DeepMind

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

ArXiv (2022)

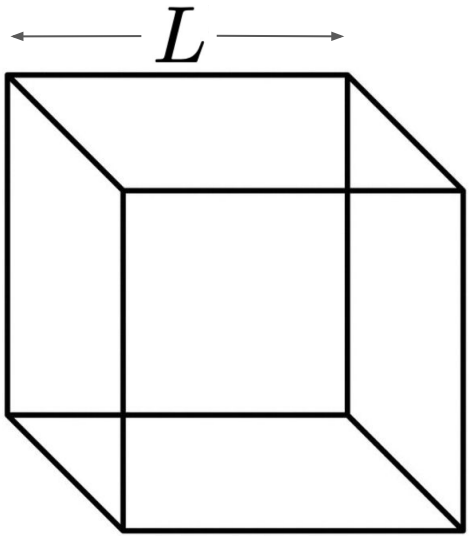
Presenters: Sourav Biswas and Ben Agro

What are Scaling Laws ?

- Describe how system properties change w.r.t. parameters
- Commonly used across various fields
  - Physics
  - Statistics
  - Biology

$$A \propto L^2$$

$$M \propto L^3$$



Prior Works on Scaling Laws in Neural Networks

- Neural networks (1989)
- Generalization of neural nets (1991)
- Classification (1993)
- General deep learning (2017)
- Generative modeling (2020)
- Language models (2020)

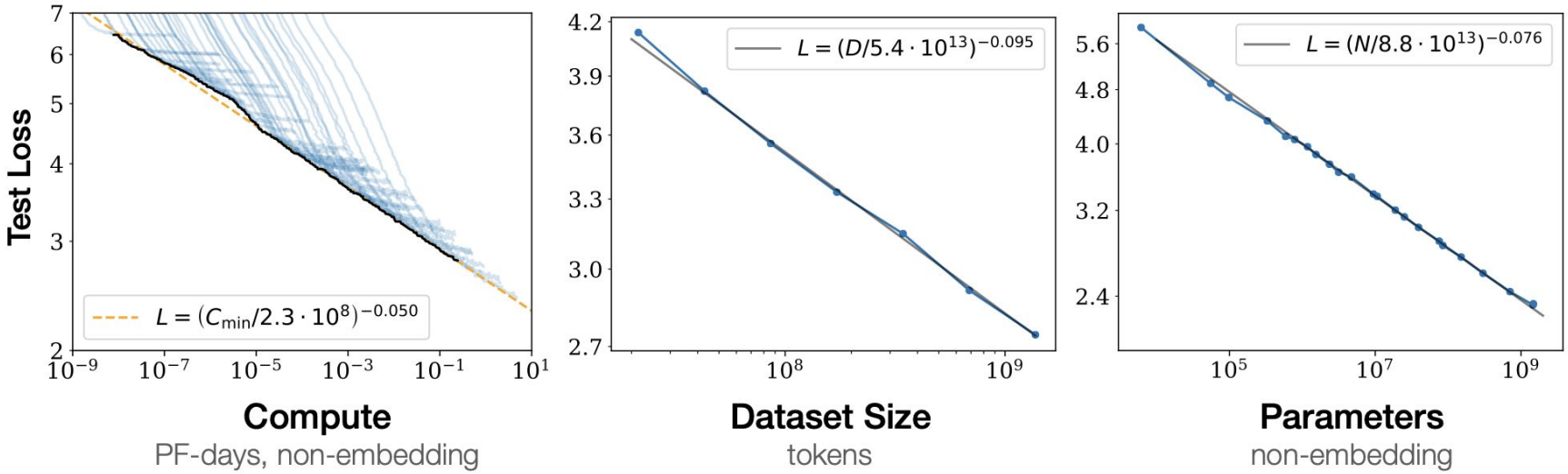
$$L \propto f(N)$$

$$L \propto g(D)$$

Ahmad et al. "Scaling and Generalization in Neural Networks: A Case Study." (1989).
Cohn et al. "Can neural networks do better than the Vapnik-Chervonenkis bounds." (1991).
Barkai et al. "Scaling Laws in Learning of Classification Tasks." (1993).
Hestness et al. "Deep Learning Scaling is Predictable, Empirically." (2017).
Henighan et al. "Scaling Laws for Autoregressive Generative Modeling." (2020).
Kaplan et al. "Scaling Laws for Neural Language Models." (2020).

Existing Work on Scaling Laws

- Kaplan et al. show a relationship between model size & performance
  - This motivated much of the progress and scaling of LLMs

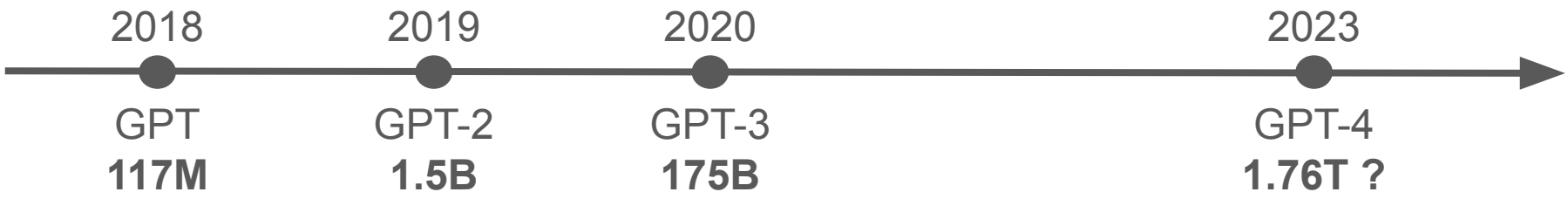


Kaplan et al. "Scaling Laws for Neural Language Models." *arXiv preprint arXiv:2001.08361* (2020).

History of LLM Progression

Layton, D. "ChatGPT — How we got to where we are today — a timeline of GPT development." https://medium.com/@dlaytonj2/chatgpt-how-we-got-to-where-we-are-today-a-timeline-of-gpt-development-f7a35dcc660e (2023).
Lubbad, M. "GPT-4 Parameters: Unlimited guide NLP's Game-Changer." https://mlubbad.medium.com/the-ultimate-guide-to-gpt-4-parameters-everything-you-need-to-know-about-nlps-game-changer-109b8767855a (2023).
Shree, P. "The Journey of Open AI GPT models." https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2 (2020).

History of LLM Progression



Layton, D. "ChatGPT — How we got to where we are today — a timeline of GPT development." https://medium.com/@dlaytonj2/chatgpt-how-we-got-to-where-we-are-today-a-timeline-of-gpt-development-f7a35dcc660e (2023).
Lubbad, M. "GPT-4 Parameters: Unlimited guide NLP's Game-Changer." https://mlubbad.medium.com/the-ultimate-guide-to-gpt-4-parameters-everything-you-need-to-know-about-nlps-game-changer-109b8767855a (2023).
Shree, P. "The Journey of Open AI GPT models." https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2 (2020).

Trends

- Larger models & more data
- Hyperparameter tuning
- Different architectures

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

Yang et al. "Tuning large neural networks via zero-shot hyperparameter transfer." (2021).
McCandlish et al. "An empirical model of large-batch training." (2018).
Shallue et al. "Measuring the effects of data parallelism on neural network training." (2018).
Zhang et al. "Which algorithmic choices matter at which batch sizes?." (2019).
Levine et al. "The depth-to-width interplay in self-attention." (2020).

Key Idea

- Motivation: LLM training is expensive;
  - Large models, with parameter count **N**
  - Large datasets, with **D** tokens
  - For a given compute budget **C** (in FLOPs)

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

Differences from Previous Works

This paper:

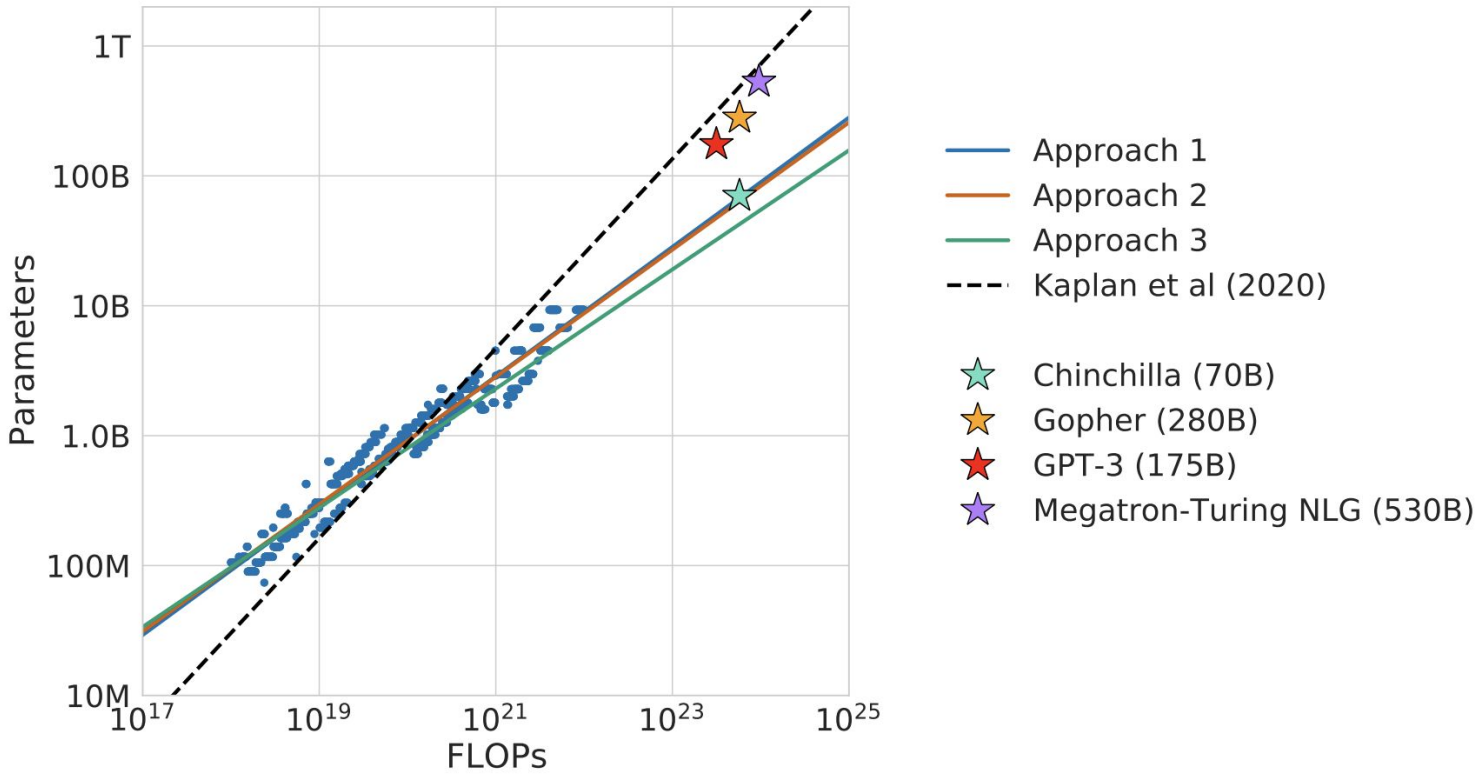
- **Varies** number of training tokens
- **Adapts** cosine LR schedule
- **Larger** models
- **Considers** embedding parameters

Kaplan et al:

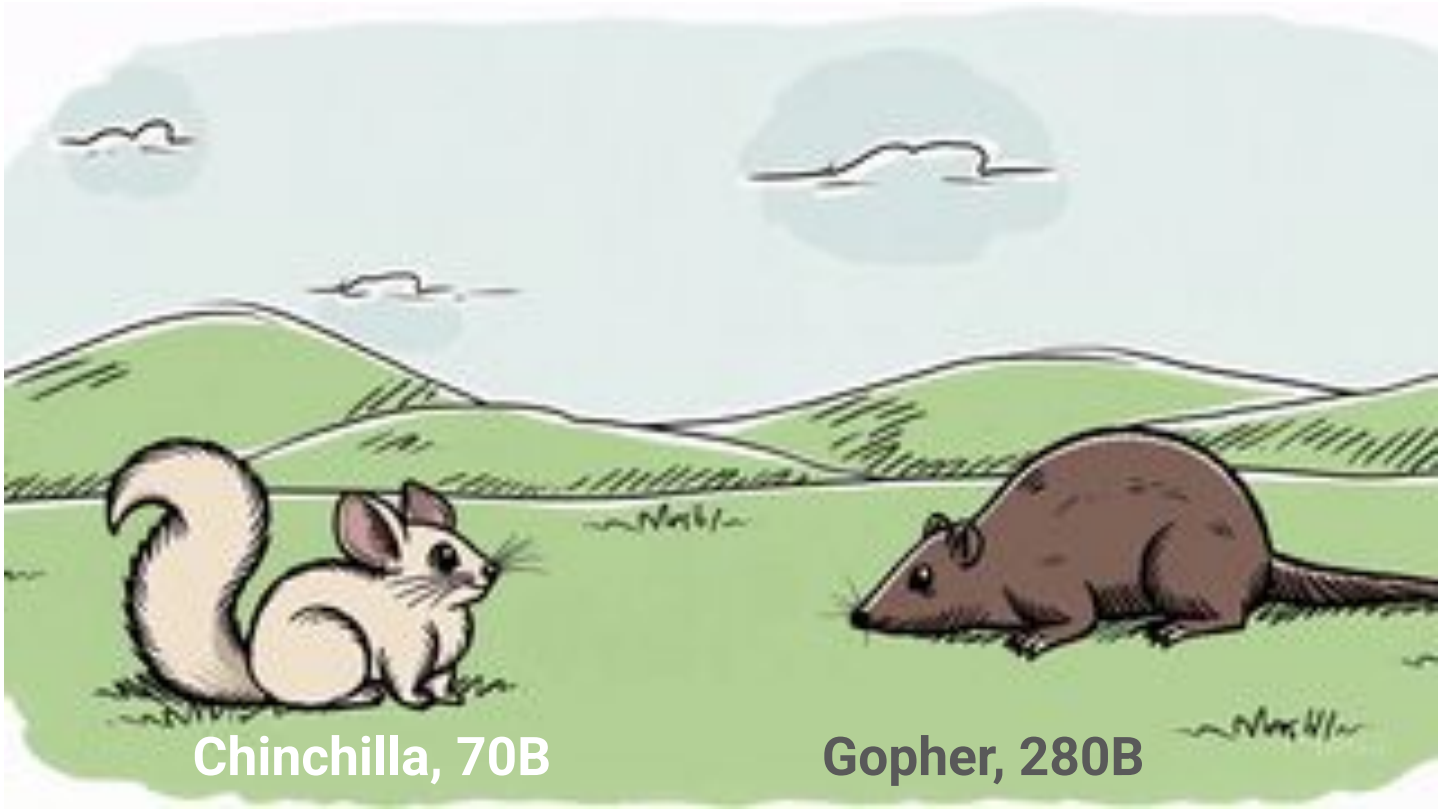
- **Fixed** number of training tokens
- **Fixed** cosine LR schedule
- **Smaller** models
- **Doesn’t count** embedding parameters

Kaplan et al. "Scaling Laws for Neural Language Models." *arXiv preprint arXiv:2001.08361* (2020).

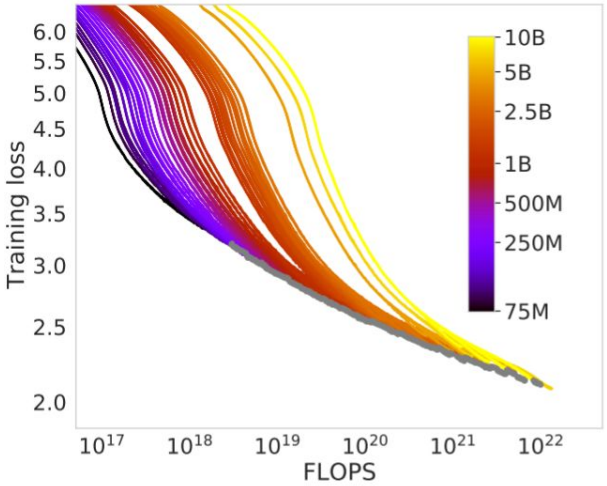
Key Idea


1T
100B
10B
1.0B
100M
10M
Parameters
10^{17}
10^{19}
10^{21}
10^{23}
10^{25}
FLOPs
Approach 1
Approach 2
Approach 3
Kaplan et al (2020)
Chinchilla (70B)
Gopher (280B)
GPT-3 (175B)
Megatron-Turing NLG (530B)

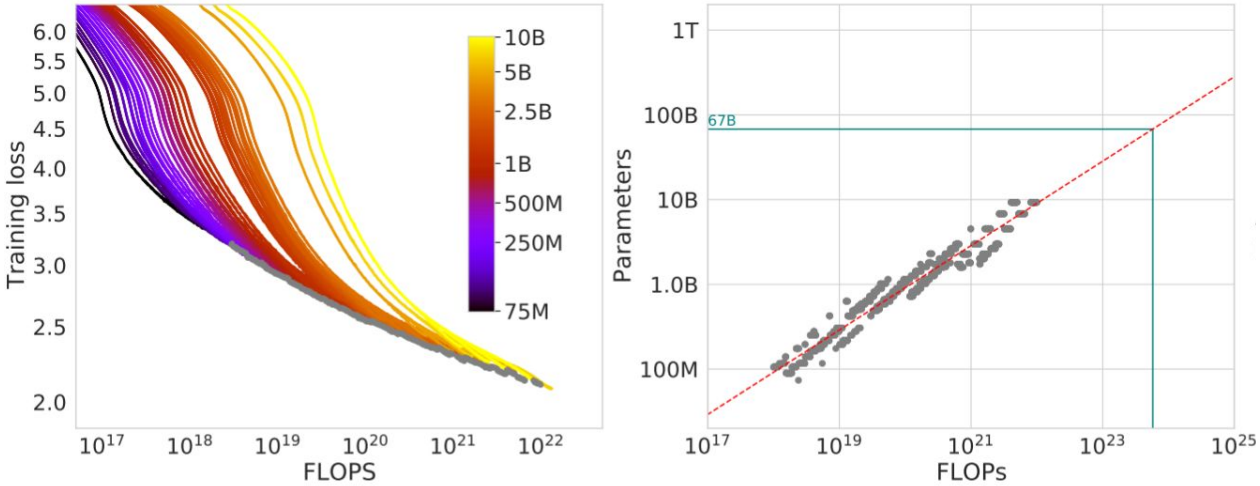
Key Idea



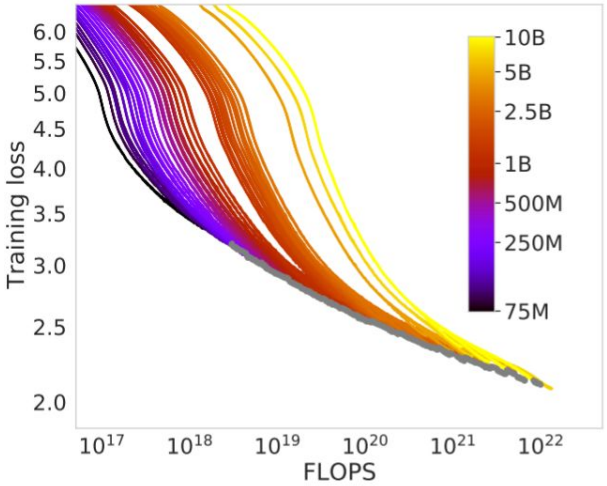
Approach 1: Fix model sizes & vary number of training tokens

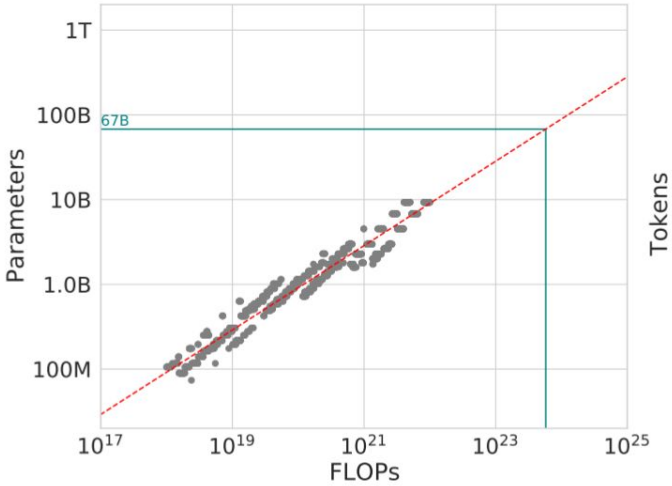


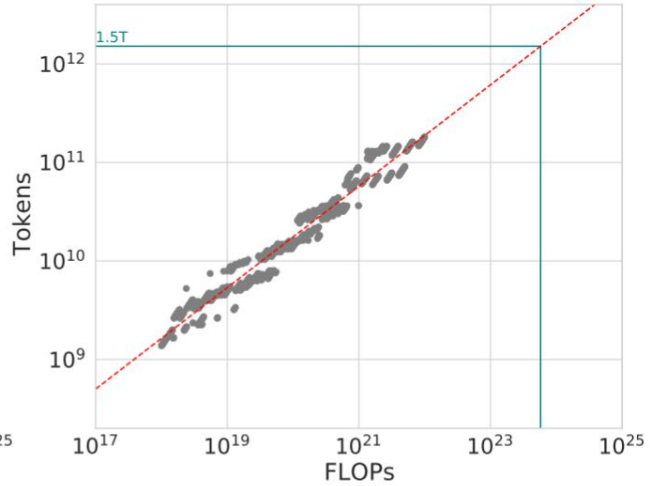
Approach 1: Minimum over training curves



Approach 1: Minimum over training curves

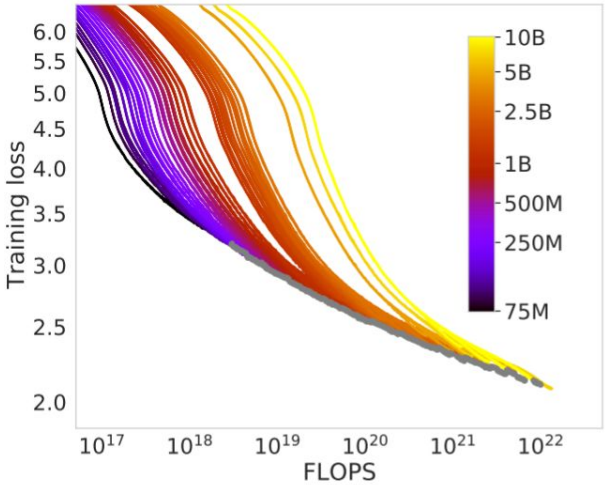


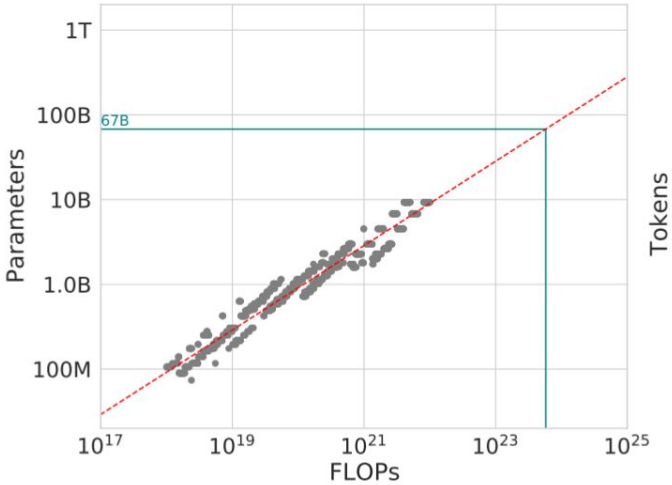


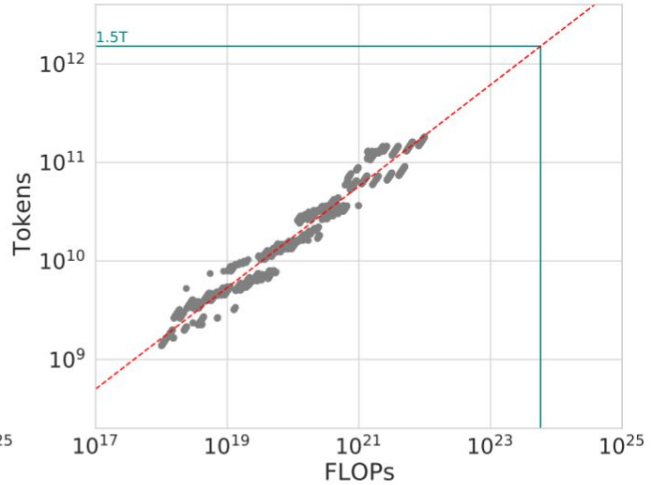


$$C \approx 6ND$$

Approach 1: Minimum over training curves

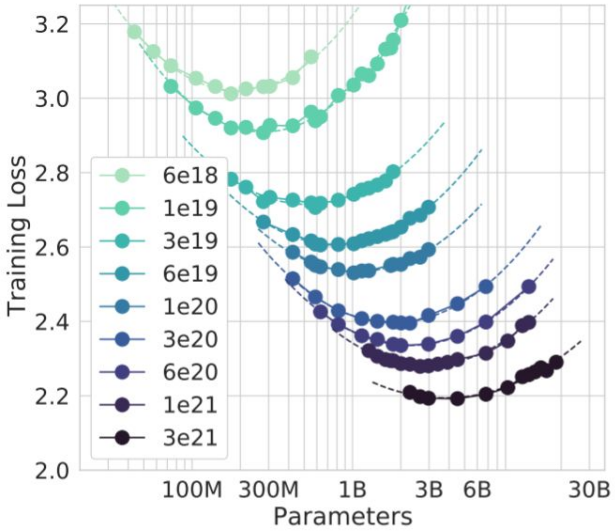






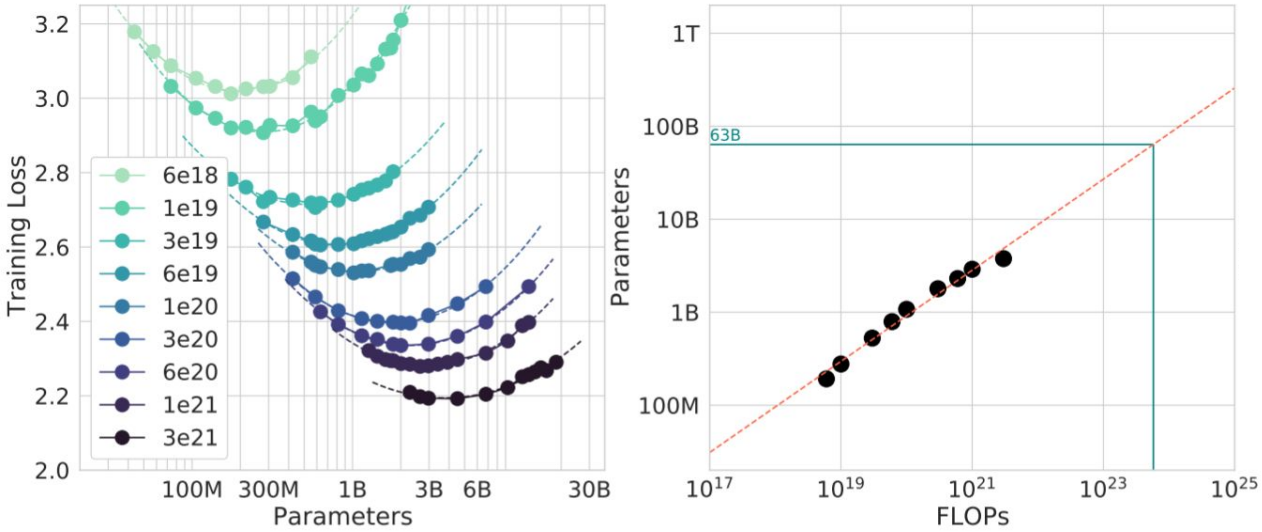
Approach	Coeff. a where $N_{opt} \propto C^a$	Coeff. b where $D_{opt} \propto C^b$
1. Minimum over training curves	0.50 (0.488, 0.502)	0.50 (0.501, 0.512)
Kaplan et al. (2020)	0.73	0.27

Approach 2: IsoFLOP Profiles

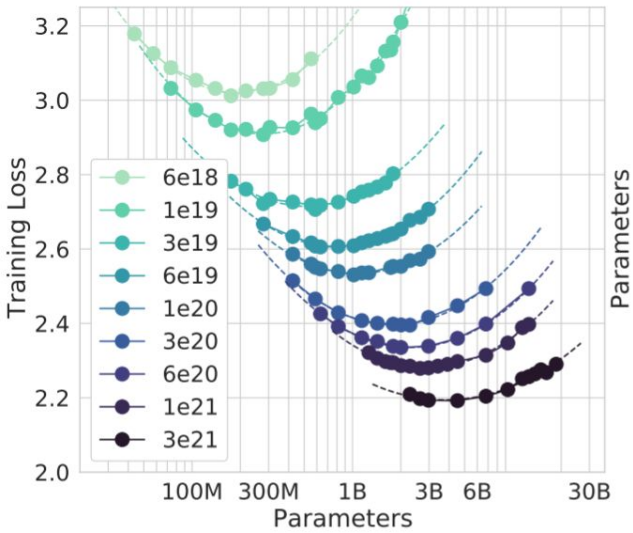


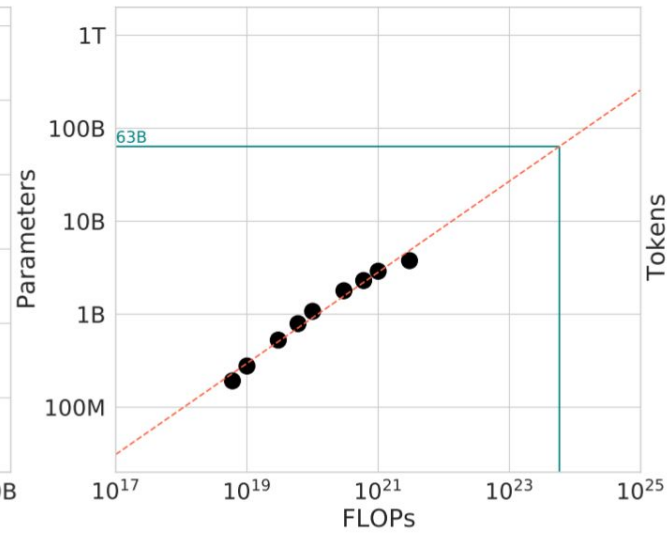
$$C \approx 6ND$$

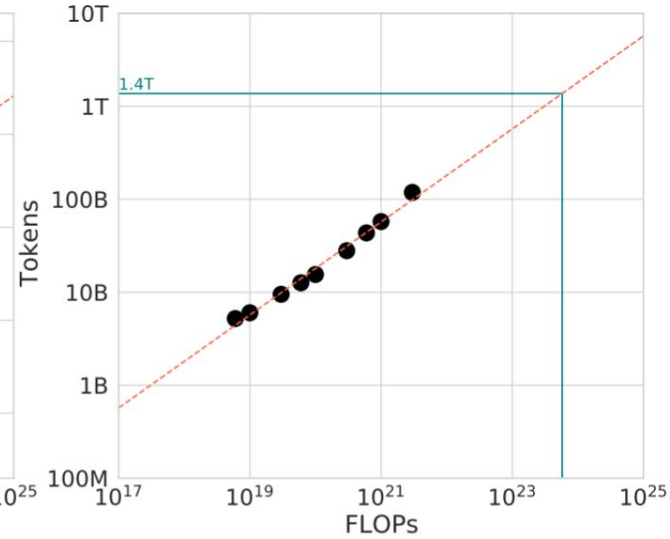
Approach 2: IsoFLOP Profiles



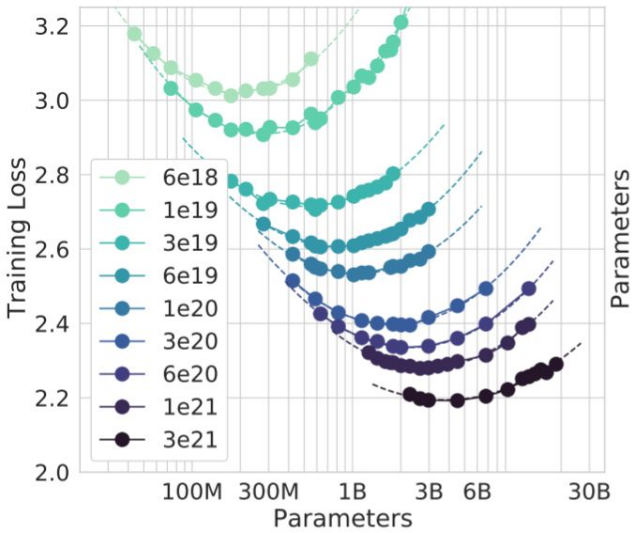
Approach 2: IsoFLOP Profiles

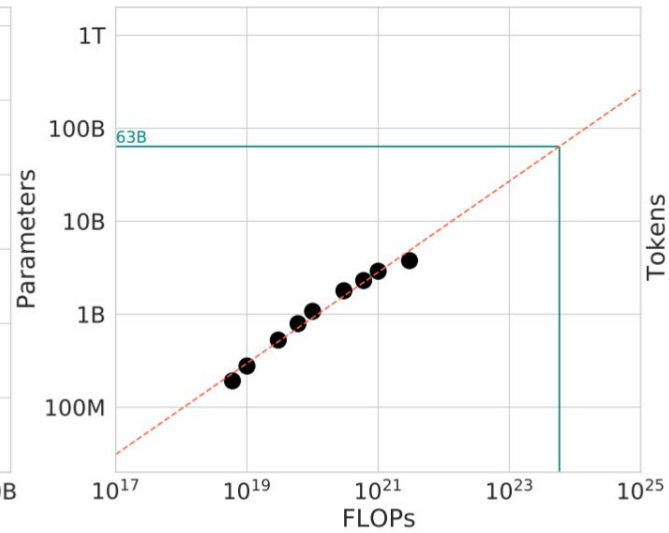


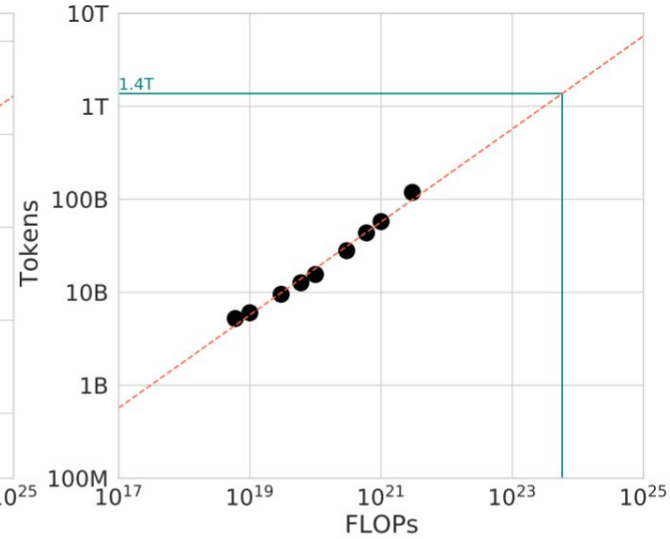




Approach 2: IsoFLOP Profiles

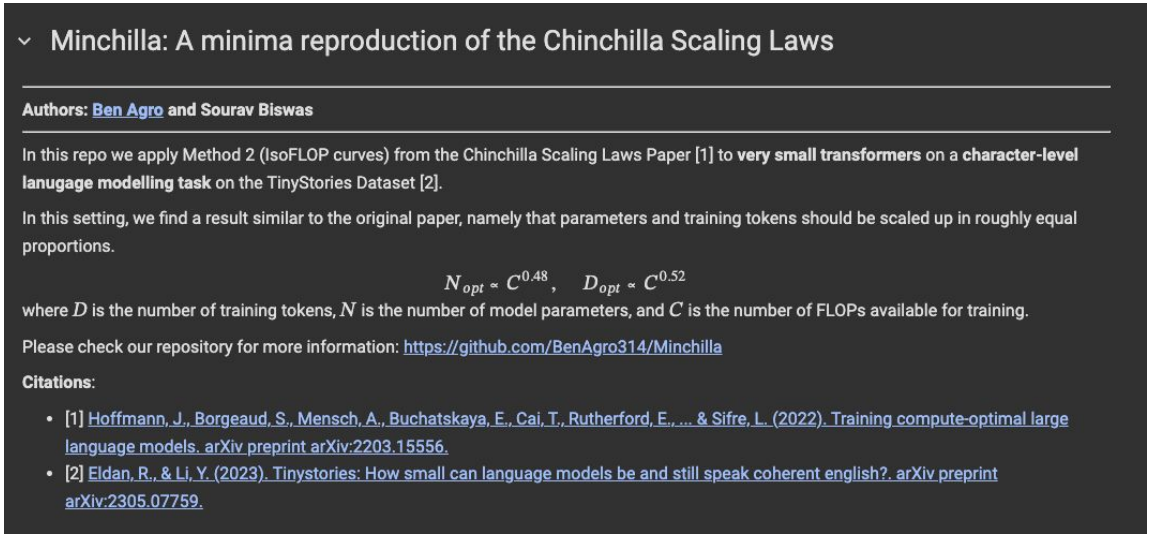






Approach	Coeff. a where $N_{opt} \propto C^a$	Coeff. b where $D_{opt} \propto C^b$
1. Minimum over training curves	0.50 (0.488, 0.502)	0.50 (0.501, 0.512)
2. IsoFLOP profiles	0.49 (0.462, 0.534)	0.51 (0.483, 0.529)
Kaplan et al. (2020)	0.73	0.27

Approach 2: MinChilla reproduction



Minchilla: A minima reproduction of the Chinchilla Scaling Laws

Authors: Ben Agro and Sourav Biswas

In this repo we apply Method 2 (IsoFLOP curves) from the Chinchilla Scaling Laws Paper [1] to **very small transformers** on a **character-level lanugage modelling task** on the TinyStories Dataset [2].

In this setting, we find a result similar to the original paper, namely that parameters and training tokens should be scaled up in roughly equal proportions.

$$N_{opt} \propto C^{0.48}, \quad D_{opt} \propto C^{0.52}$$

where D is the number of training tokens, N is the number of model parameters, and C is the number of FLOPs available for training.

Please check our repository for more information: https://github.com/BenAgro314/Minchilla

Citations:

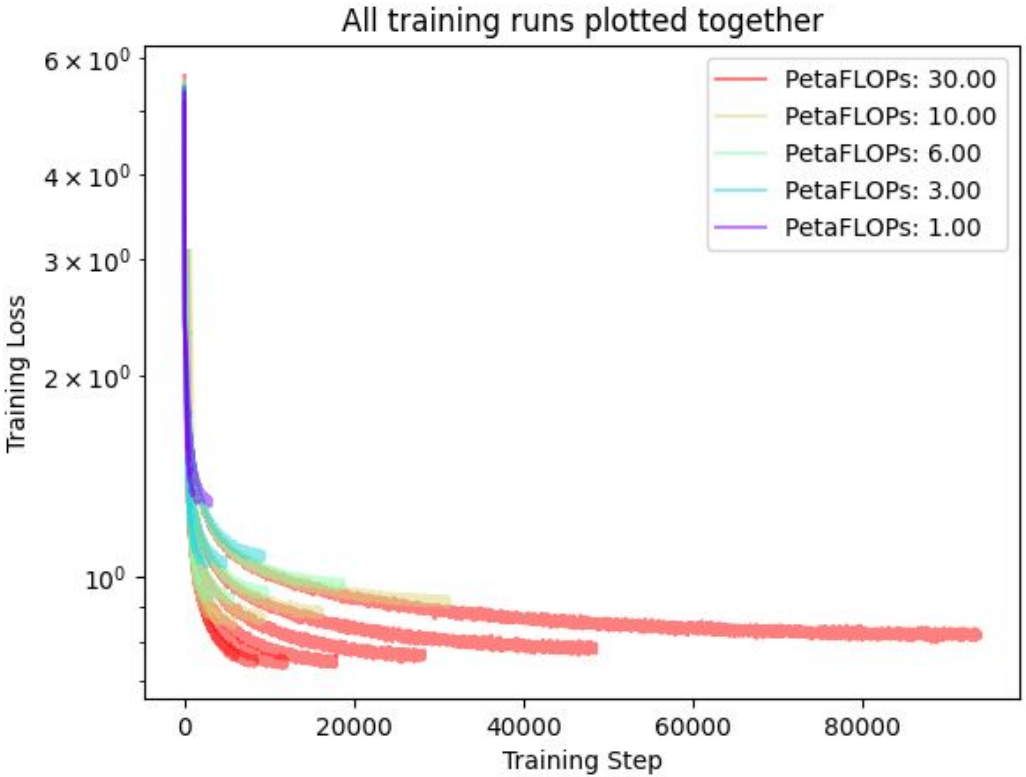
- [1] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.
- [2] Eldan, R., & Li, Y. (2023). Tinystories: How small can language models be and still speak coherent english?. arXiv preprint arXiv:2305.07759.

Tiny Stories Dataset:

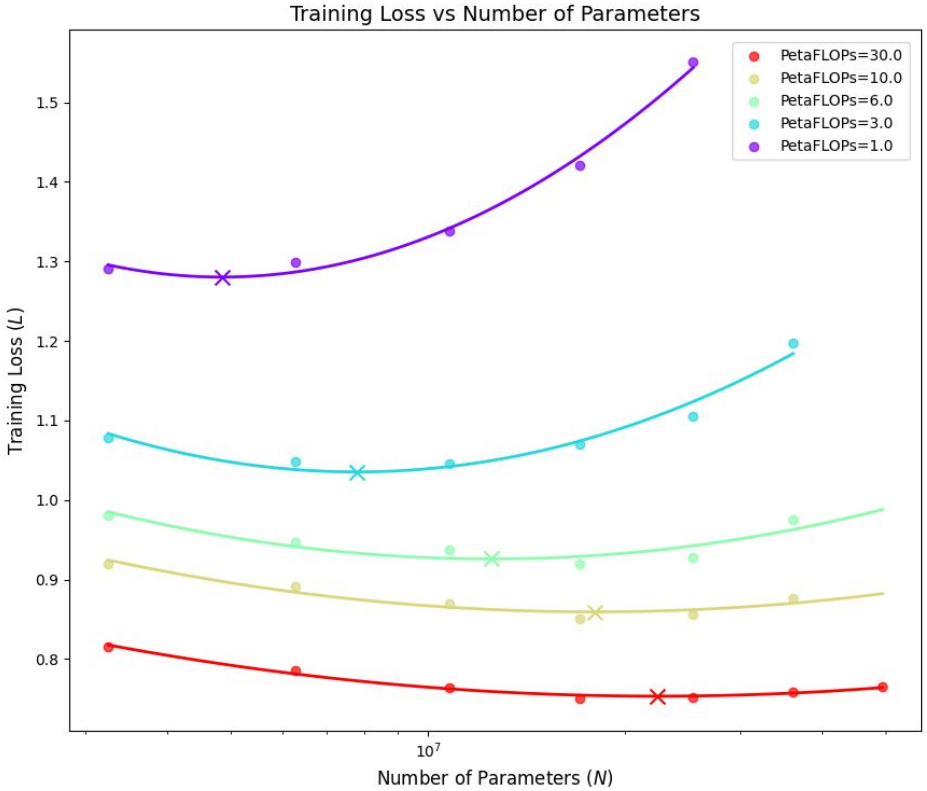
```
One day, a little girl named Lily found a
needle in her room. She knew it was
difficult to play with it because it was
sharp. Lily wanted to share the needle
with her mom, so she could sew a button
on her shirt. Lily went to her mom and
said, "Mom, I found this needle. Can you
share it with me and sew my shirt?" Her
mom smiled and said, "Yes, Lily, we can
share the needle and fix your shirt."
Together, they shared the needle and
sewed the button on Lily's shirt. It was
not difficult for them because they were
sharing and helping each other. After
they finished, Lily thanked her mom for
sharing the needle and fixing her shirt.
They both felt happy because they had
shared and worked together.
```

Eldan, Ronen, and Yuanzhi Li. "Tinystories: How small can language models be and still speak coherent english?." *arXiv preprint arXiv:2305.07759* (2023).

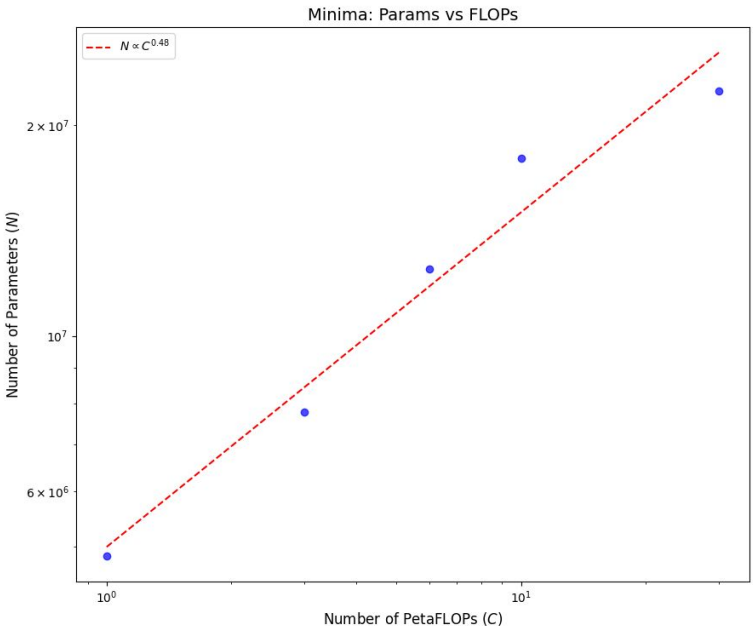
Approach 2: MinChilla reproduction

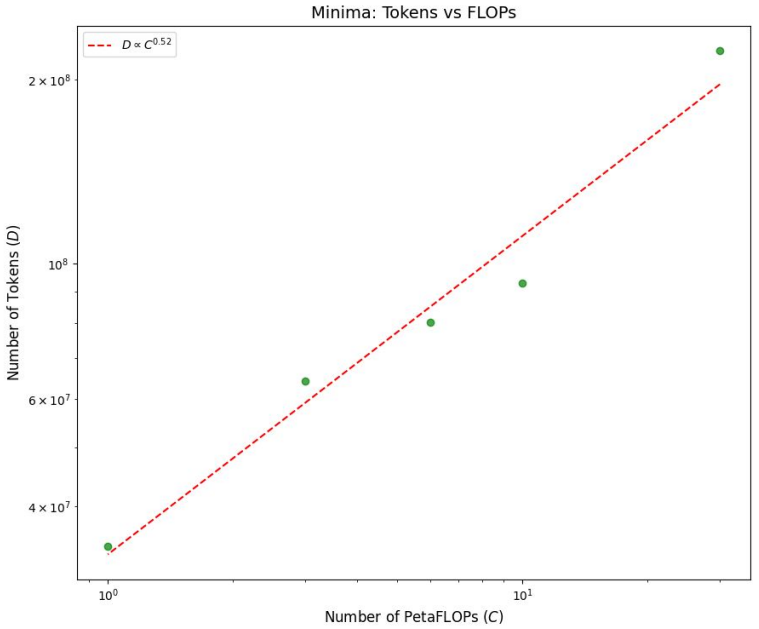


Approach 2: IsoFLOP Profiles



Approach 2: IsoFLOP Profiles





Source	Coeff. a where $N_{opt} \propto C^a$	Coeff. b where $D_{opt} \propto C^b$
Paper (Method 2)	0.49	0.51
Ours	0.48	0.52

Approach 3: Fitting a Parametric Function

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

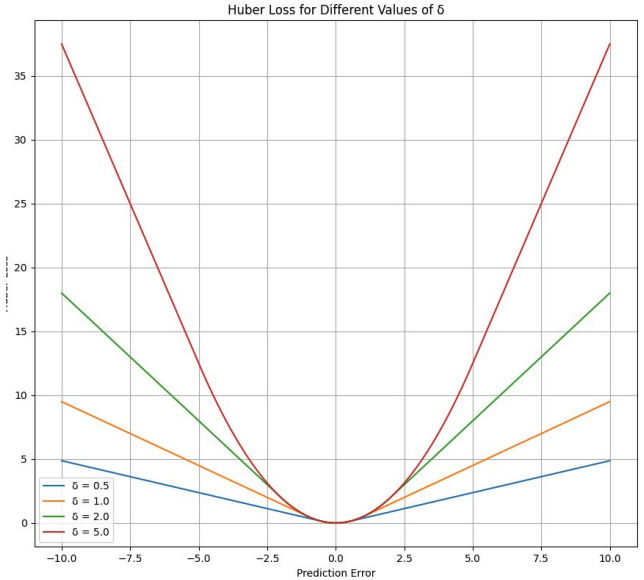
Entropy of natural text (minimum possible loss for any generative process)

Transformer under-performs ideal generative process

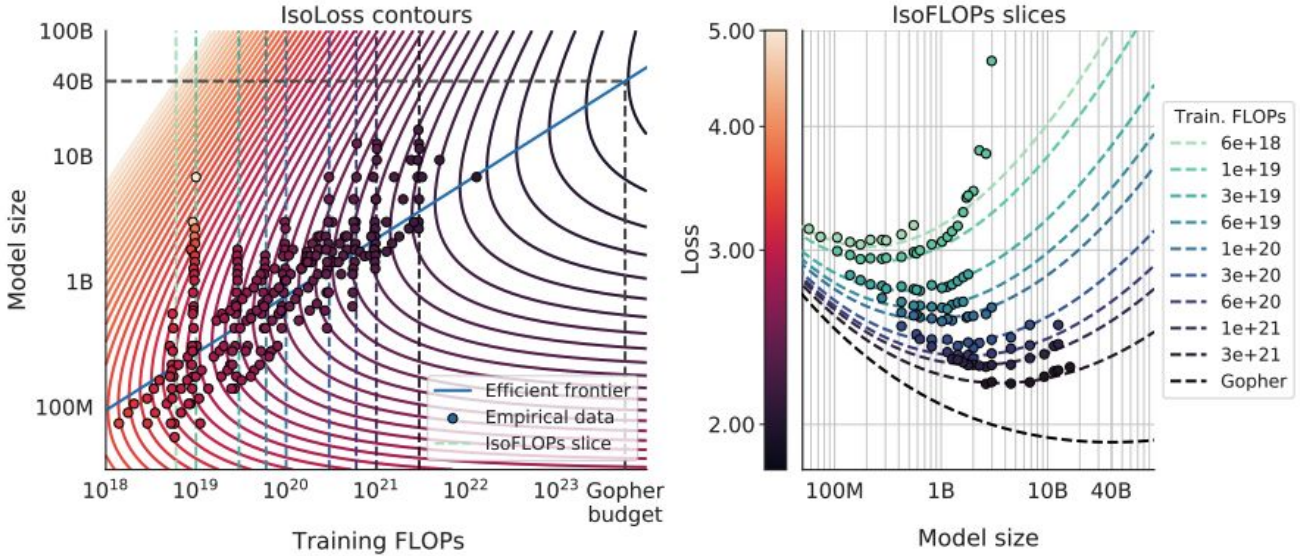
Transformer is not trained to convergence

Approach 3: Fitting a Parametric Function

$$\hat{L}(N, D) \triangleq E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} \qquad \min_{A,B,E,\alpha,\beta} \sum_{\text{Runs } i} \text{Huber}_{\delta}\left(\log \hat{L}(N_i, D_i) - \log L_i\right)$$



Approach 3: Fitting a Parametric Function



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

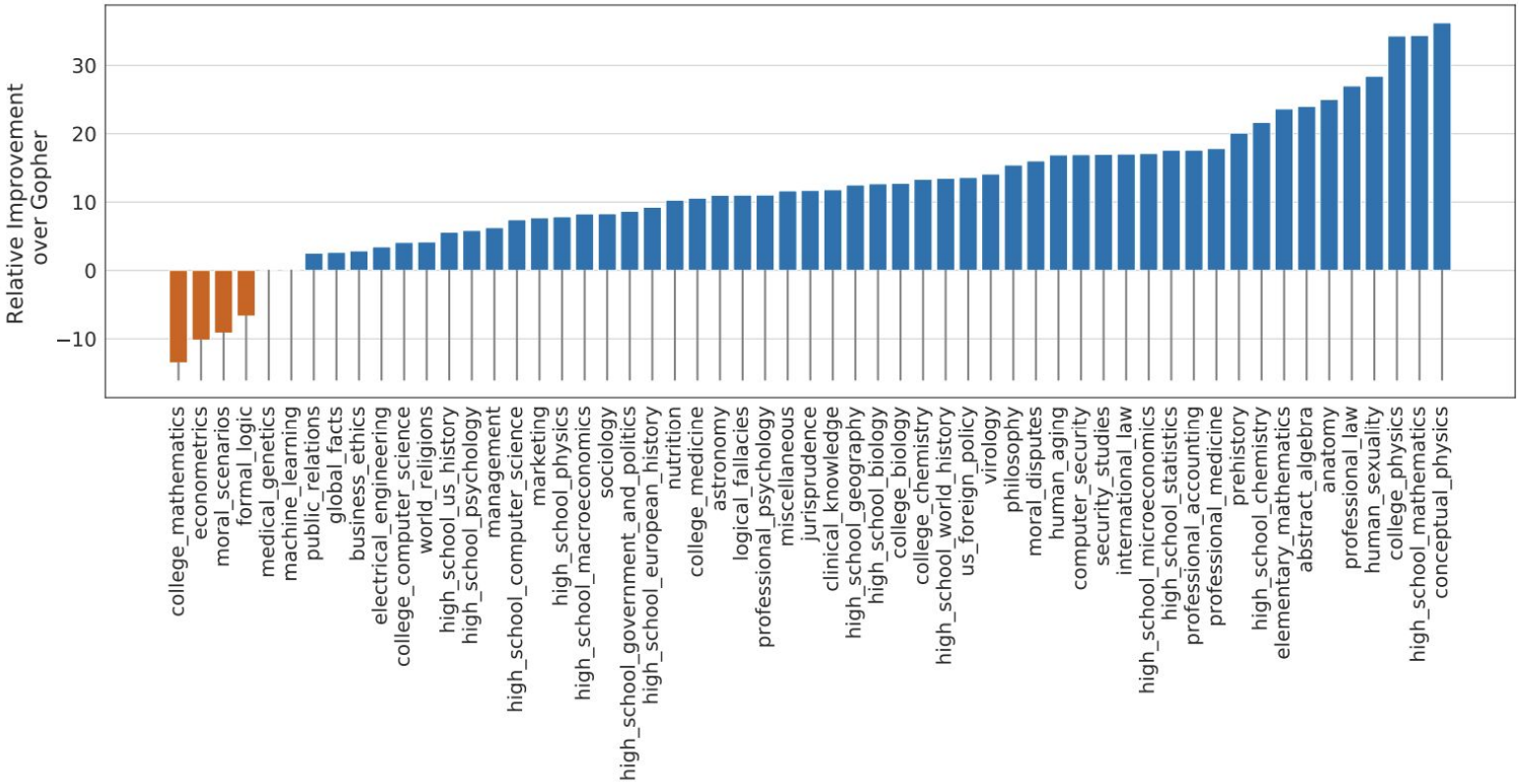
Optimal Model Scaling

- Gopher LLM: 280B parameters, trained on 300B tokens. **Non optimal**
- Test this hypothesis by training a compute-optimal version of the Gopher LLM, using the same number of FLOPs

Parameters	FLOPs	FLOPs (in *Gopher* 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

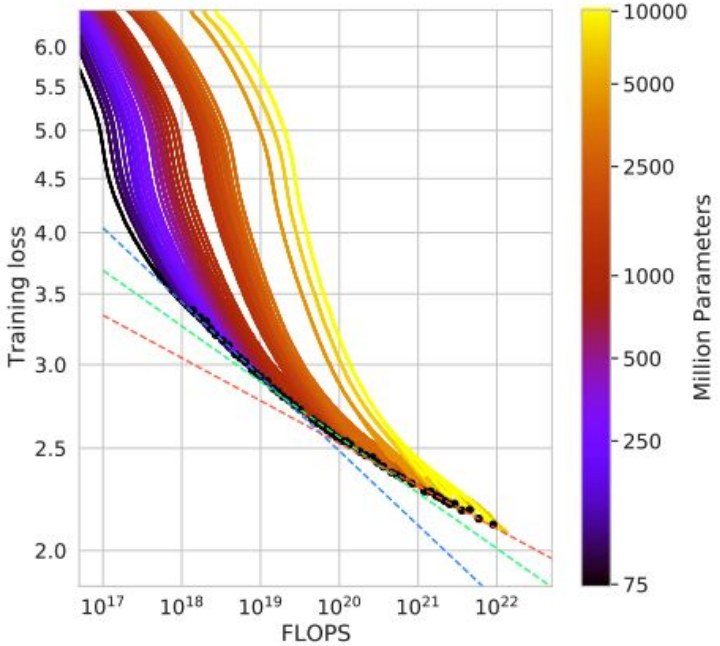
Rae, J. W., et al. "van den, Hendricks, LA, Rauh, M., Huang, P.-S.,... Irving, G.(2022)." *Scaling language*.

Chinchilla 70B vs Gopher 280B



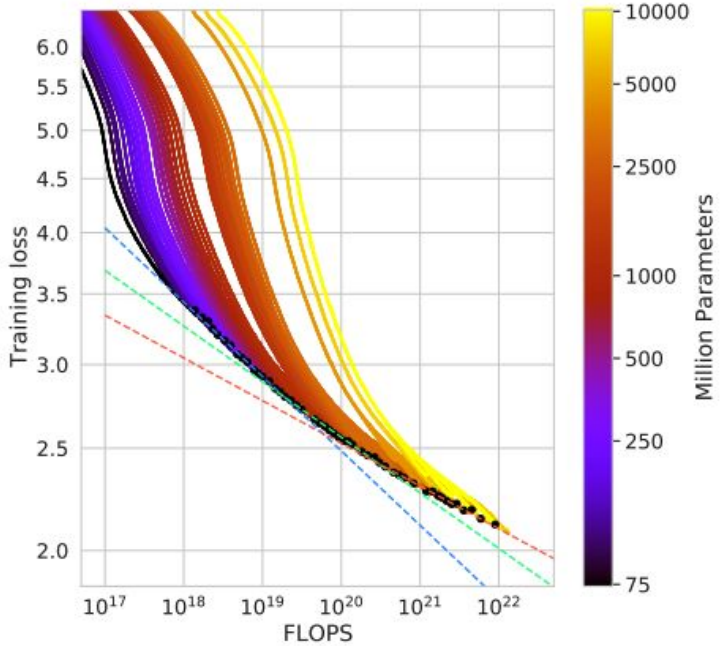
Discussion and Conclusion

- Parameters and tokens should be increased equally
- Models circa chinchilla were oversized



Scope and Limitations

- Assume a power law, but observe concavity
- Single epoch training
- Token quality
- Autoregressive transformers on next token prediction
- Training-time scaling laws, no inference time scaling laws
- An empirical observation, no theory



Bahri, Yasaman, et al. "Explaining neural scaling laws." *Proceedings of the National Academy of Sciences* 121.27 (2024): e2311878121.

Thank You For Listening!