Compute-optimal large language models

DeepMind 2022, https://arxiv.org/abs/2203.15556

summarized by Michael Scherbela
Deep Learning Seminar
Jan 18, 2023

universität wien

1) Find large language model scalling law

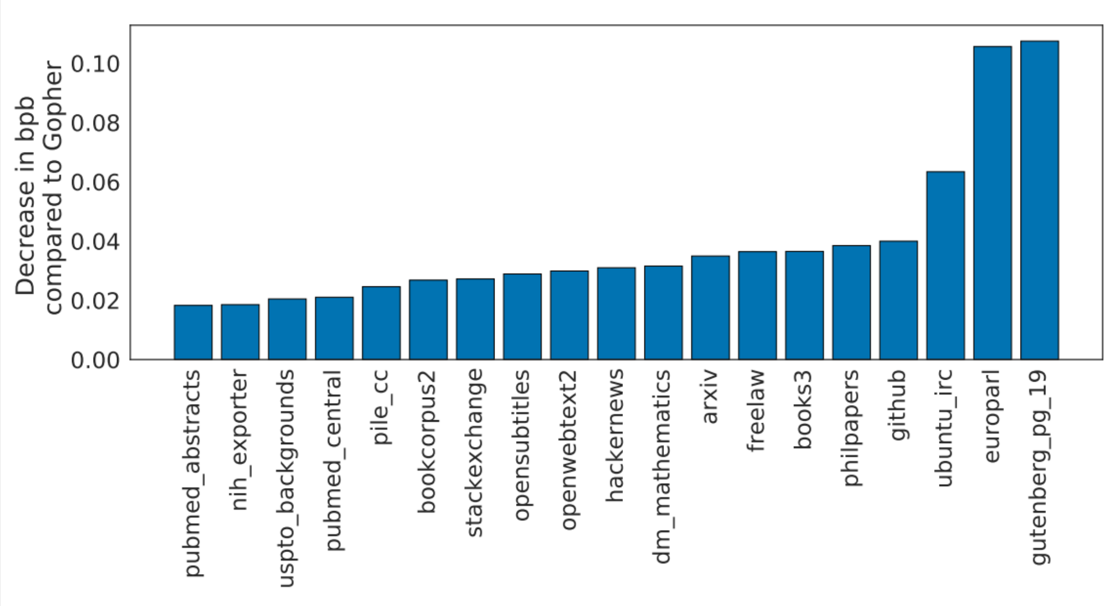
Compute-cost to train a LLM depends on 2 key choices:

- **Model-size:** Nr of trainable parameters
- **Data-size:** Nr of tokens processed during training

?

To train a larger LLM:
How much should we increase model-size vs data-size?

2) Train compute-optimal language model „Chinchilla“

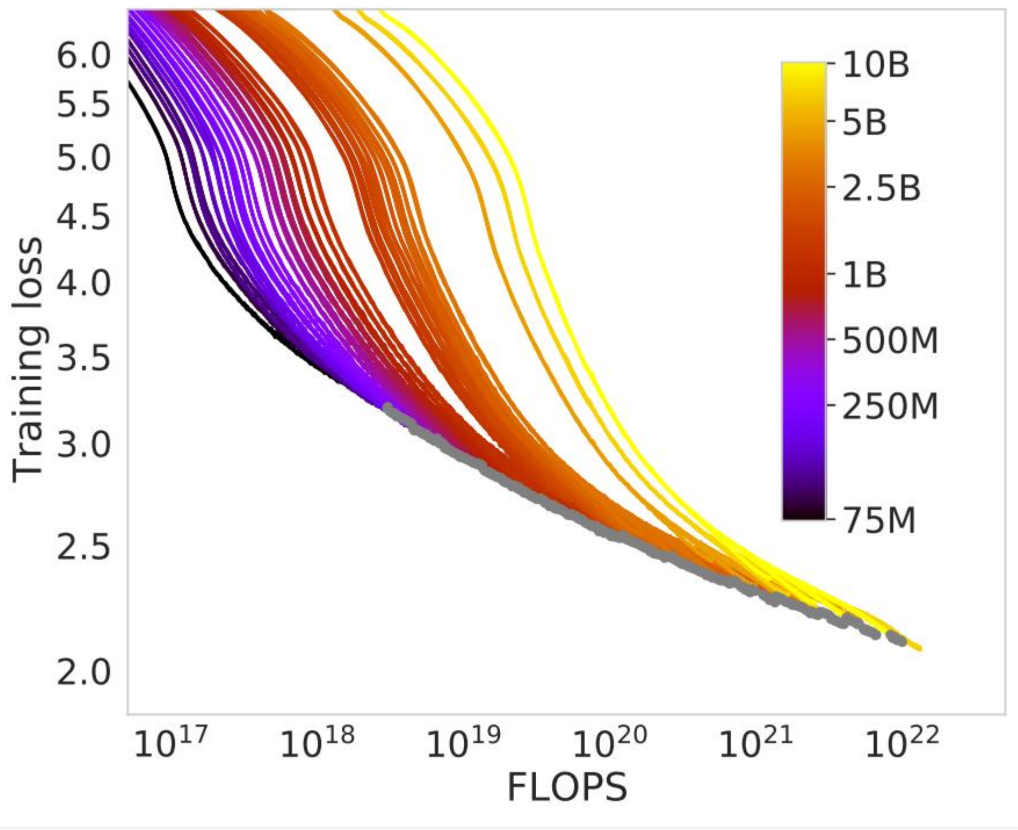


!

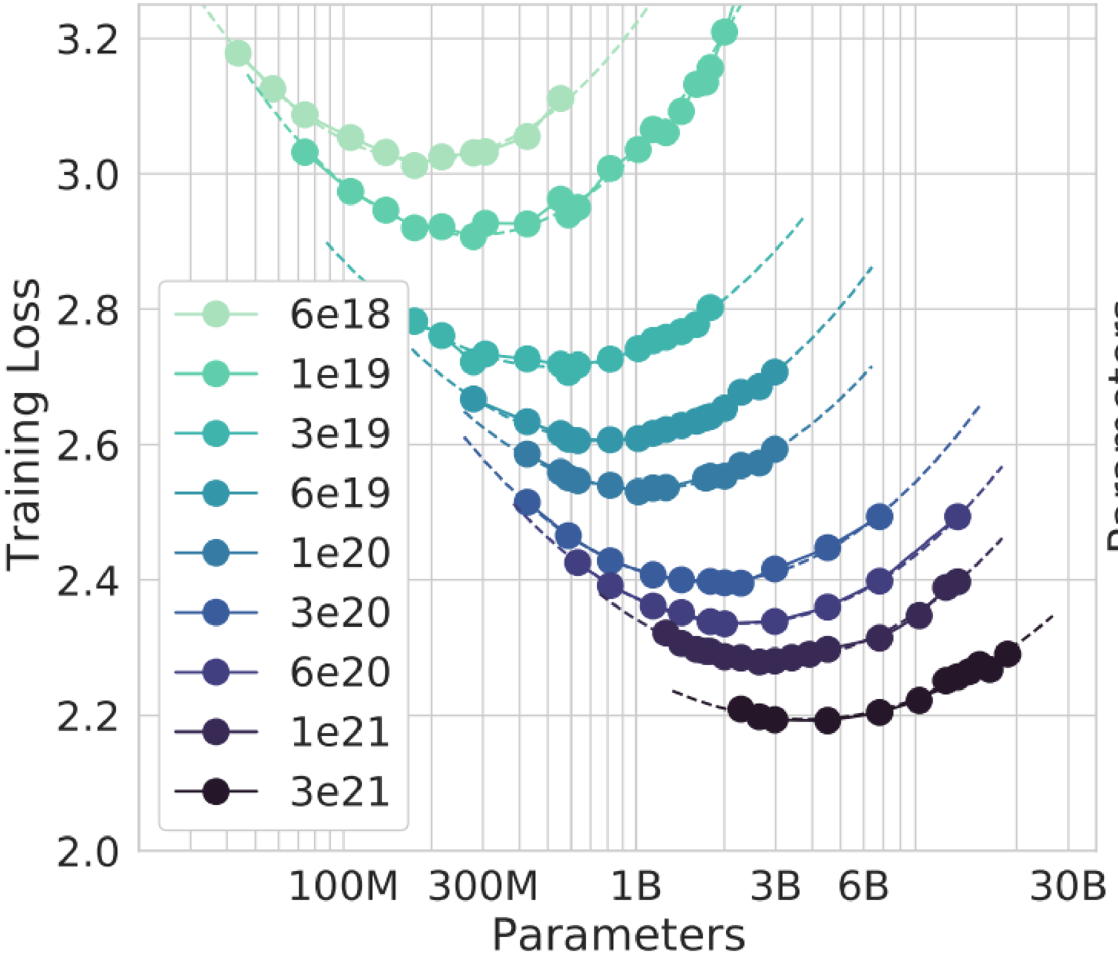
4x smaller model beats GPT3 and Gopher on all benchmarks

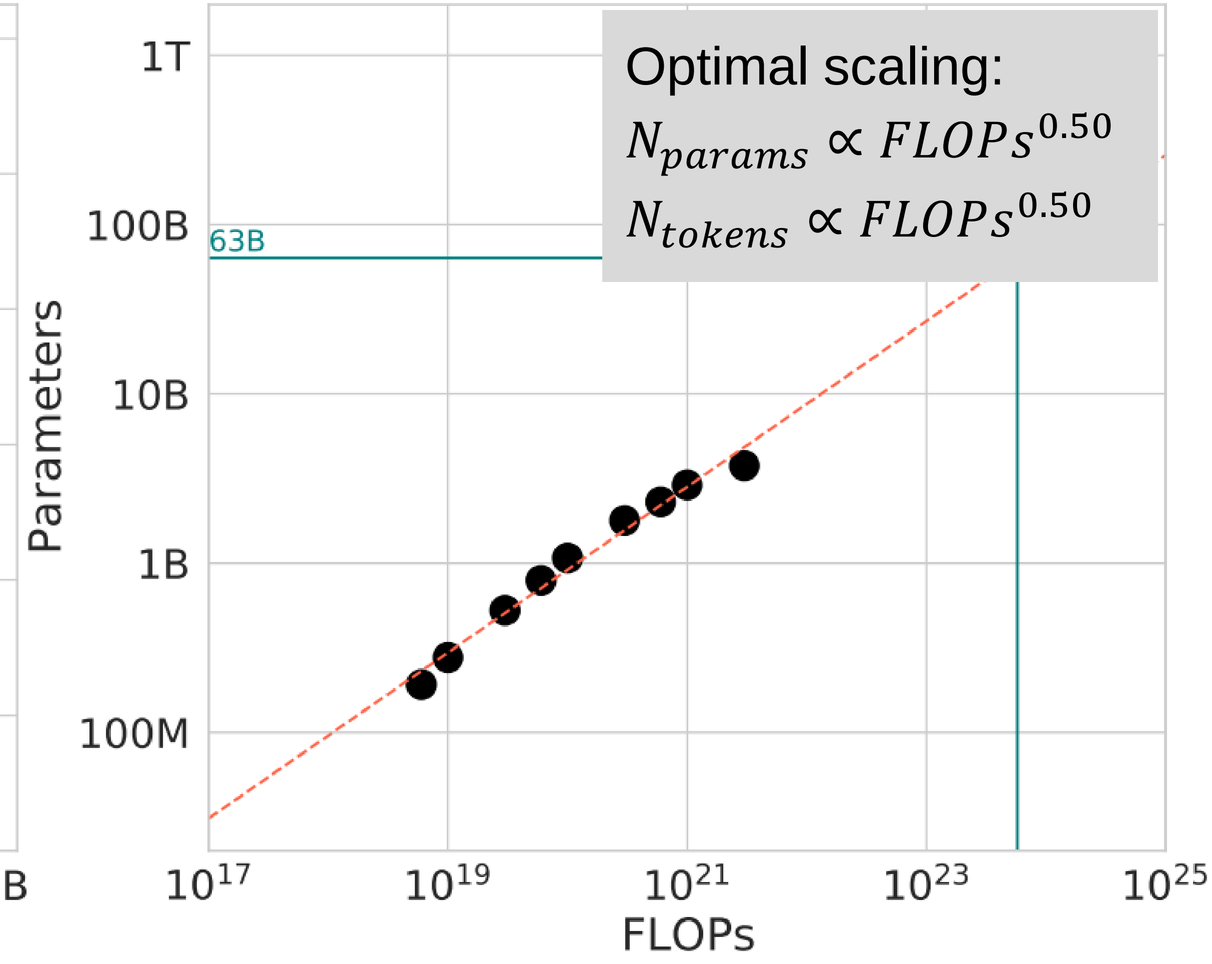
Some background on typical LLM (pre-)training

- **Single epoch:**
  - Each training sample is only used once
  - Training-loss is unbiased estimator of test-loss
  - Training-loss is a good proxy for downstream performance
- **Fixed training length:** Nr of training steps fixed in advance, due to cosine-LR-schedul
- **Compute scales linearly with parameters:** Parameters and compute dominated by K/Q/V-matmuls in attention layers

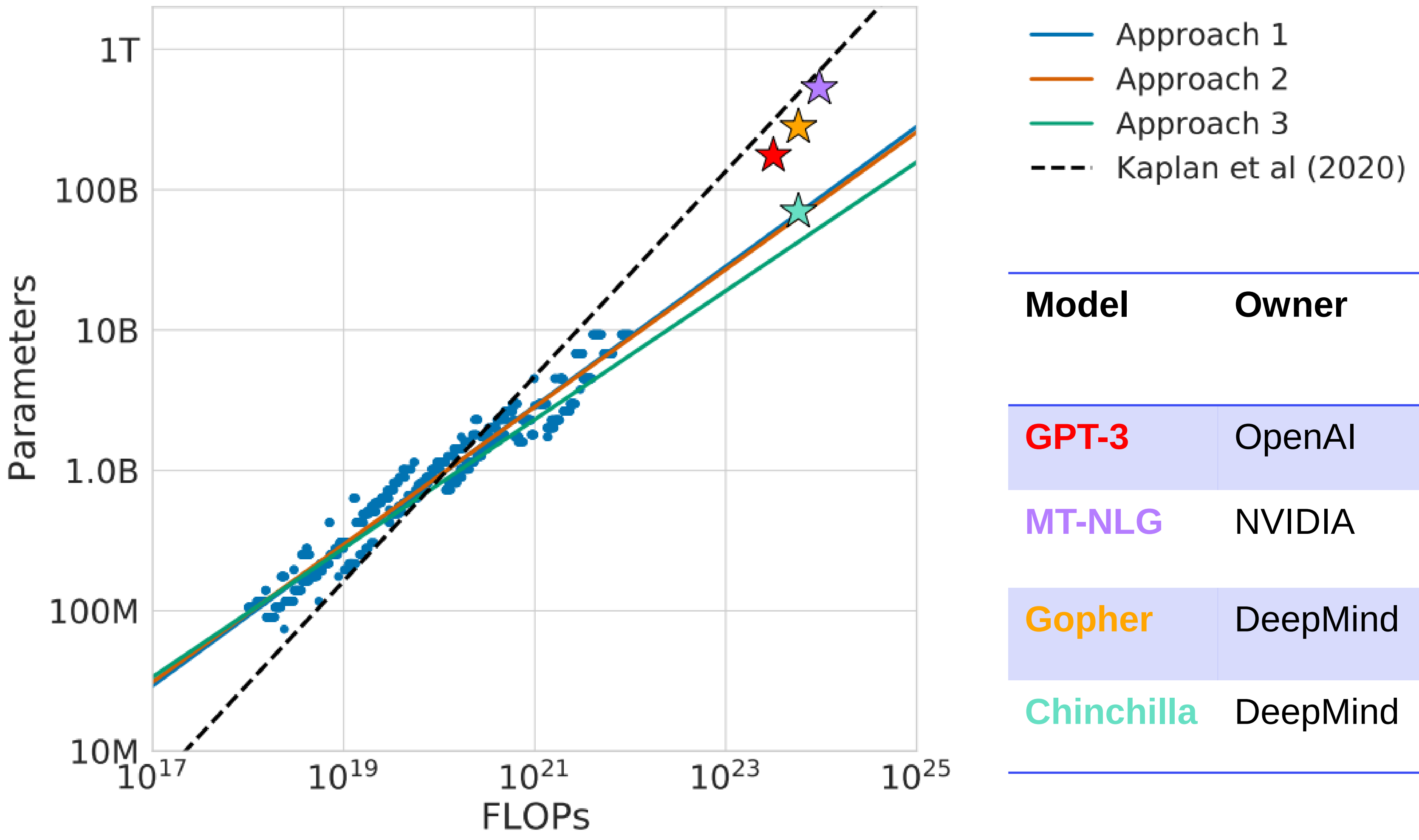


Methodology: Train a lot of models and find optimal model size as a function of compute-budget



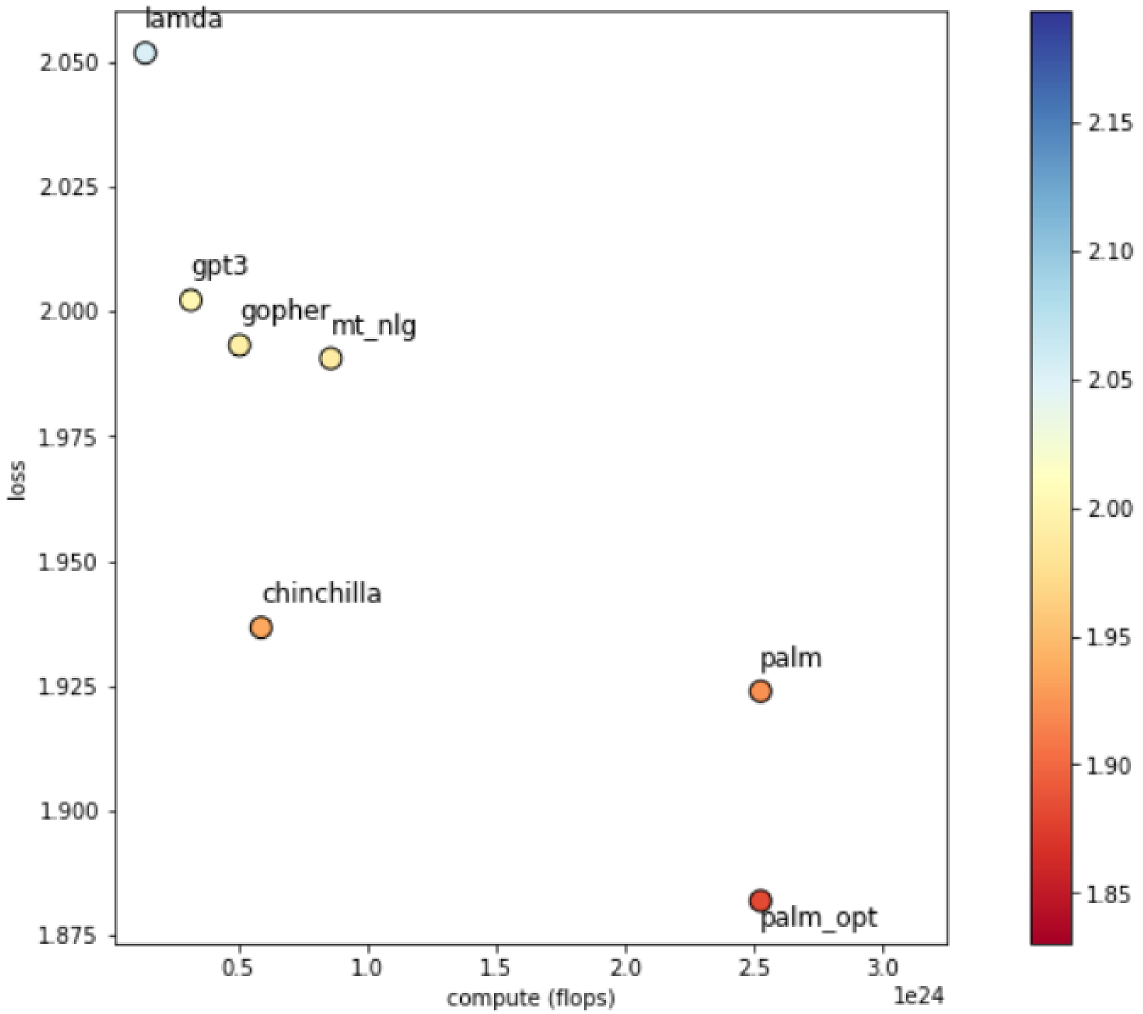


Current LLMs are too big and under-trained, because they followed a wrong scaling law (Kaplan et al 2020)



Model	Owner	Params (bn)	Tokens (bn)	MLLU Score
GPT-3	OpenAI	175	300	44%
MT-NLG	NVIDIA	530	270	
Gopher	DeepMind	280	300	60%
Chinchilla	DeepMind	70	1400	68%

PaLM could be substantially better, if it had been smaller and trained on more data



Scaling to trillions of parameters?

Max Welling
@wellingmax

In my 2018 keynote at ICML I showed this curve and predicted that in 2025 we would have models with **100 trillion parameters**. We might get there sooner...

Required resources for compute-optimal LLM

Params (bn)	FLOPs vs. Gopher	Tokens (trillions)
70	1x	1.4
175	7x	4
520	59x	11
1,000	221x	21
10,000	22,515x	216
100,000	225,159x	2162

How much text is there actually?

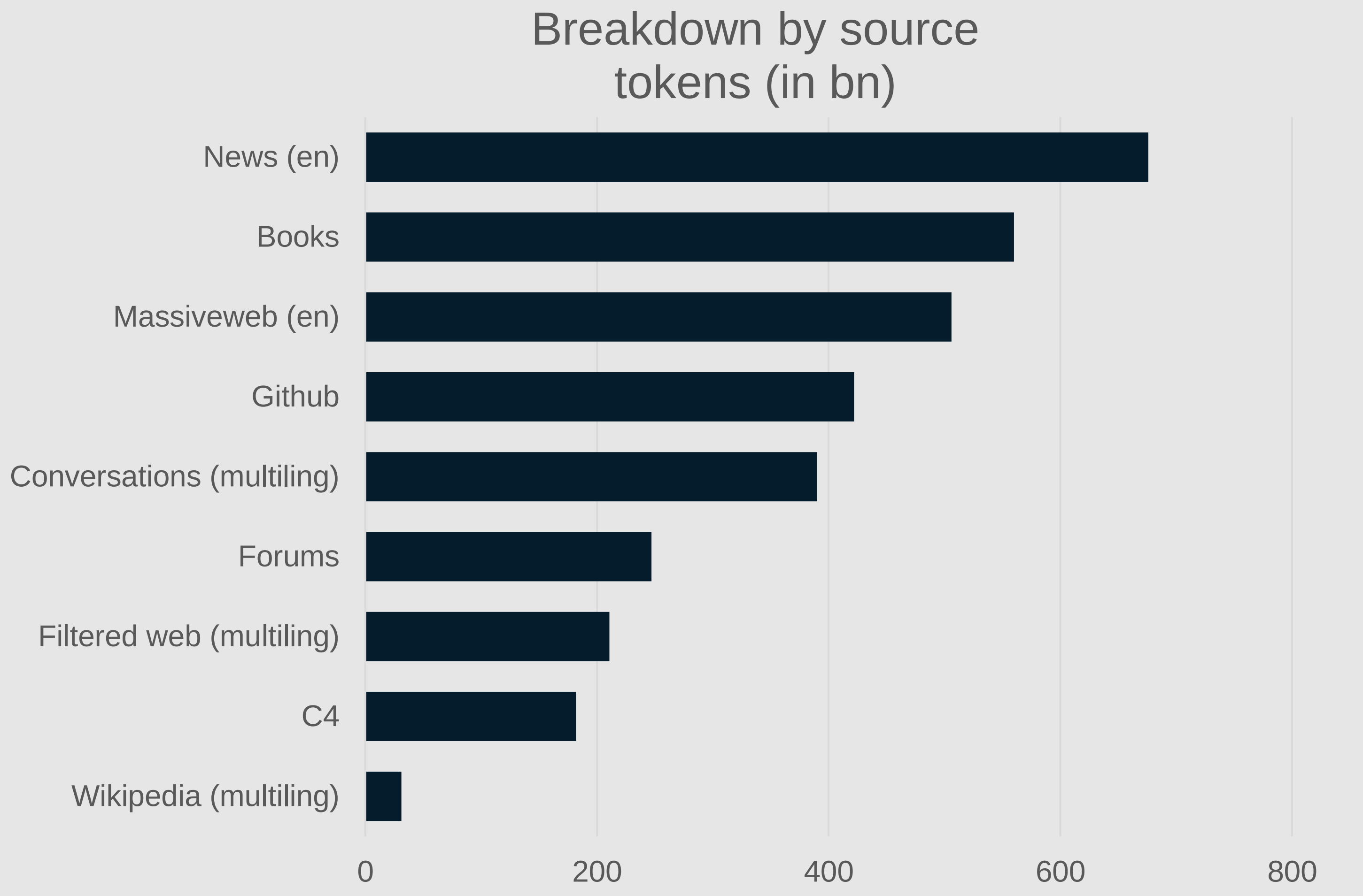
Order-of-magnitude estimations:

2 tn

tokens in MassiveText

3.2 tn

tokens in high-quality data



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