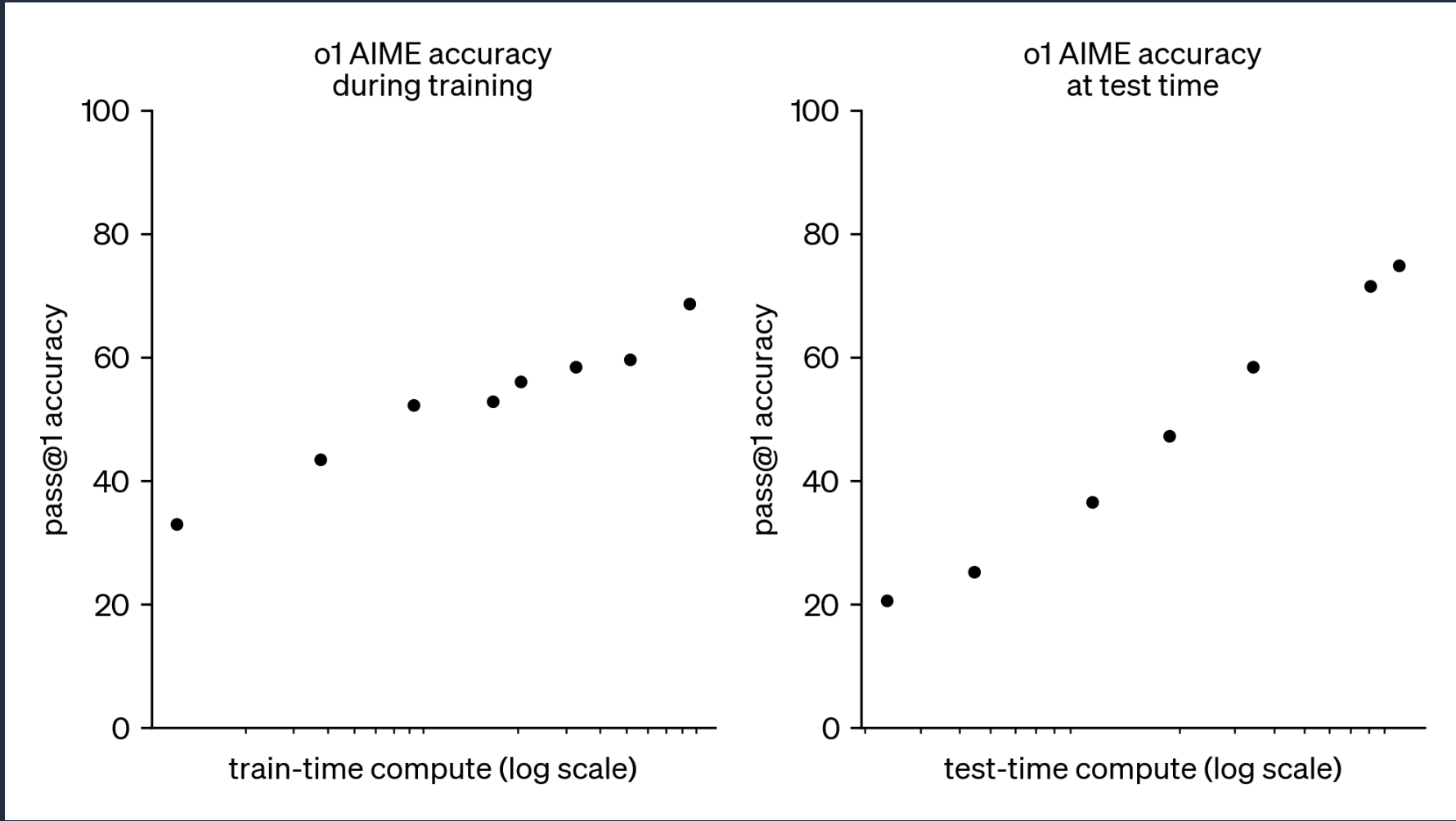




© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

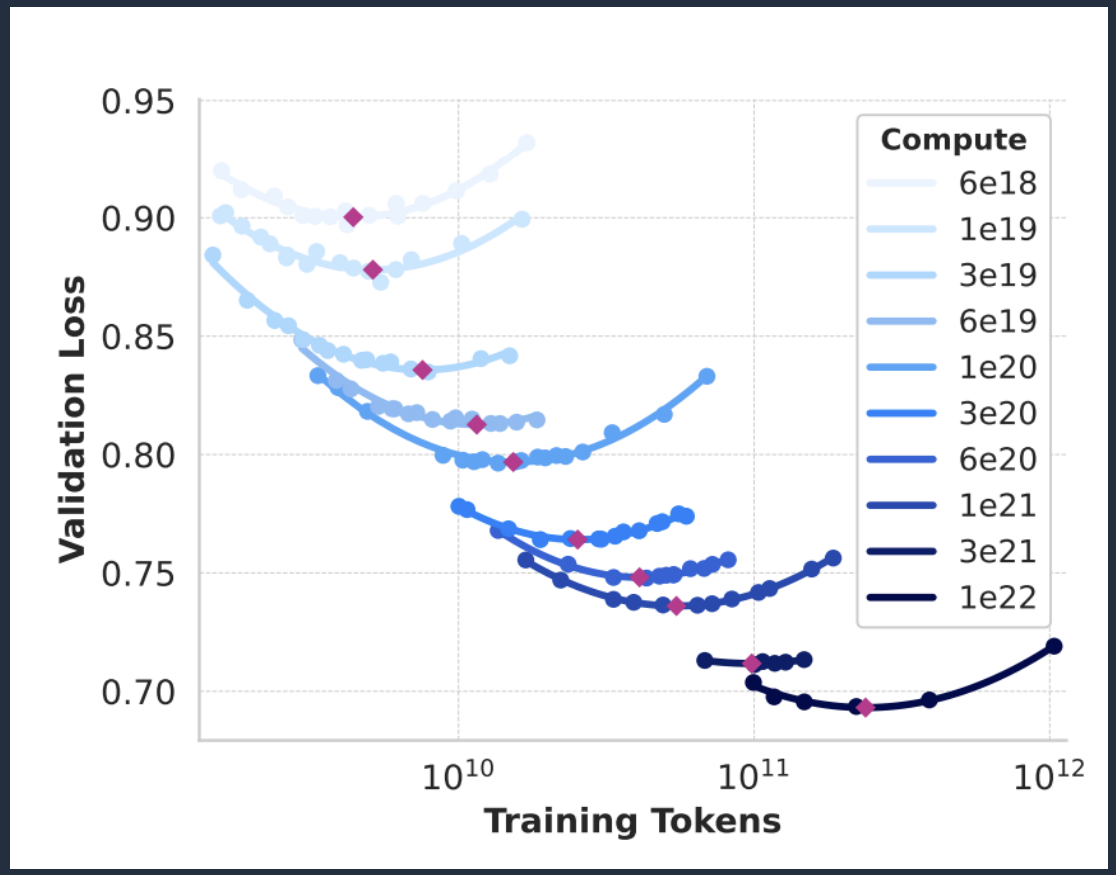
OpenAI o1

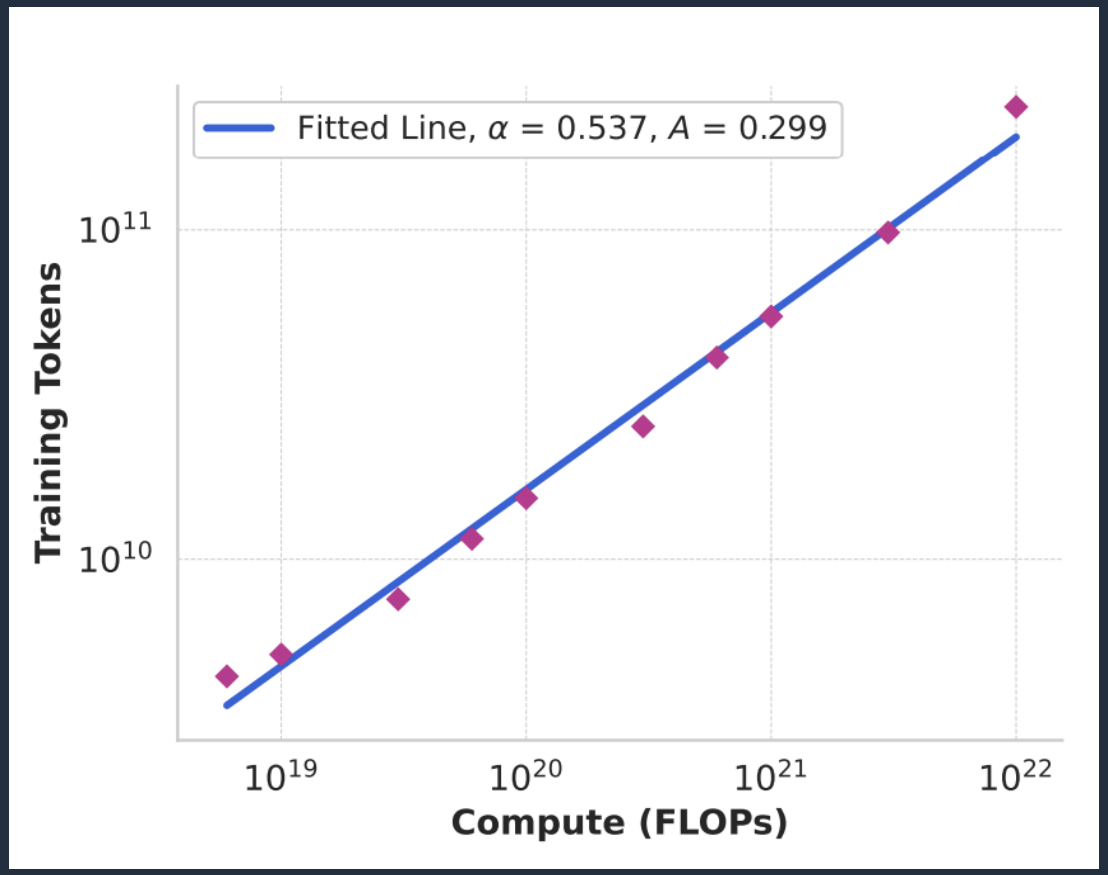
https://openai.com/index/learning-to-reason-with-llms/



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The Llama 3 Herd of Models
Dubey et al., 2024





© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.