Chinchilla scaling law:

L = LM test loss (avg. cross-entropy loss)

D = dataset size (# training tokens)

N = # model parameters (FF layers, Self-attn, etc)

C = compute budget, FLOPS(N, D), deterministic fn of data size and model size

Given a fixed FLOPS budget C,

find

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

Loss w/ perfect LM

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

after training a lot of diff models and fitting this eqn, we get

$$\alpha = 0.34, \ \beta = 0.28, \ E = 1.69$$
$$A \sim B \sim 400$$

Two models; same compute C

- Gopher (280B params, 300B tokens)
- Chinchilla (70B params, 1.4T tokens)

$L(\text{Gopher}) = 1.993$

$L(\text{Chinchilla}) = 1.936$

to put in perspective, according to these scaling laws, with this compute budget C, no model trained w/ 300B tokens could ever be better than Chinchilla!