Chinchilla scaling law:

\[ L = \text{LM test loss (avg. cross-entropy loss)} \]
\[ D = \text{dataset size (# training tokens)} \]
\[ N = \text{# model parameters (PE layers, Self-attn, etc.)} \]
\[ C = \text{compute budget, deterministic} \]
\[ \text{FLOPS}(N, D) \]

Given a fixed FLOPS budget \( C \), find

\[ \arg\min_{N, D} L(N, D) \]
\[ \text{s.t. FLOPS}(N, D) = C \]

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

Loss w/ perfect LM
g) contribution of model size
g) contribution of data
after training a lot of other 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 (708B 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!