

Scaling laws

Marco Kuhlmann

Department of Computer and Information Science

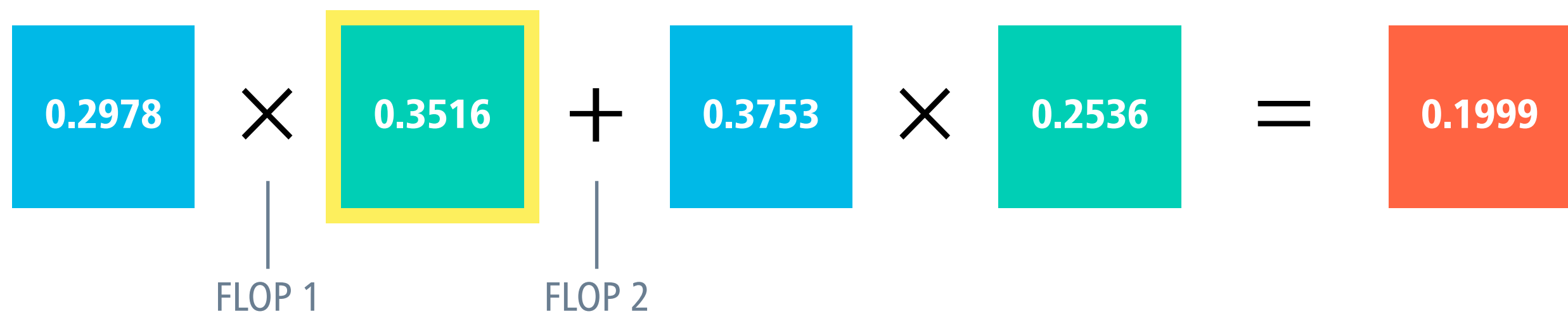
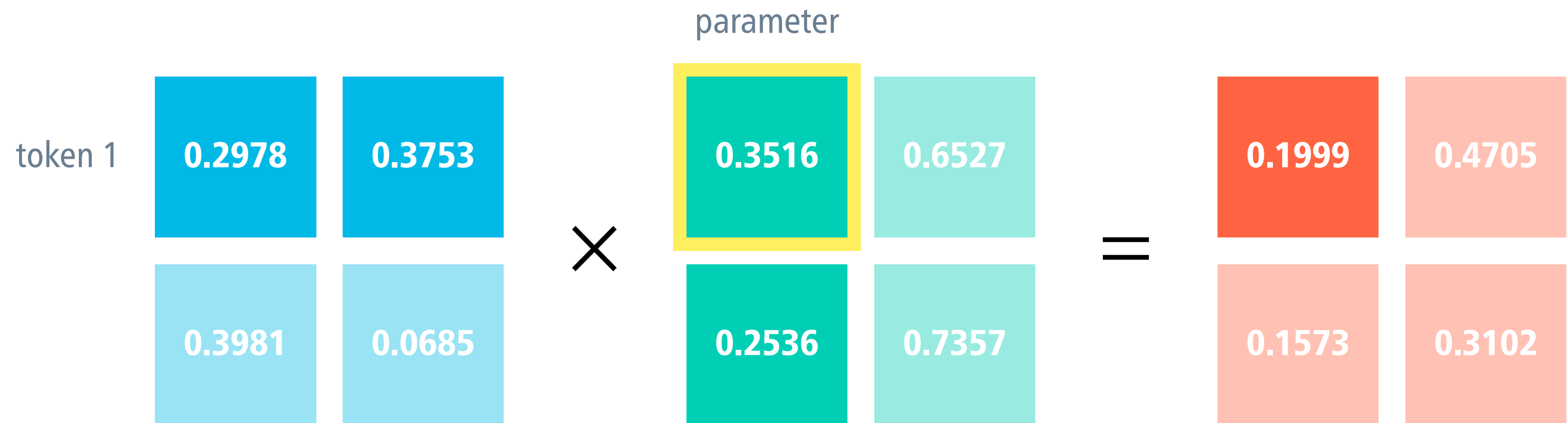
Scaling laws in language modelling

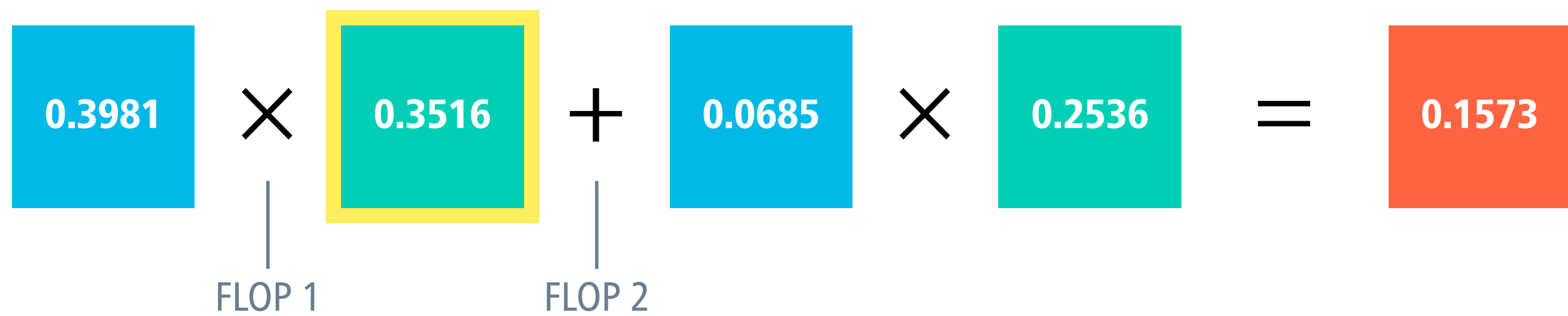
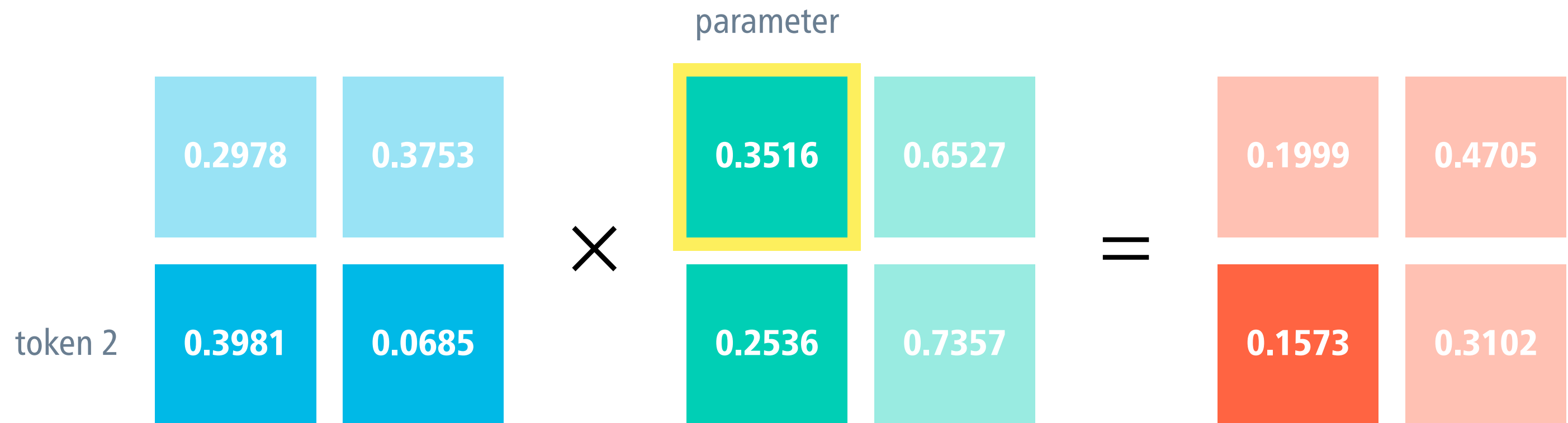
- **Scaling laws** describe how model performance improves as we increase key factors such as model size and training data size.
- Empirical results suggest that performance improvements obey a power law: performance increases, but at a diminishing rate.
cf. Heap's law
- Scaling laws can help developers answer many practically relevant questions about resource allocation.

Computational cost

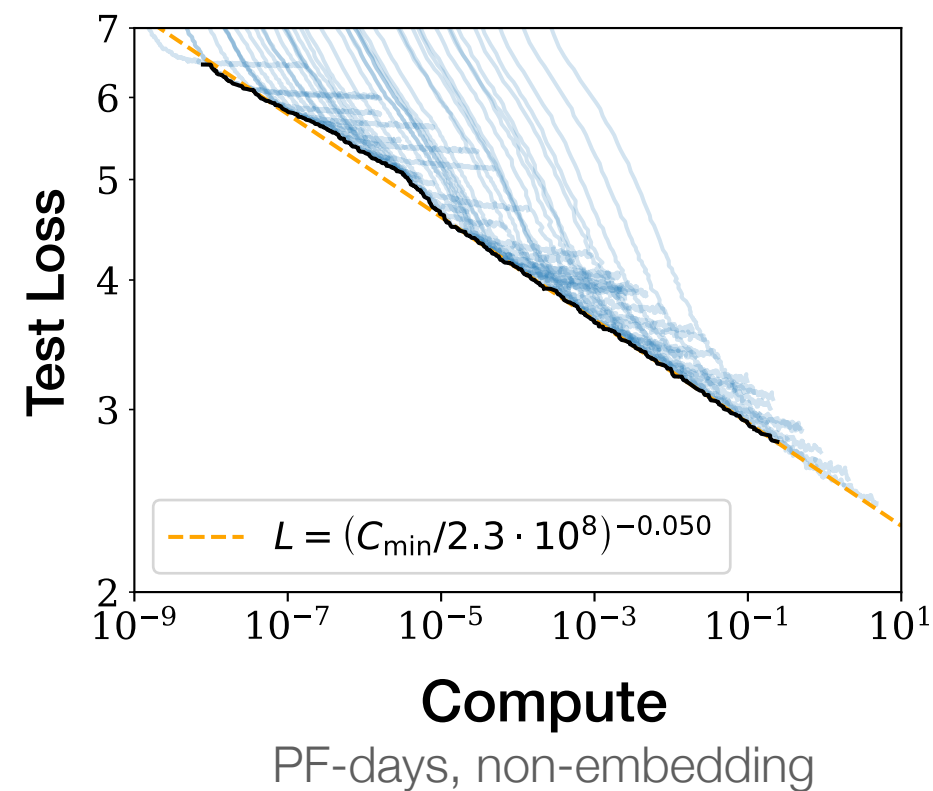
- The cost of language model training is a function of the number of model parameters, P , and the number of training tokens, T .
- The standard unit for measuring computational cost is the number of **floating point operations (FLOPs)**.
- For the Transformer architecture, a useful approximation for the computational cost C is $C \cong 6PT$.



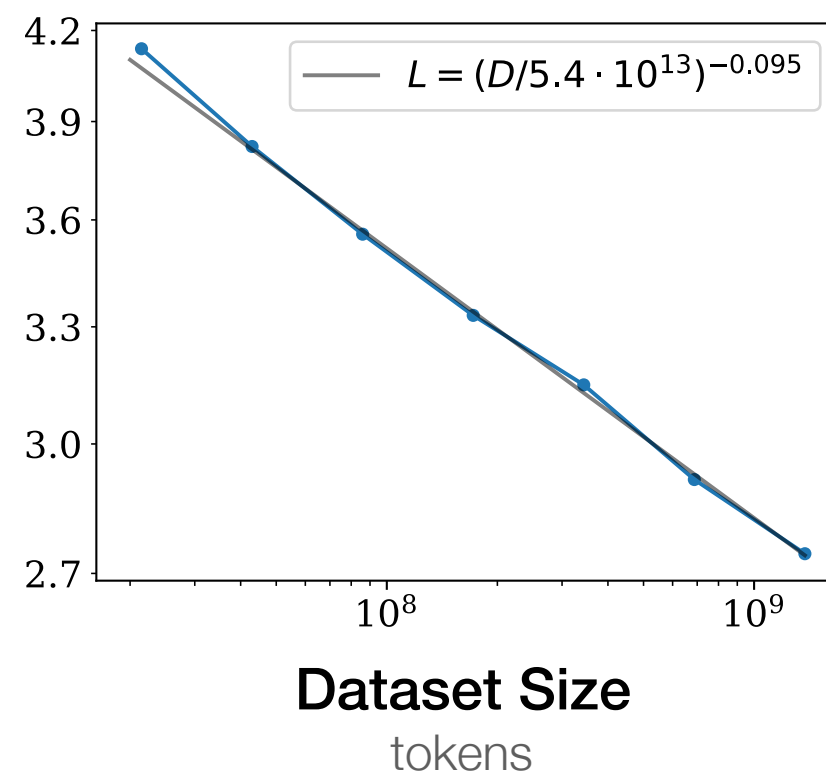




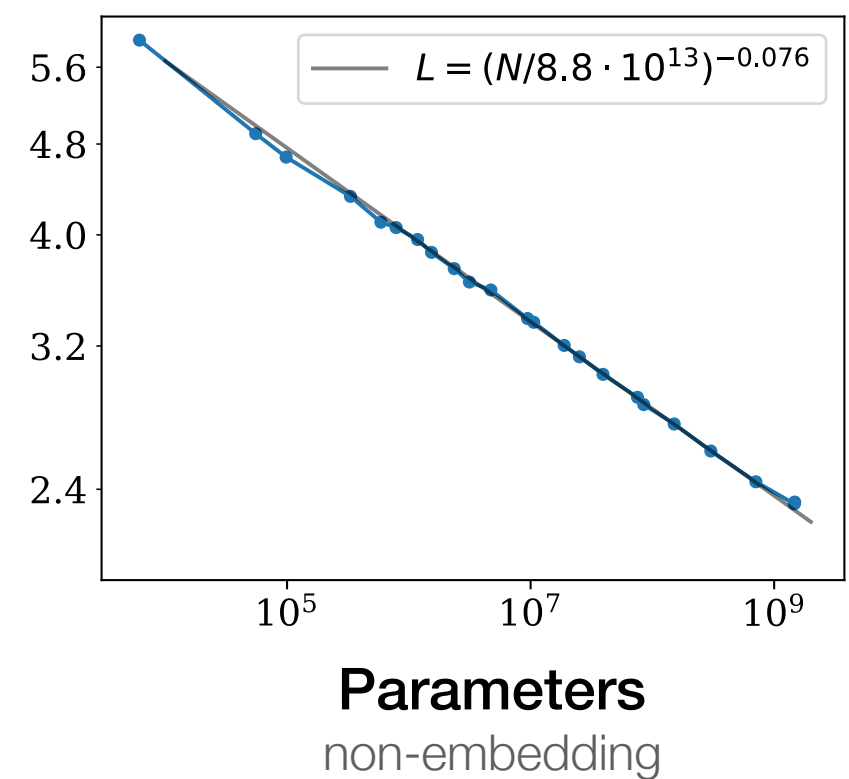
Performance improves smoothly with scale



Performance improves smoothly as we increase **compute**



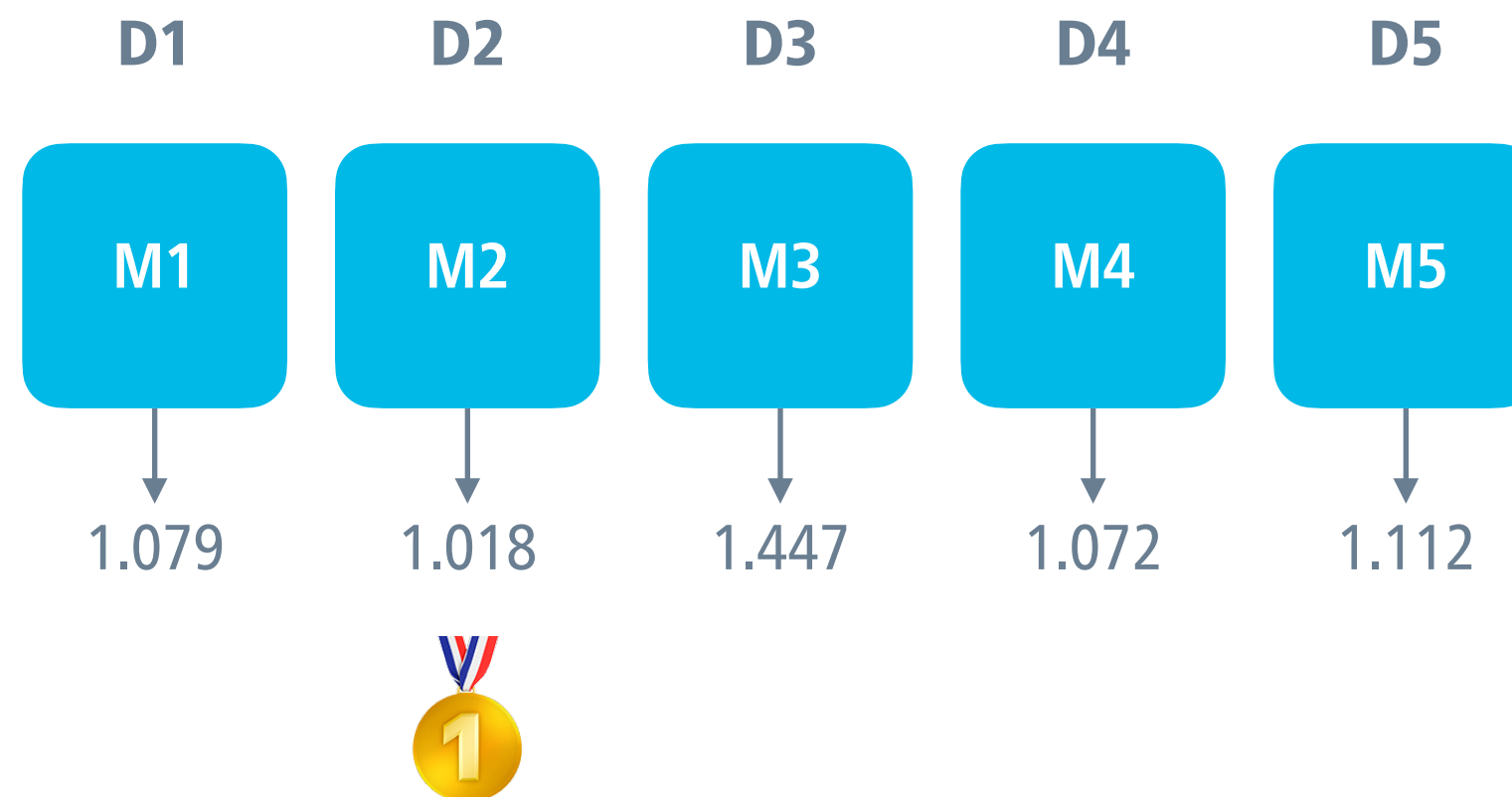
Performance improves smoothly as we increase **dataset size**



Performance improves smoothly as we increase **model size**

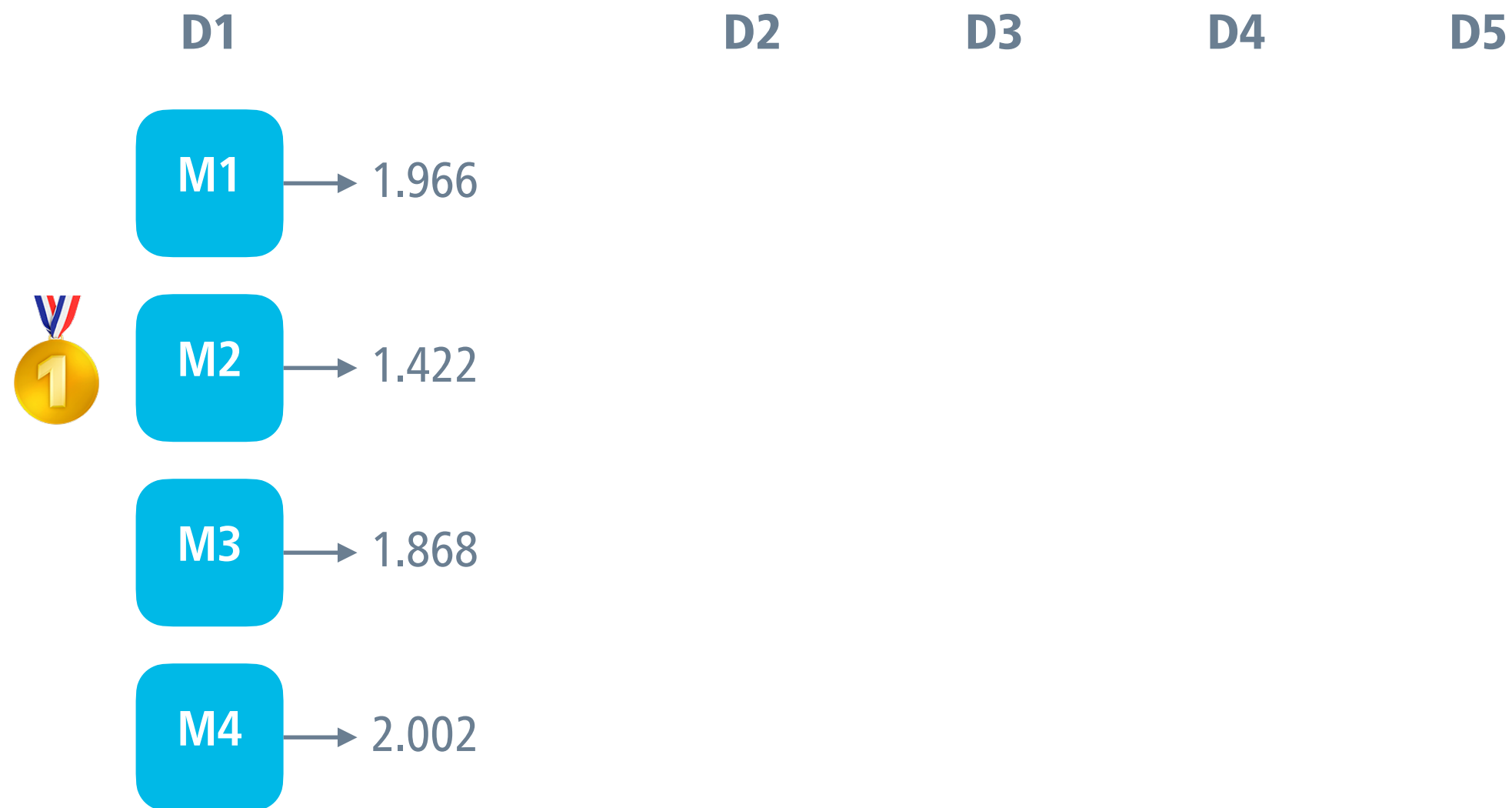
Kaplan et al. (2020)

Putting scaling laws into practice



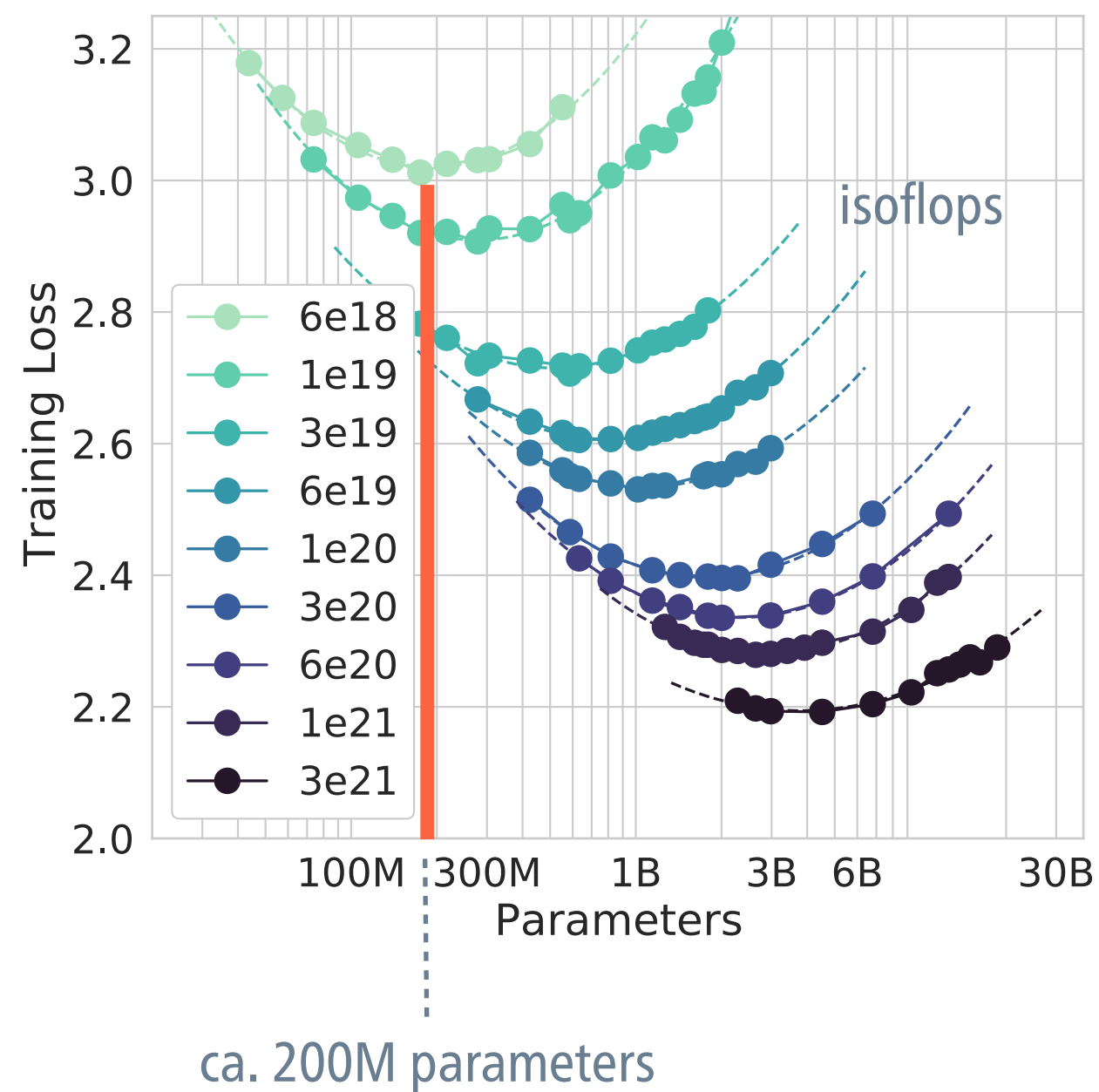
old paradigm: train a few models, select the best one

Putting scaling laws into practice

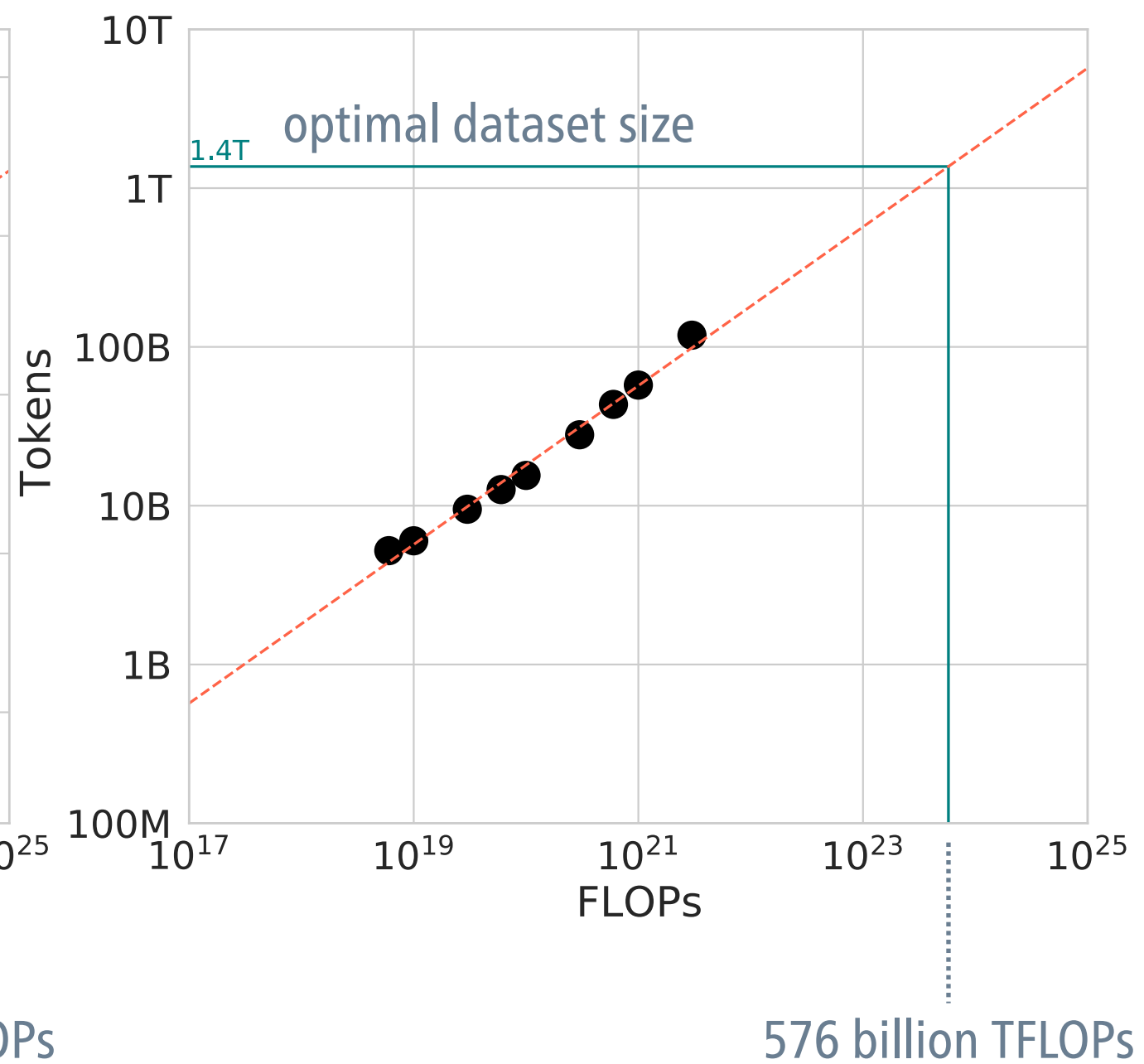
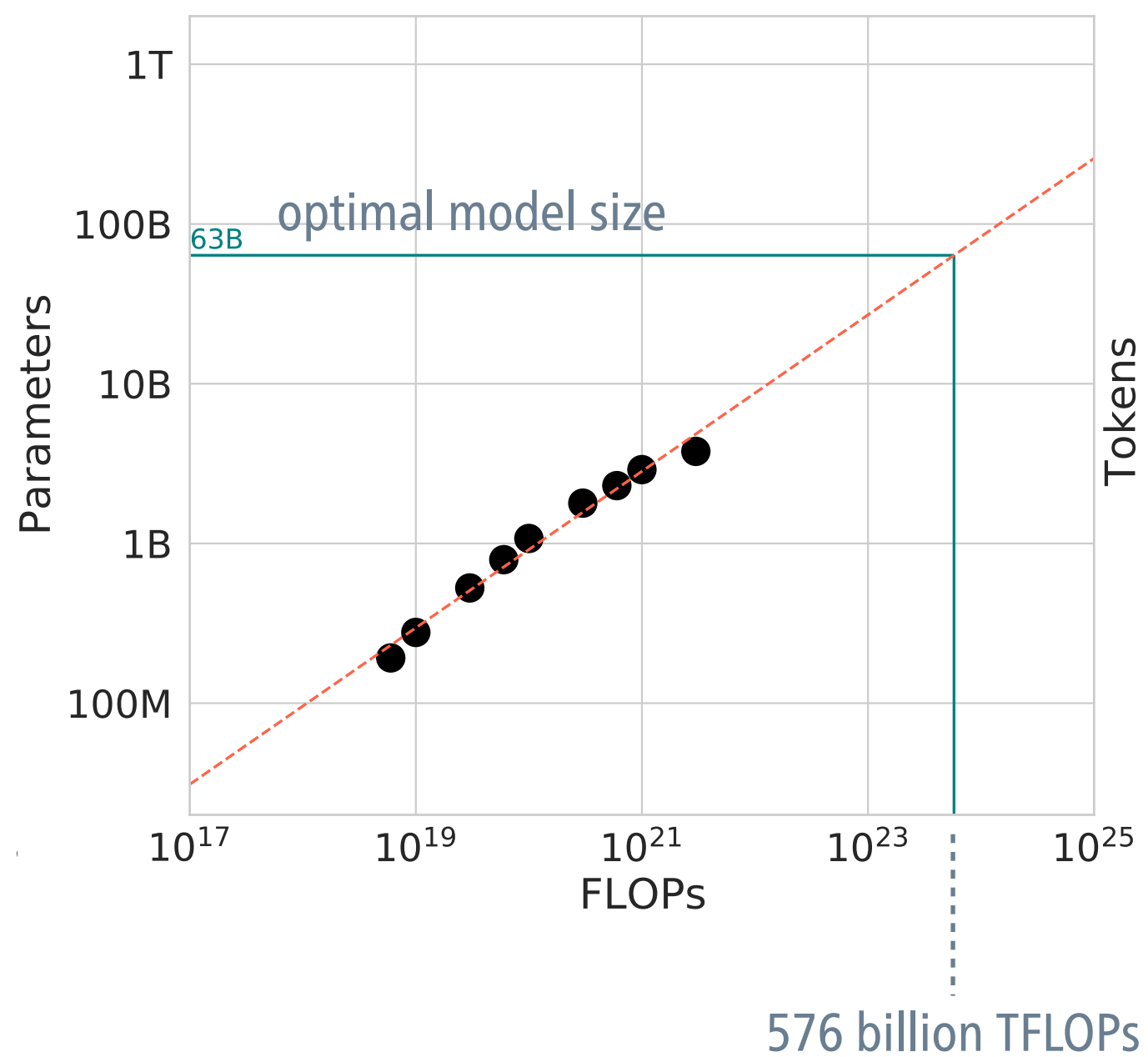


new paradigm: train many small models, up-scale the best one

Compute-optimal models



Compute-optimal models



Large language models can be too large

