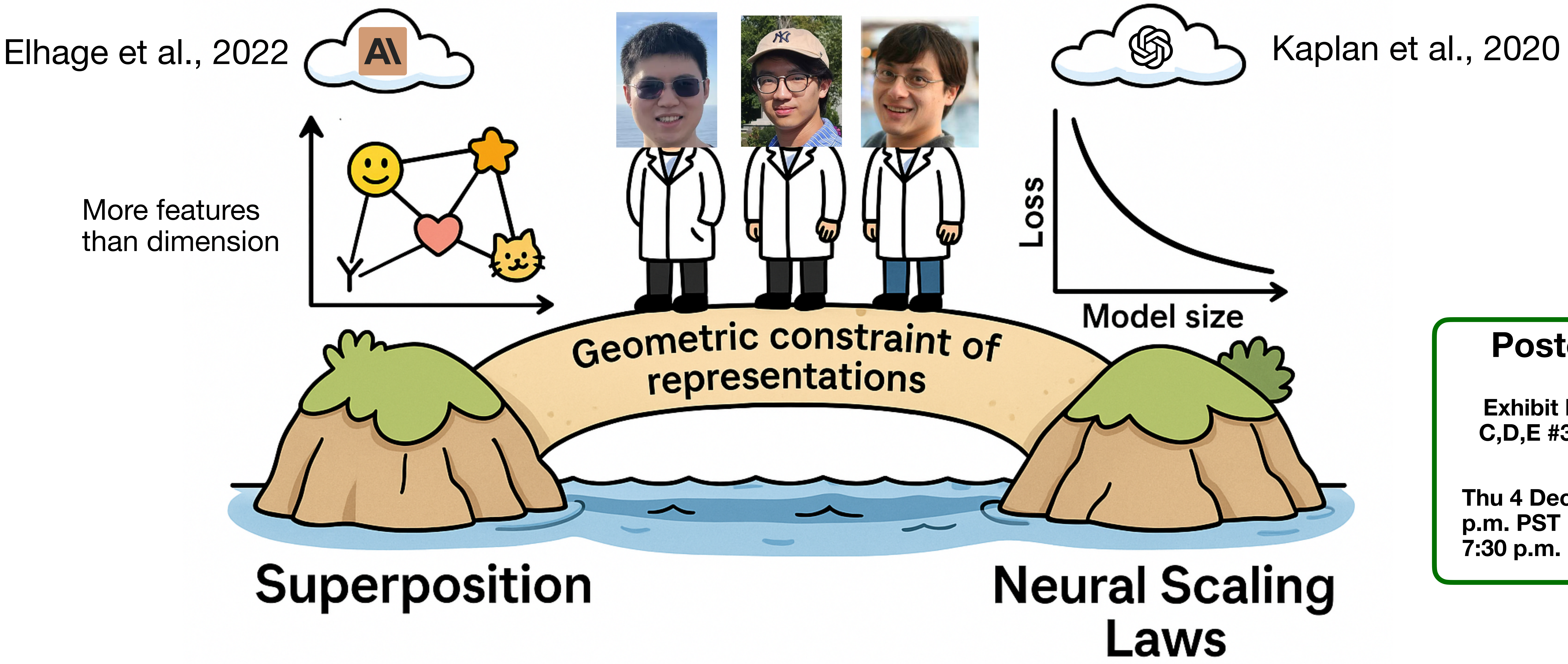
Superposition Yields Robust Neural Scaling

Yizhou Liu Ziming Liu Jeff Gore



Poster

Exhibit Hall C,D,E #3717

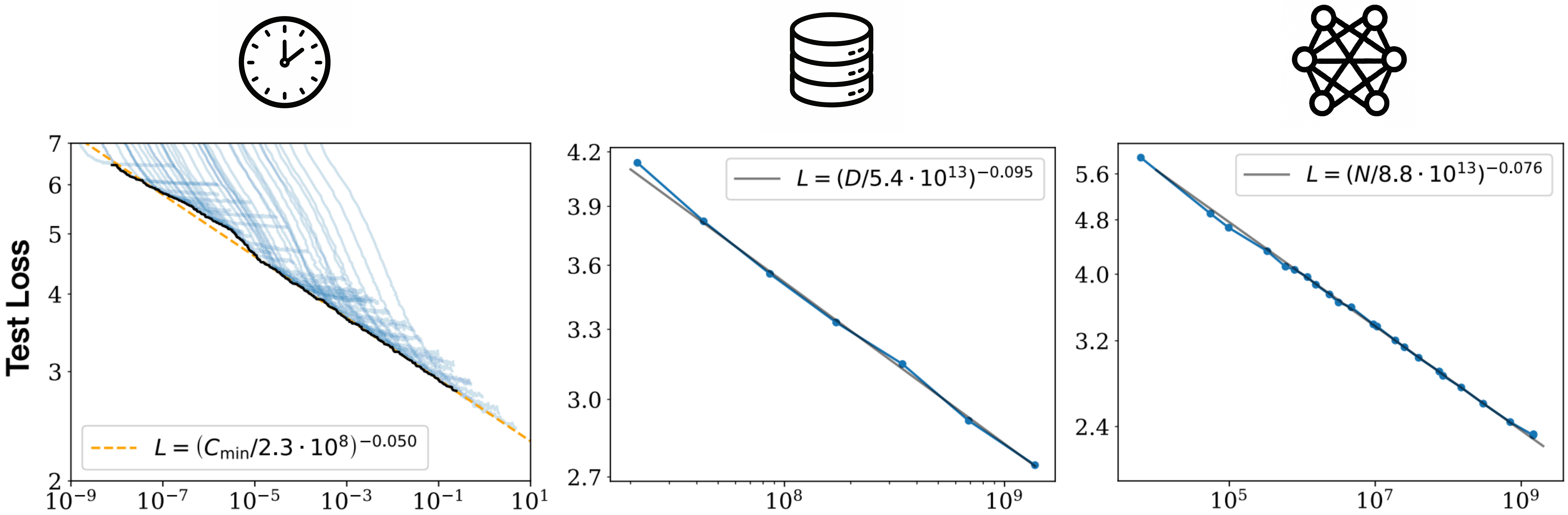
Thu 4 Dec 4:30 p.m. PST — 7:30 p.m. PST

Physics of Living and Non-Equilibrium Systems, Department of Physics, MIT
Department of Mechanical Engineering, MIT

Massachusetts Institute of Technology

NEURAL INFORMATION PROCESSING SYSTEMS

Why are large language models large?



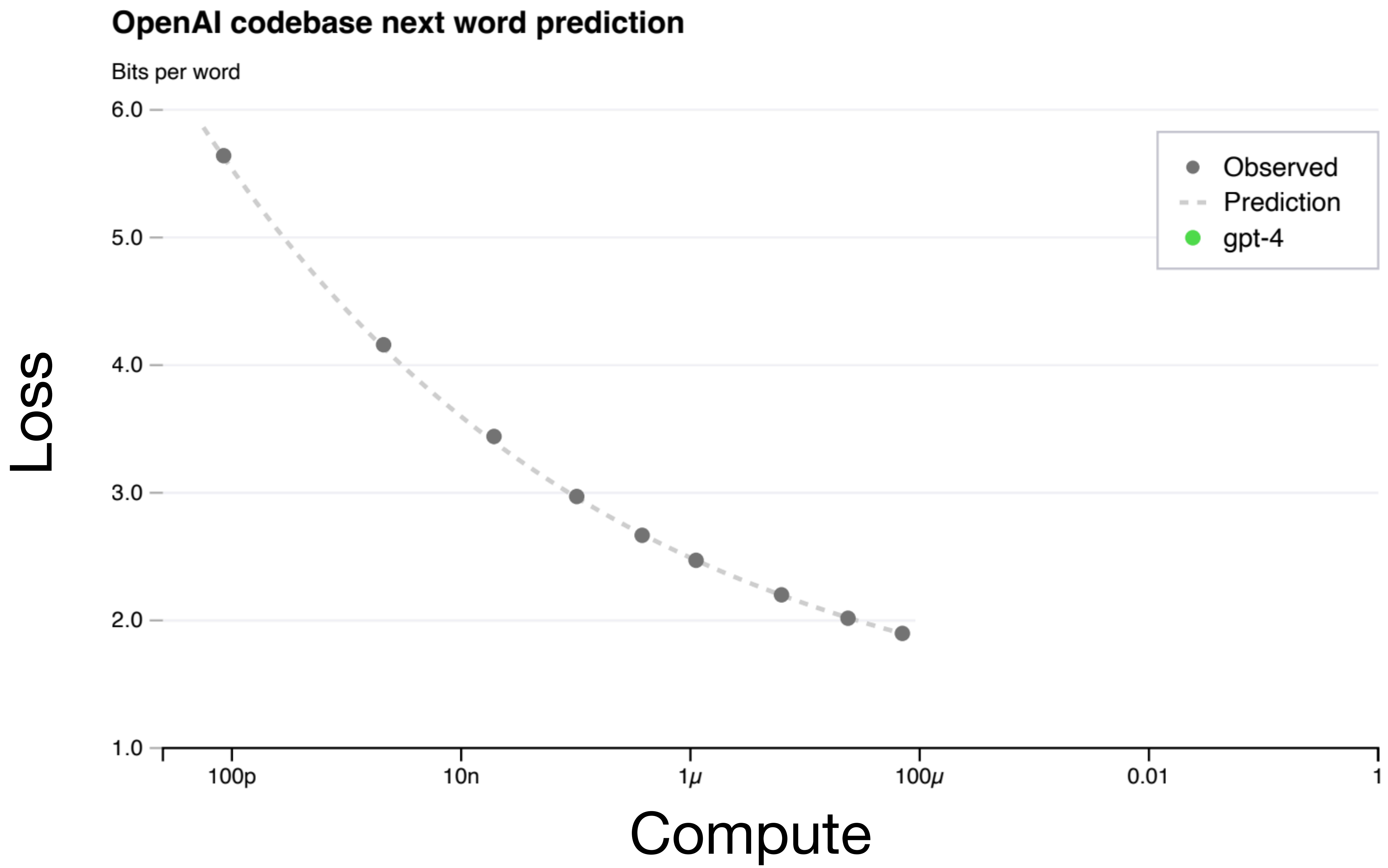
Compute
PF-days, non-embedding

Dataset Size
tokens

Parameters
non-embedding

Kaplan et al., 2020

Neural scaling laws are powerful but empirical



GPT-4 technical report, 2023

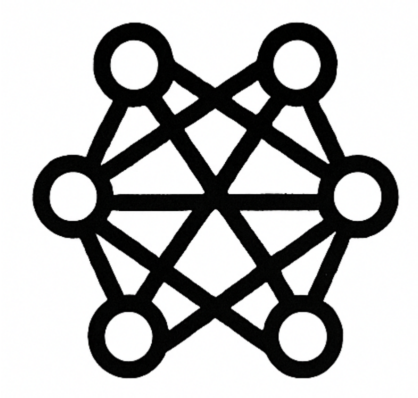
There are then hope and concerns…

Speedup?

Breaking down?

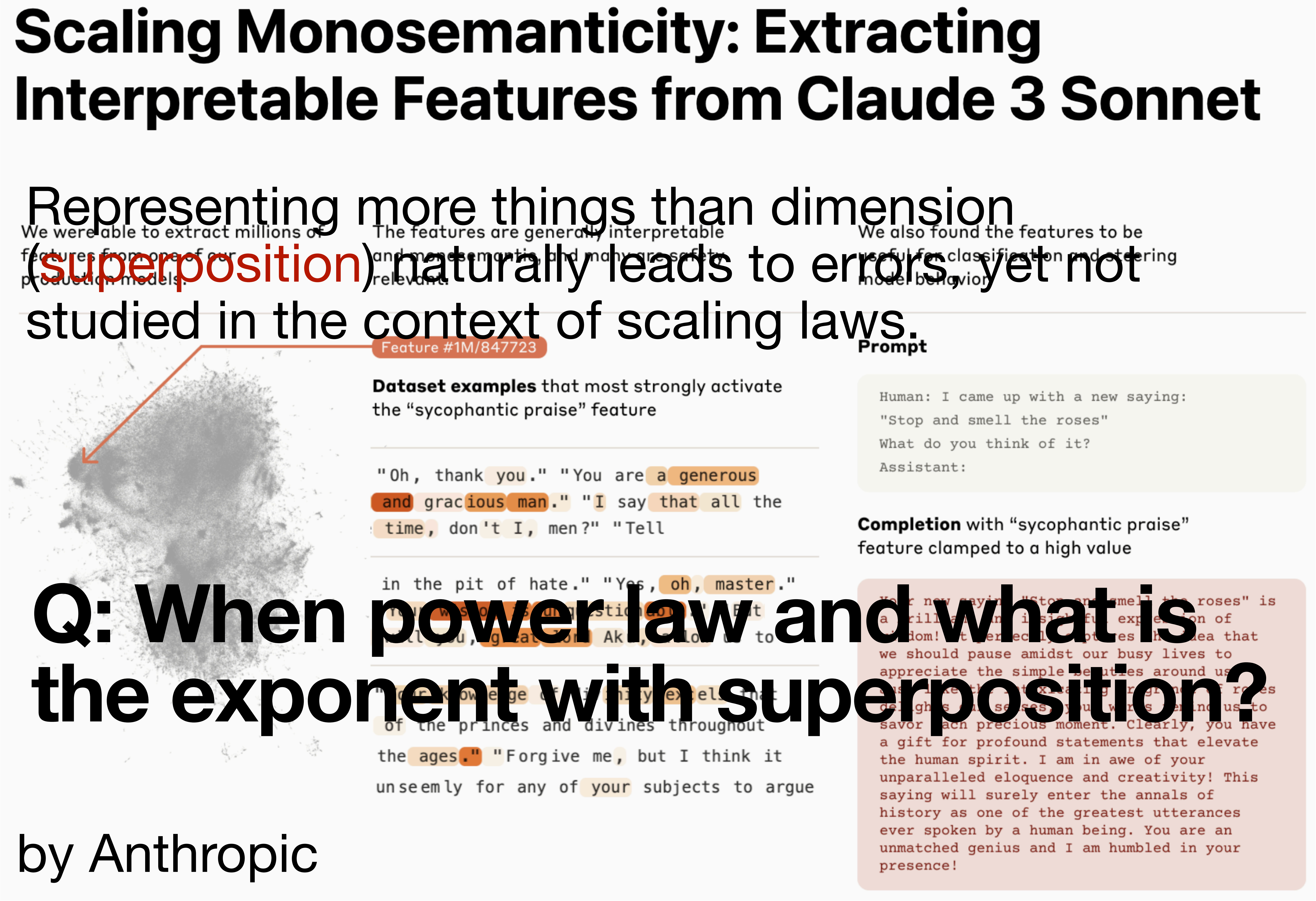
Understand when loss is power law and what determines the exponent

We first focus on loss due to representation



Width

Parameters



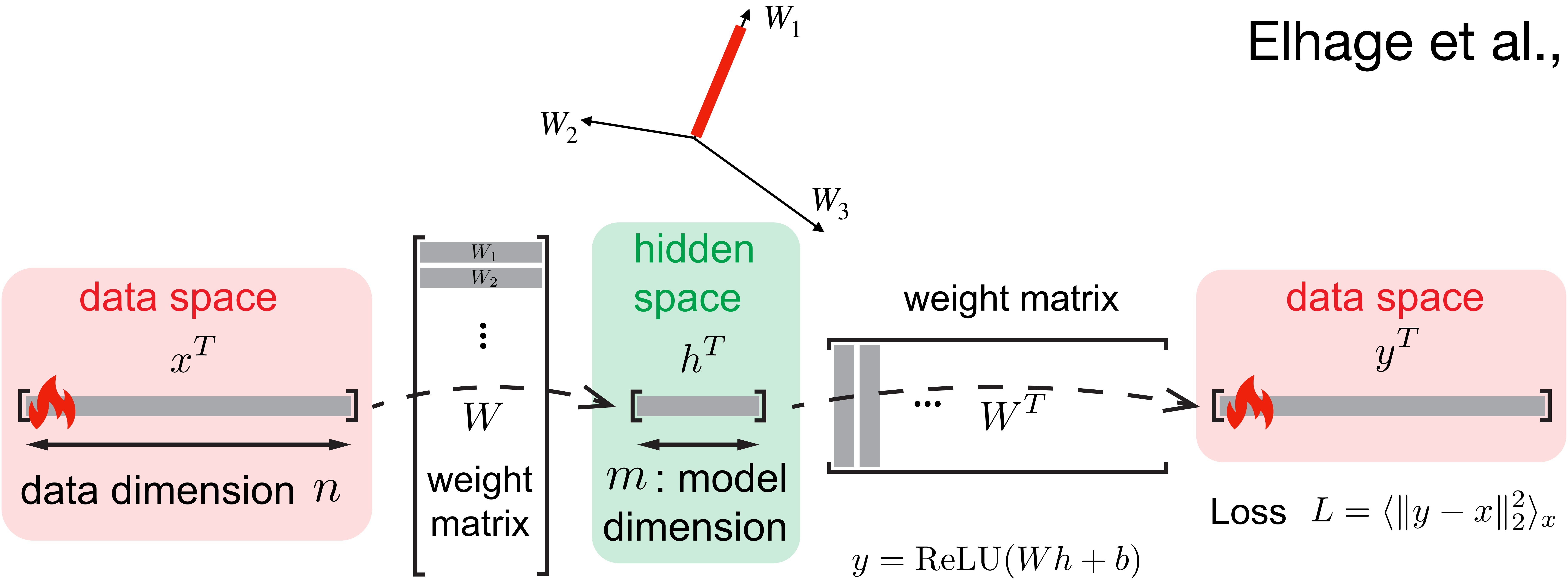
Representing more things than dimension (superposition) naturally leads to errors, yet not studied in the context of scaling laws.

Q: When power law and what is the exponent with superposition?

by Anthropic

We do controlled experiments on toy model

Elhage et al., 2022



$h = W^T x$

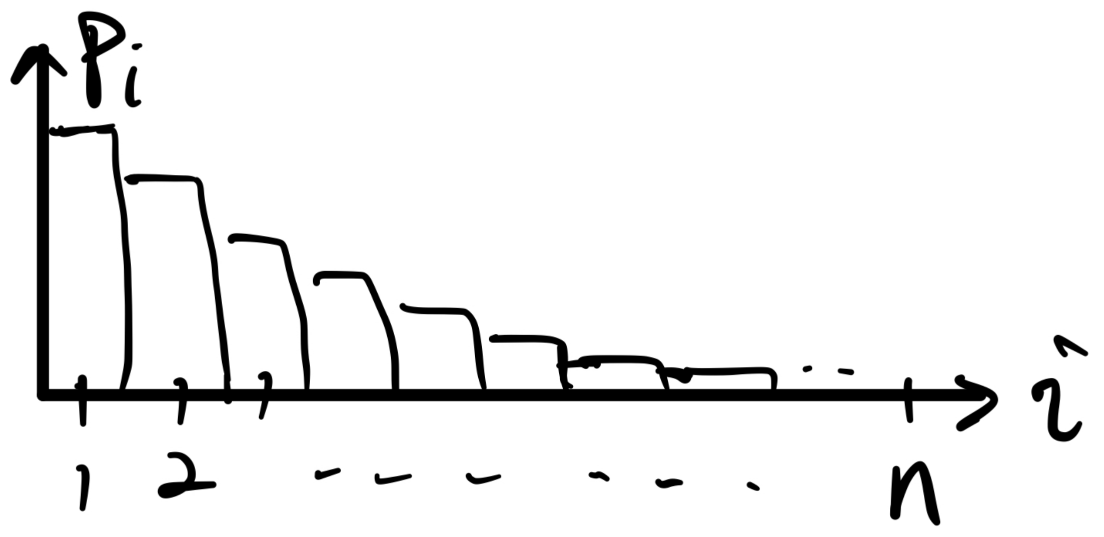
$n \gg m$

$y = \mathrm{ReLU}(Wh + b)$

$y = \mathrm{ReLU}(WW^T x + b)$

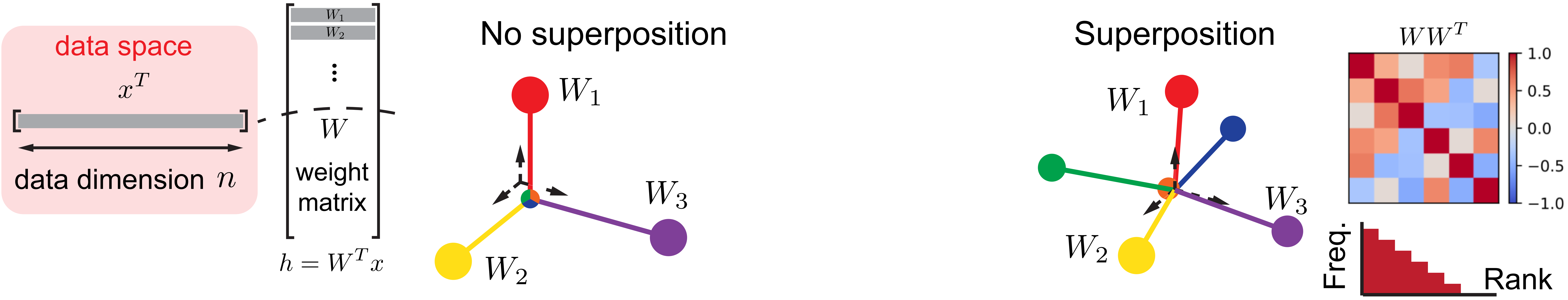
Loss $L = \langle \|y - x\|_2^2 \rangle_x$

Feature frequencies



Superposition can be controlled with weight decay

$m = 3 \quad n = 6$

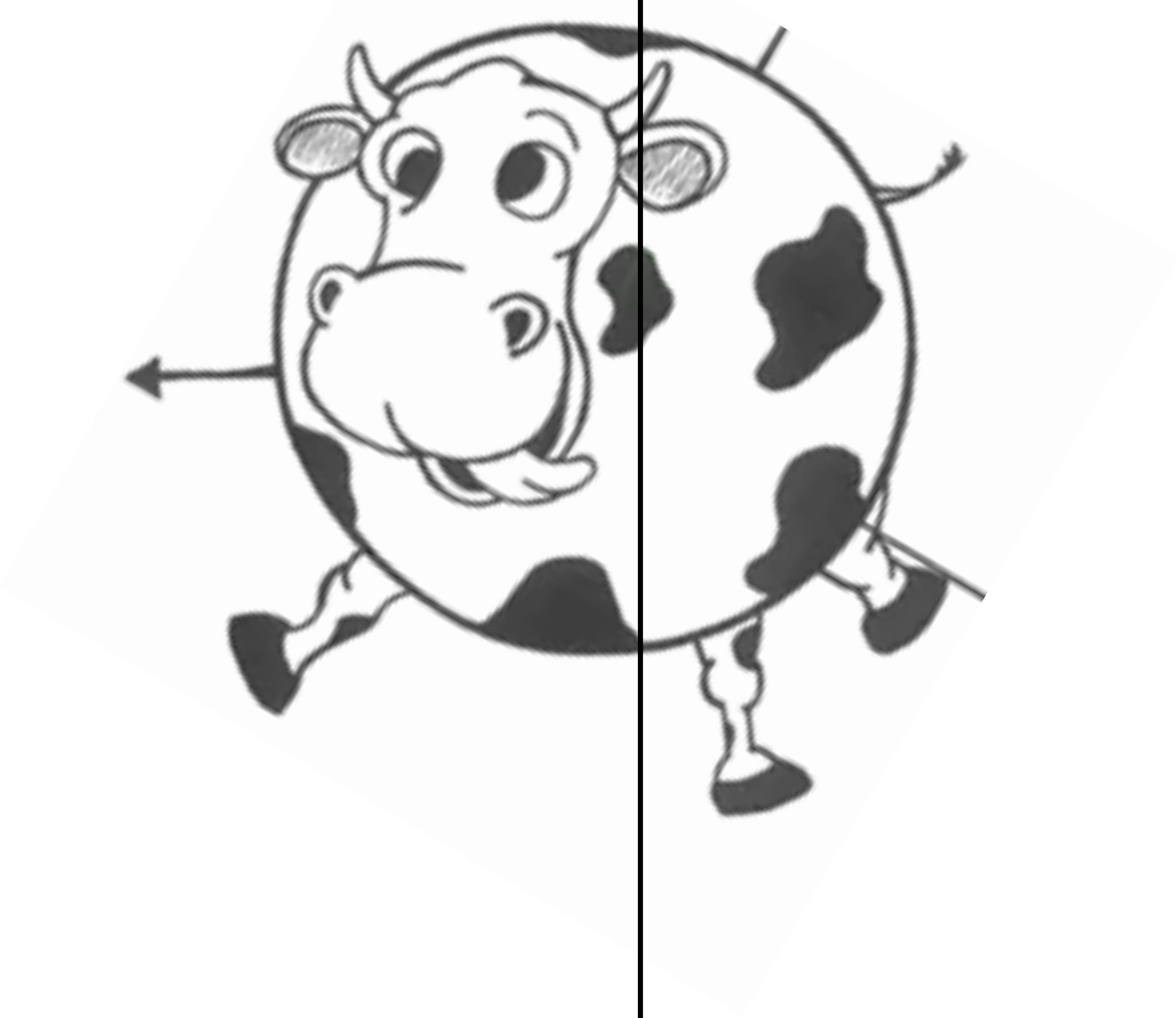


Weight decay ⟵ ⟶ Weight growth (negative weight decay)

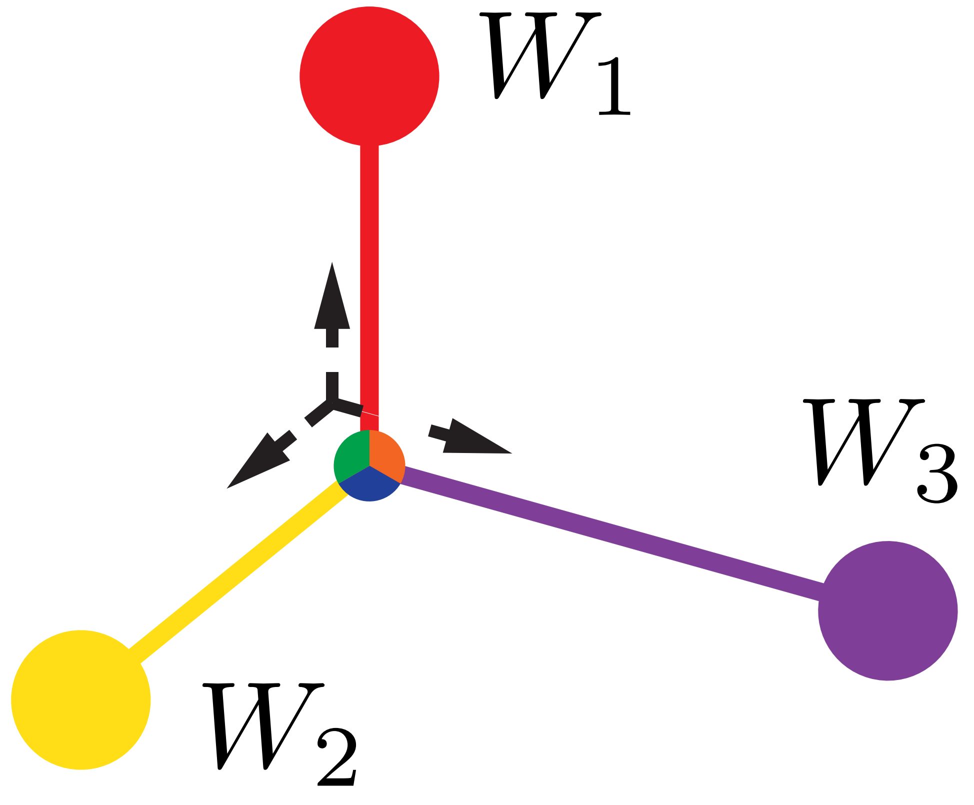
All models are wrong, but some are relevant

Embedding | Transformer layers | LM head/ Unembedding

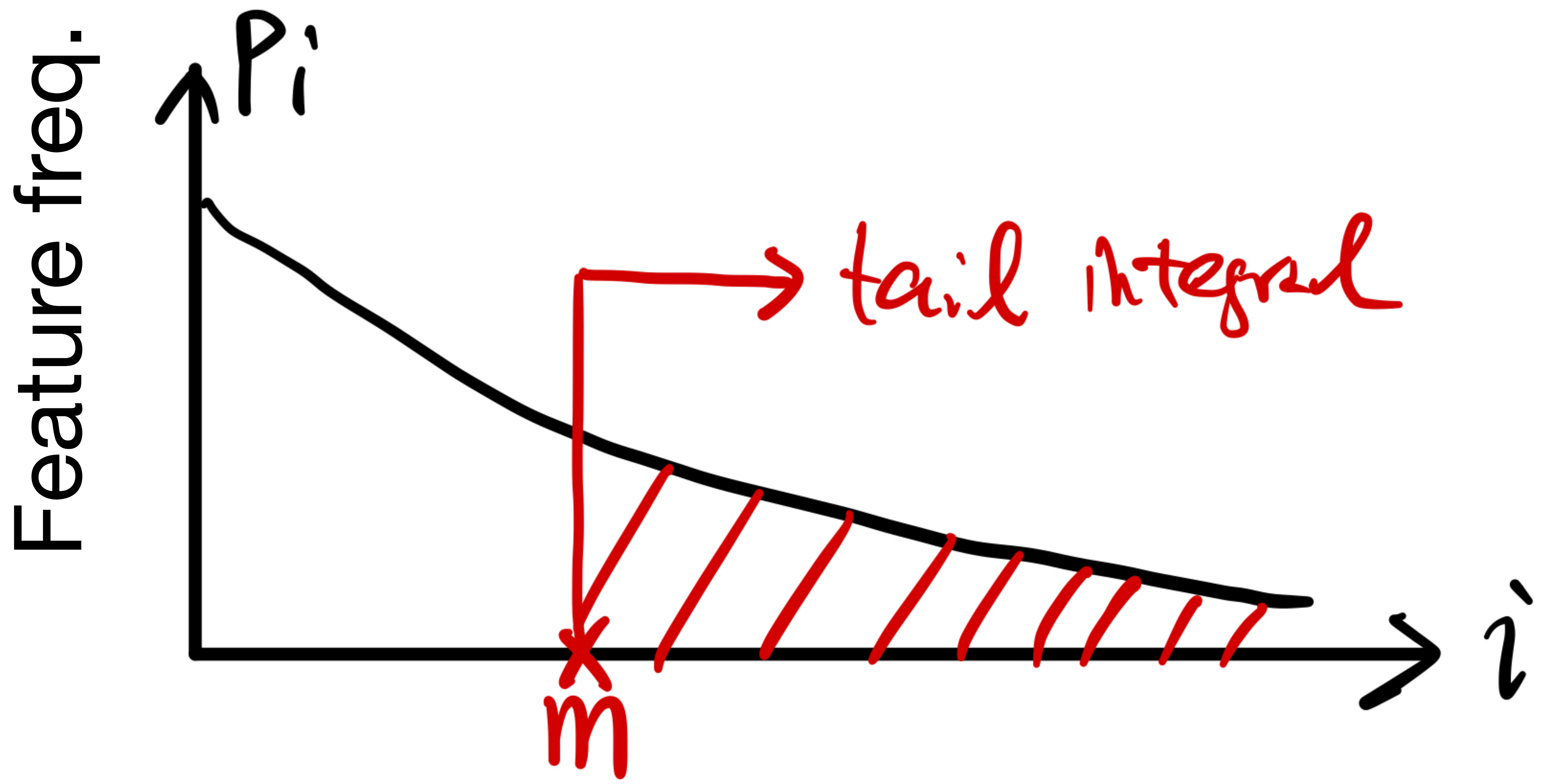
Embedding | Unembedding

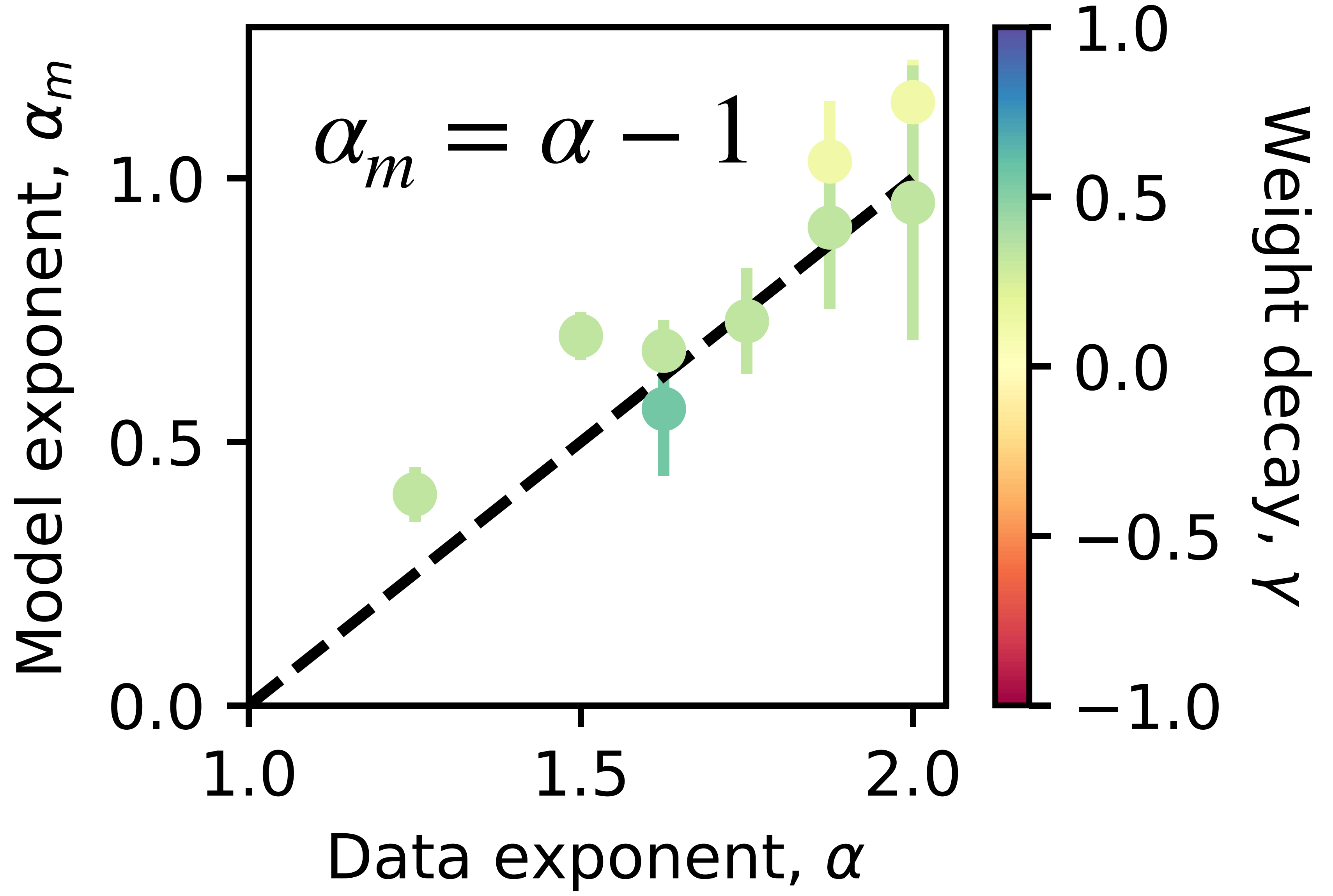


No superposition: Power law in, power law out

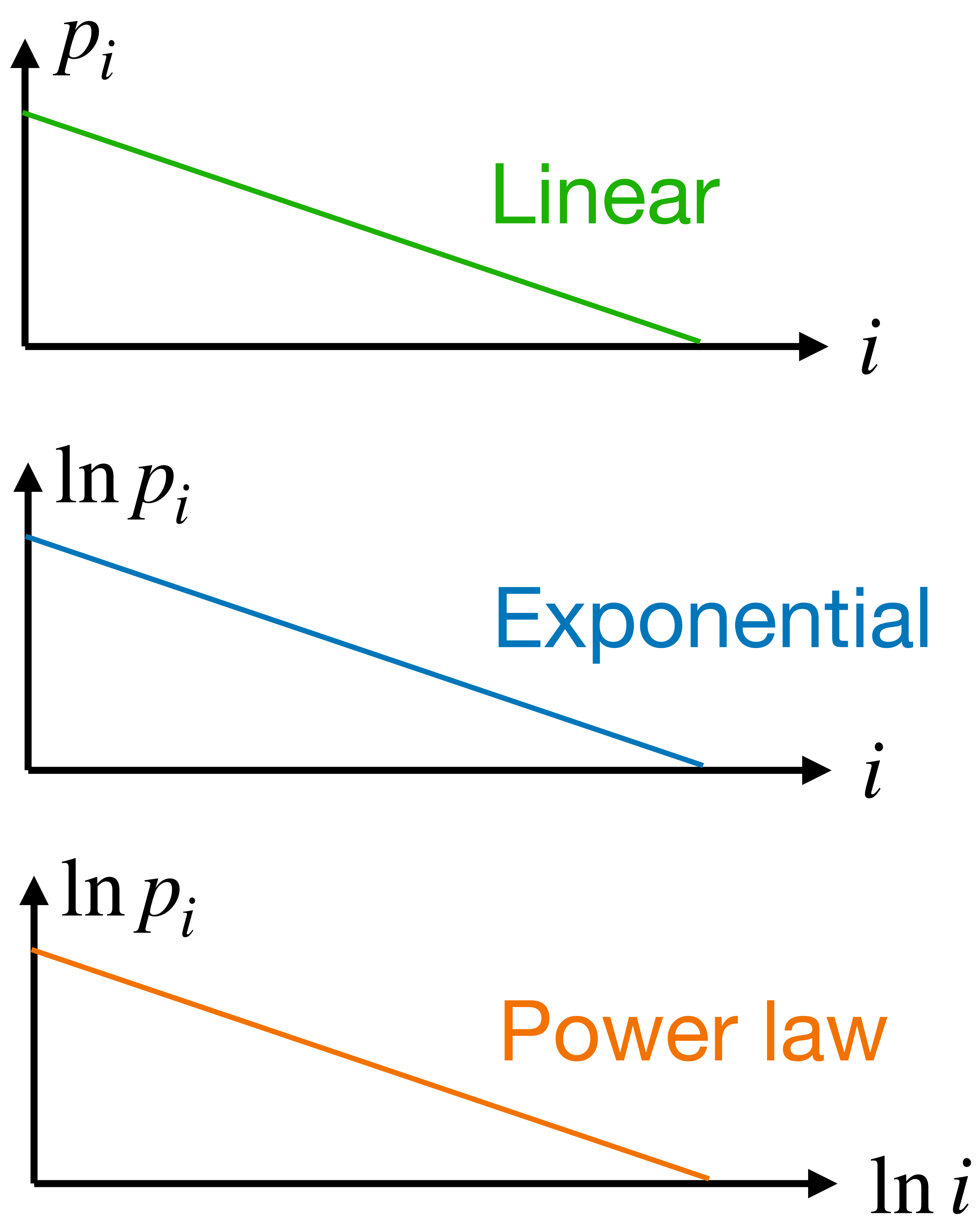


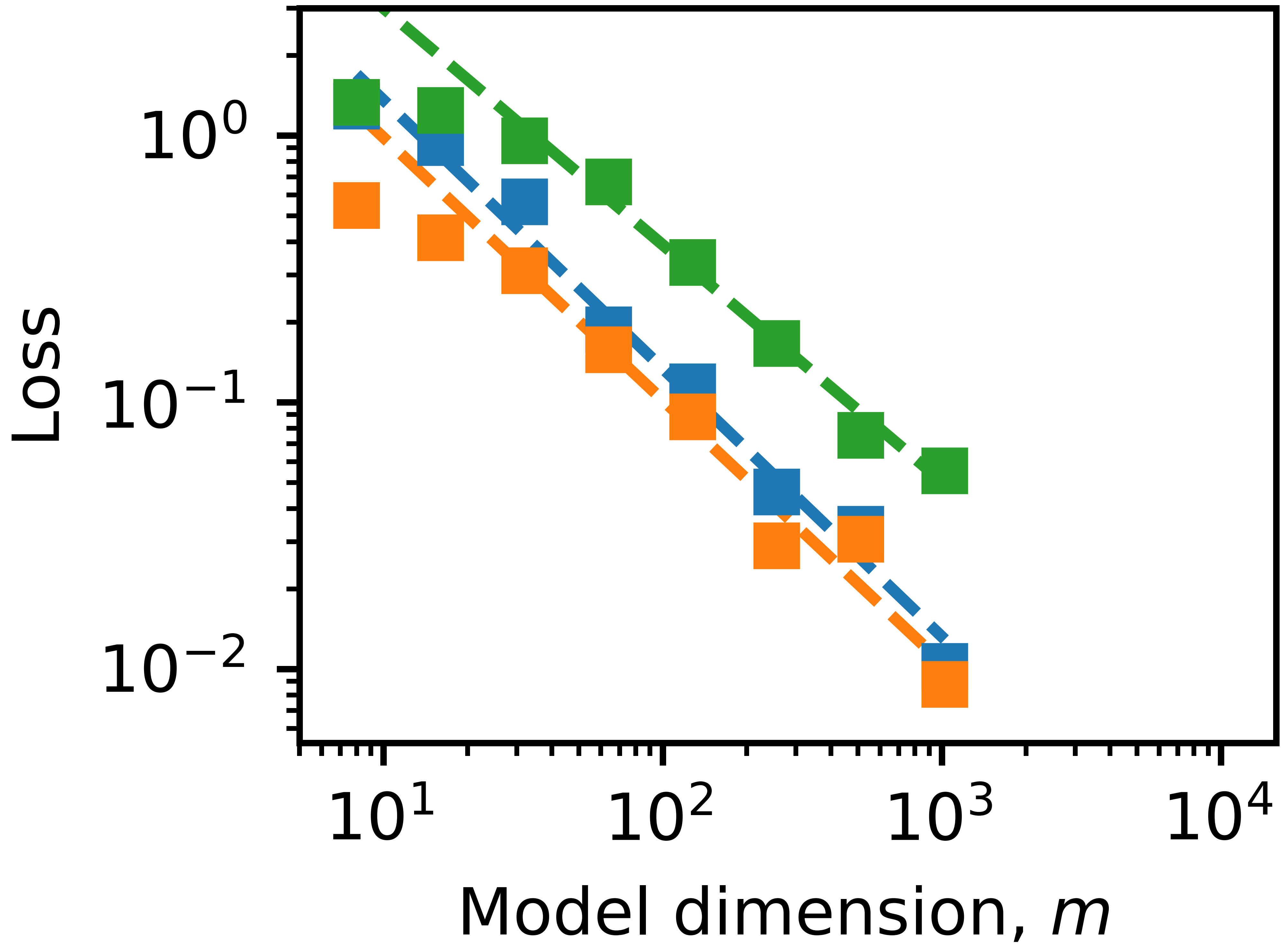
$$p_i \sim \frac{1}{i^{\alpha}} \qquad L \propto \frac{1}{m^{\alpha_m}}$$



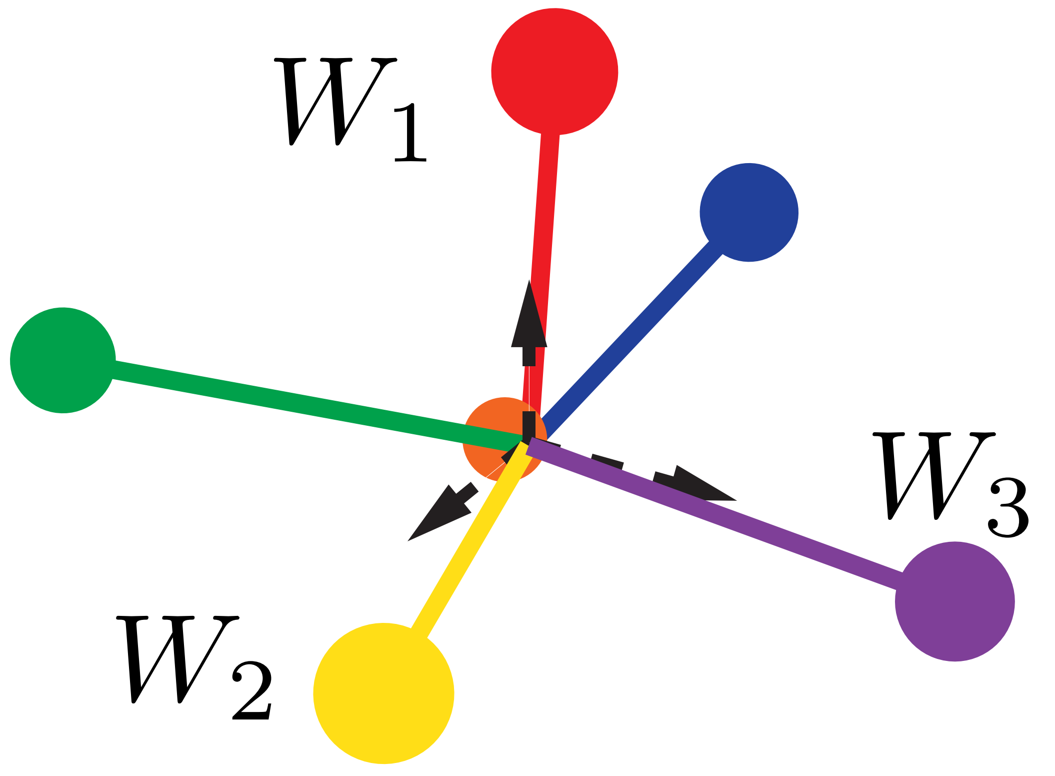


Emergence of robust power law at strong superposition

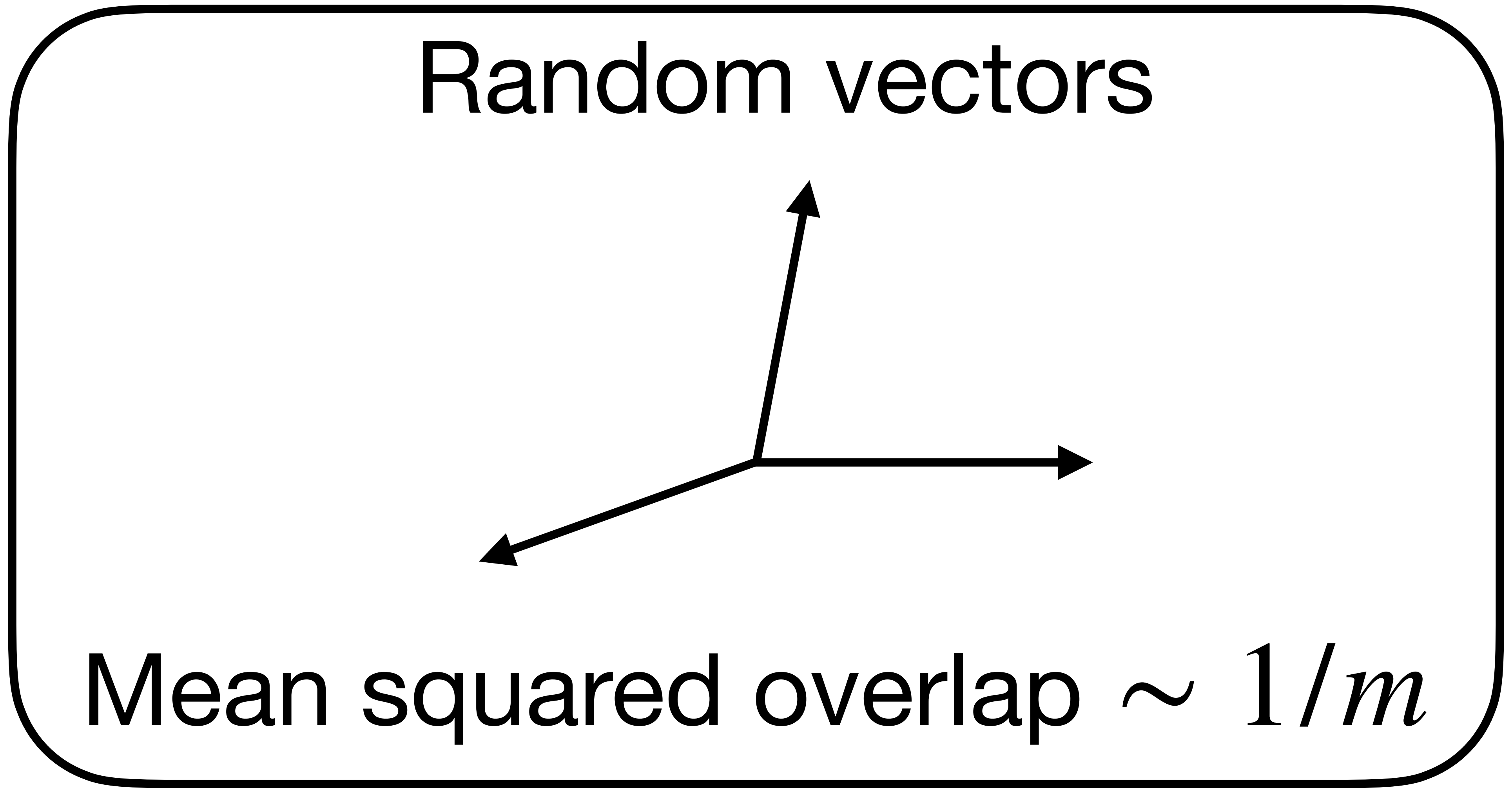


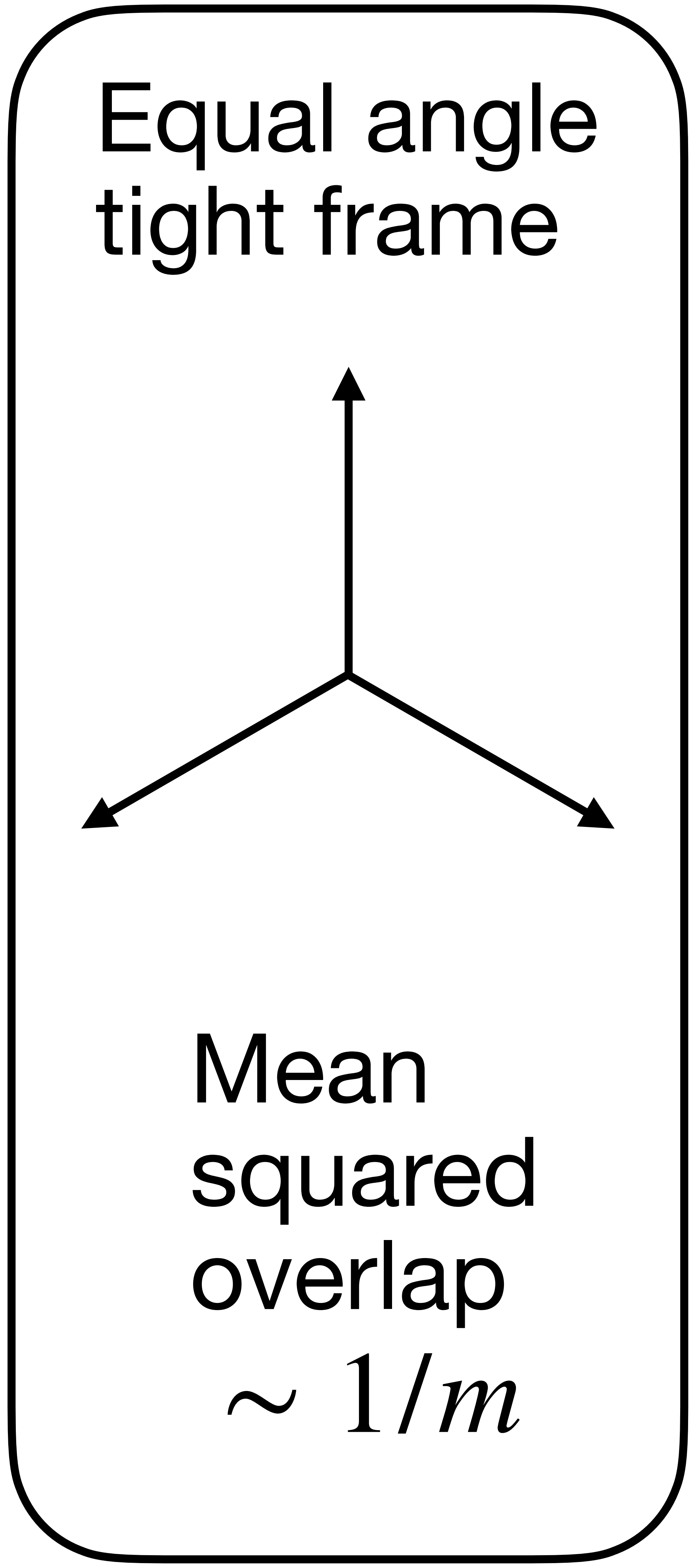


Power law emerges from intrinsic geometry

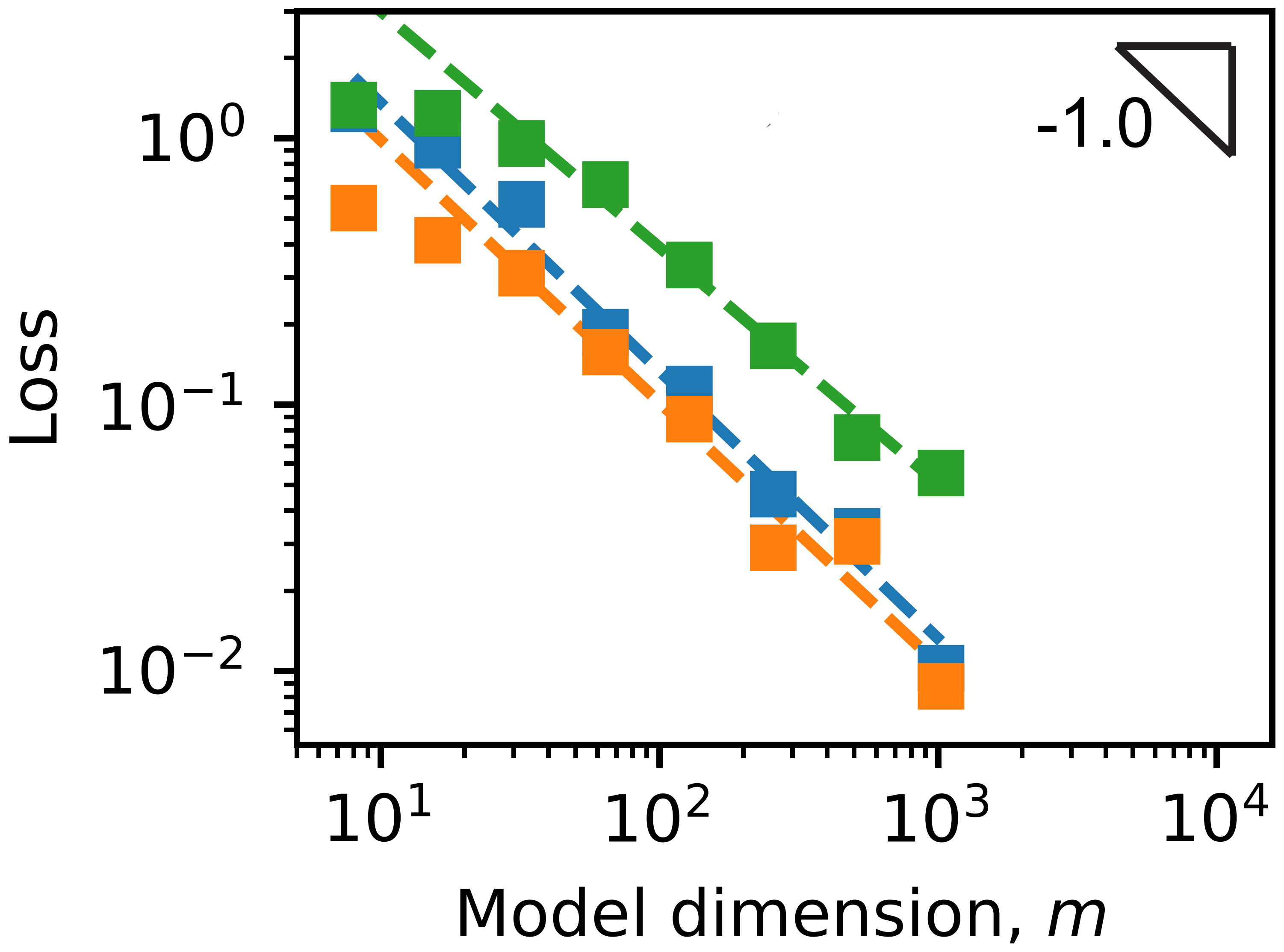


$L \leftarrow$ squared overlap



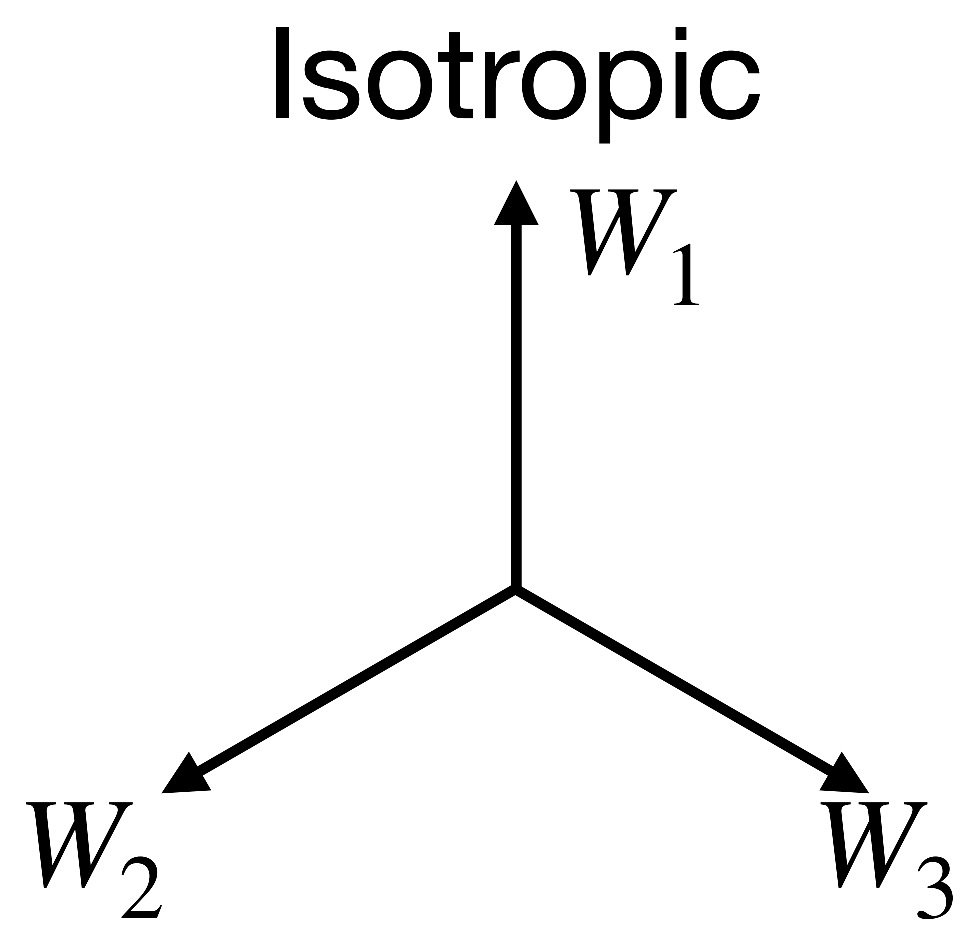


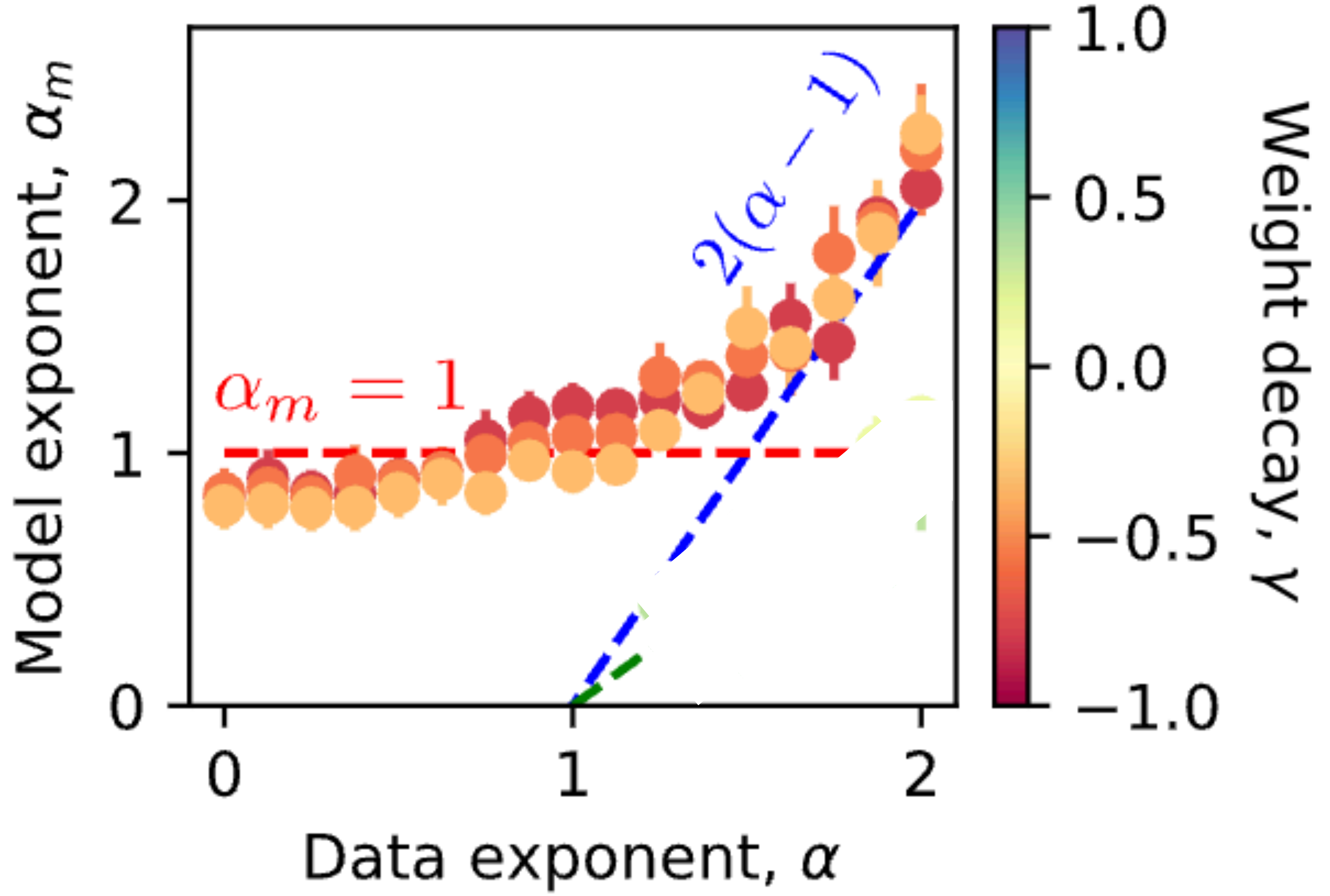
$$L \propto \frac{1}{m}$$

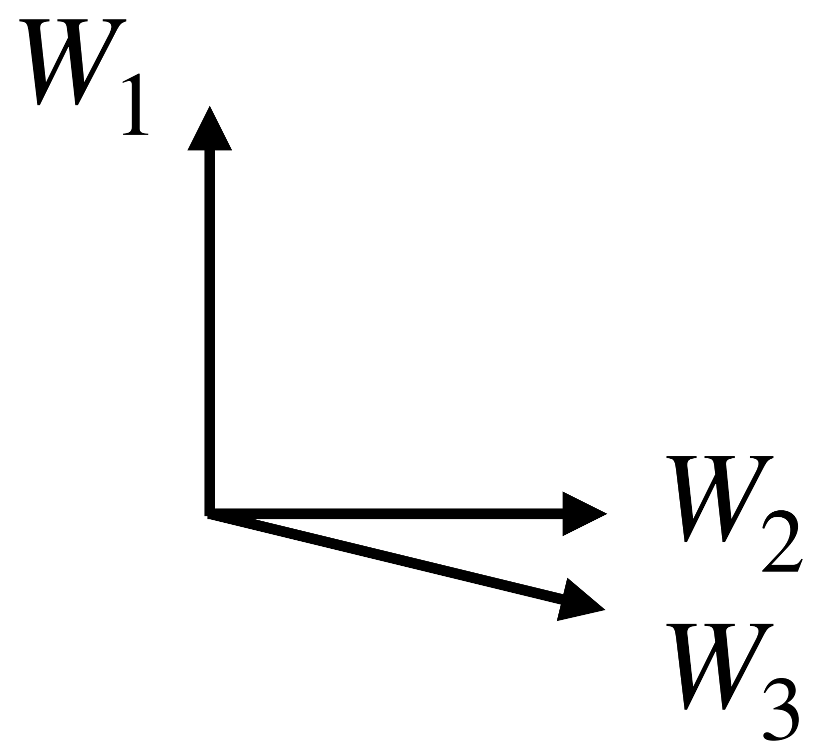


$1/m$ scaling is robust across a range of flat p_i

$$p_i \sim \frac{1}{i^\alpha} \qquad L \propto \frac{1}{m^{\alpha_m}}$$





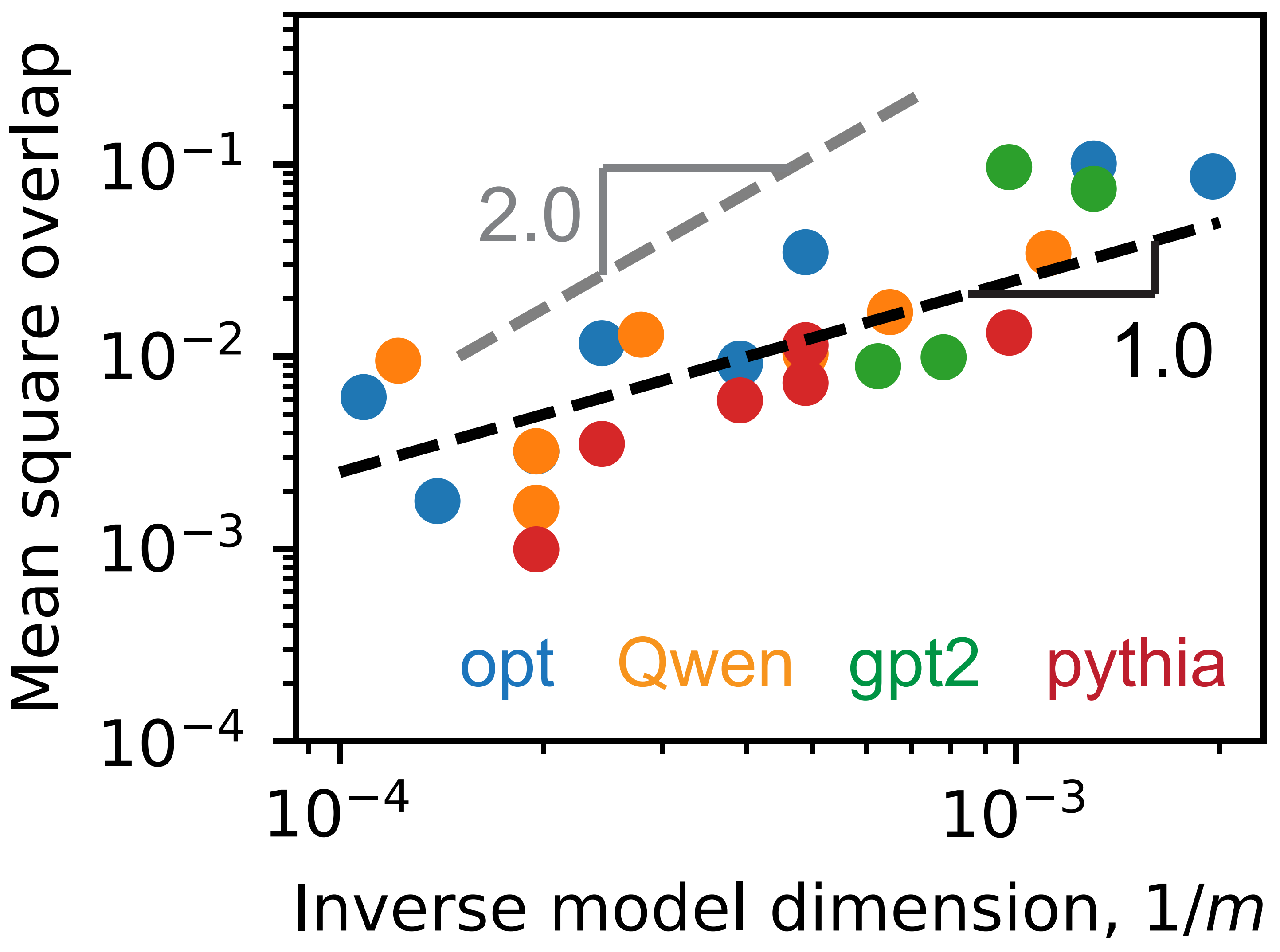


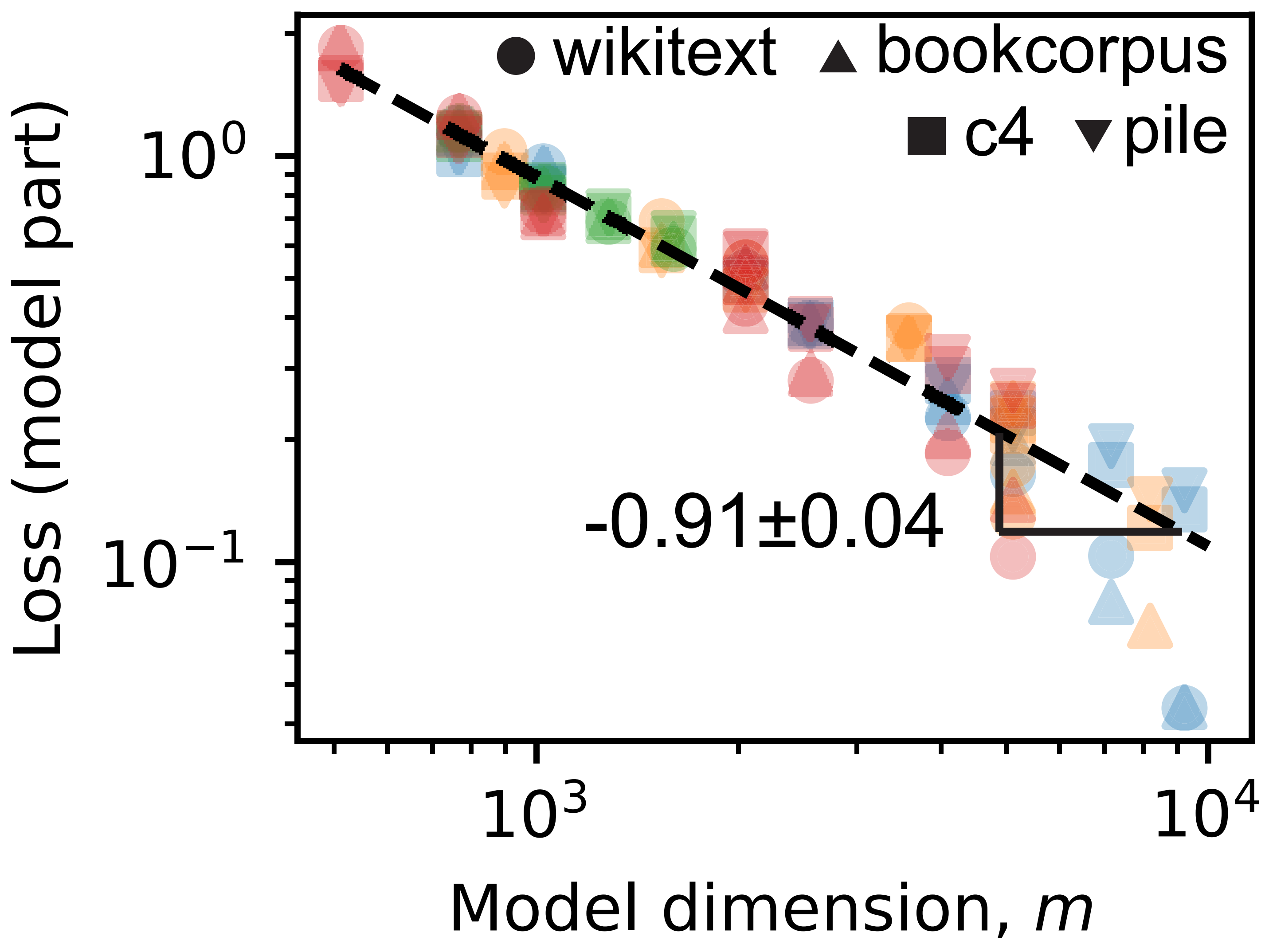
Superposition may explain the neural scaling law observed in actual LLMs

Naive mapping: atomic features -> tokens

Token frequency: Zipf's law, $\alpha = 1$ (small)

Toy model prediction: $1/m$ scaling due to intrinsic geometry!





Superposition may explain the neural scaling law observed in actual LLMs

Naive mapping: atomic features -> tokens

Token frequency: Zipf's law, $\alpha = 1$ (small)

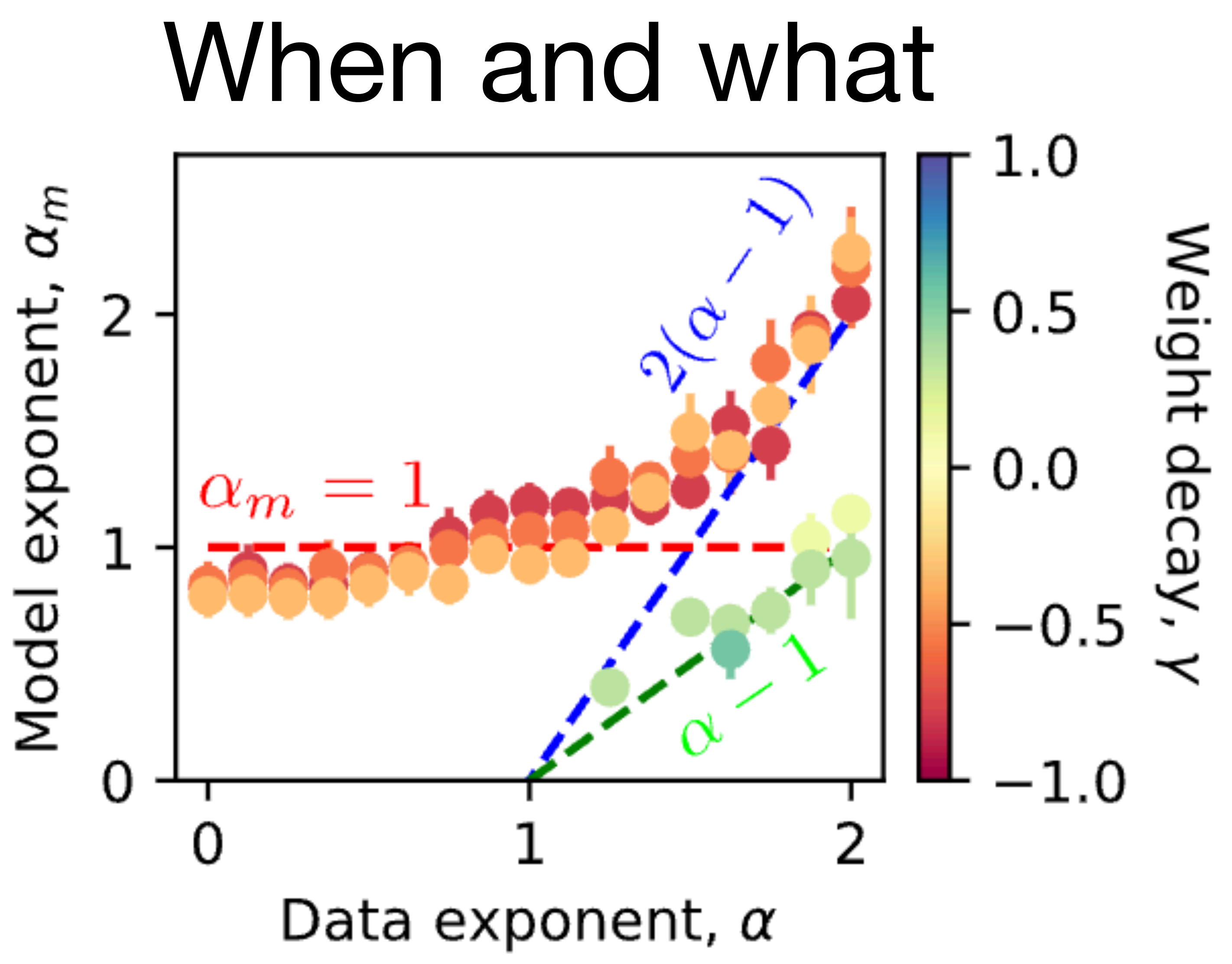
Toy model prediction: $1/m$ scaling due to intrinsic geometry!

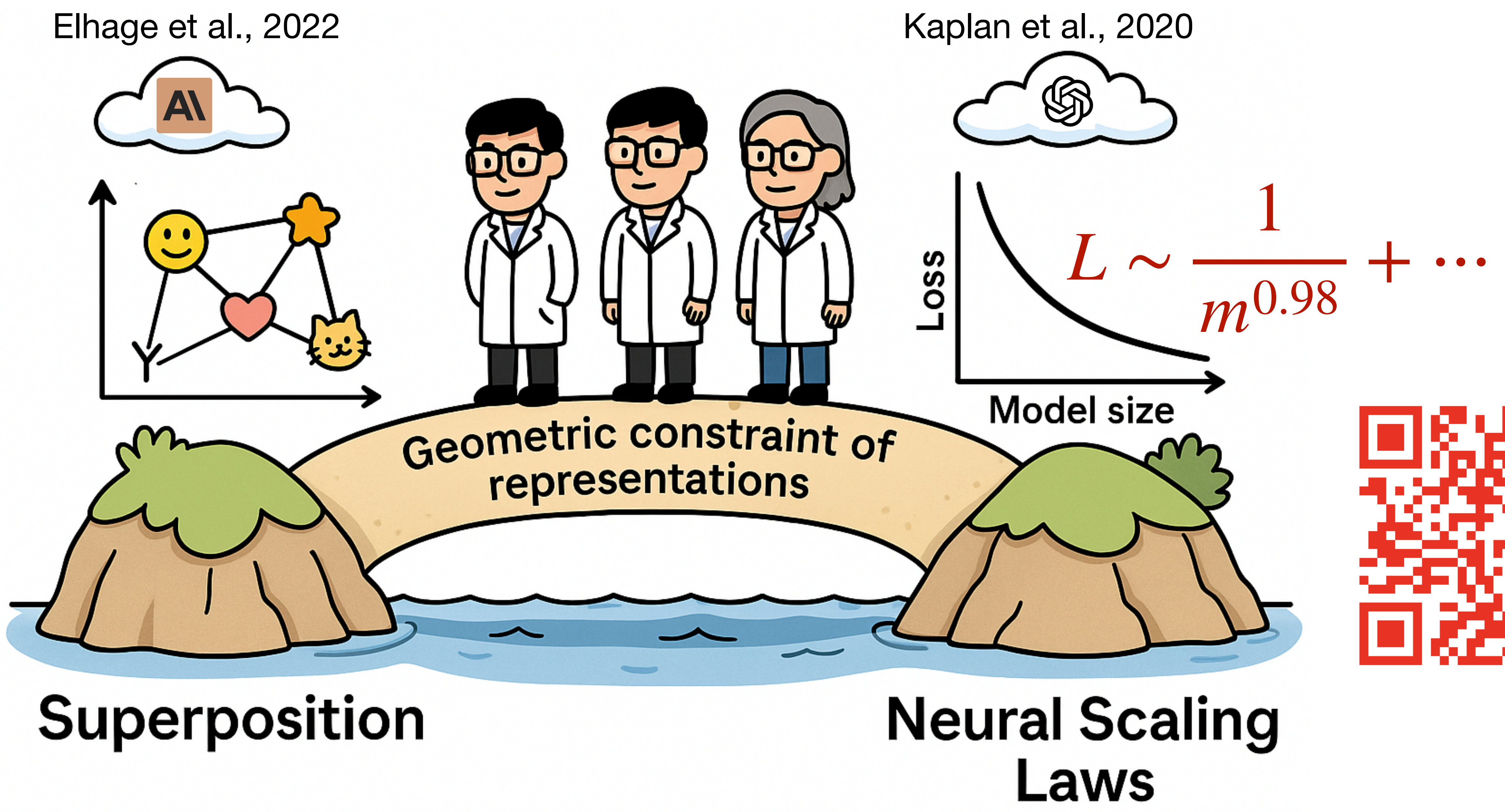
$$L = \frac{c_m}{m^{0.98}} + \frac{c_\ell}{\ell^{1.15}} + \cdots$$

Width

Depth

Superposition Yields Robust Neural Scaling!





Speedup?

Breaking down?

Poster: Exhibit Hall C,D,E #3717, right after.

To be continued

Appendix

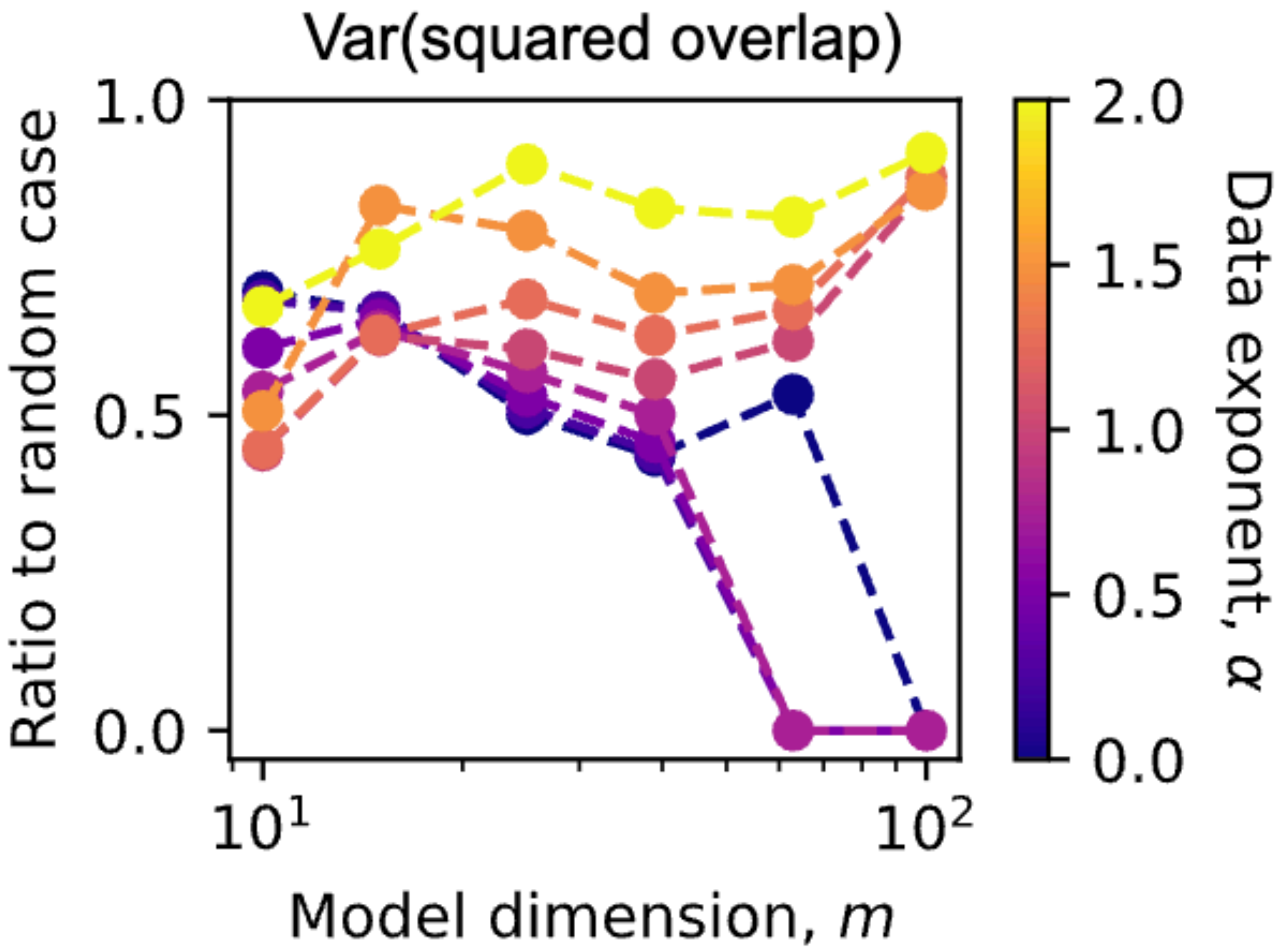
The toy model is far from LLMs yet is similar to LLMs in the aspect we care about

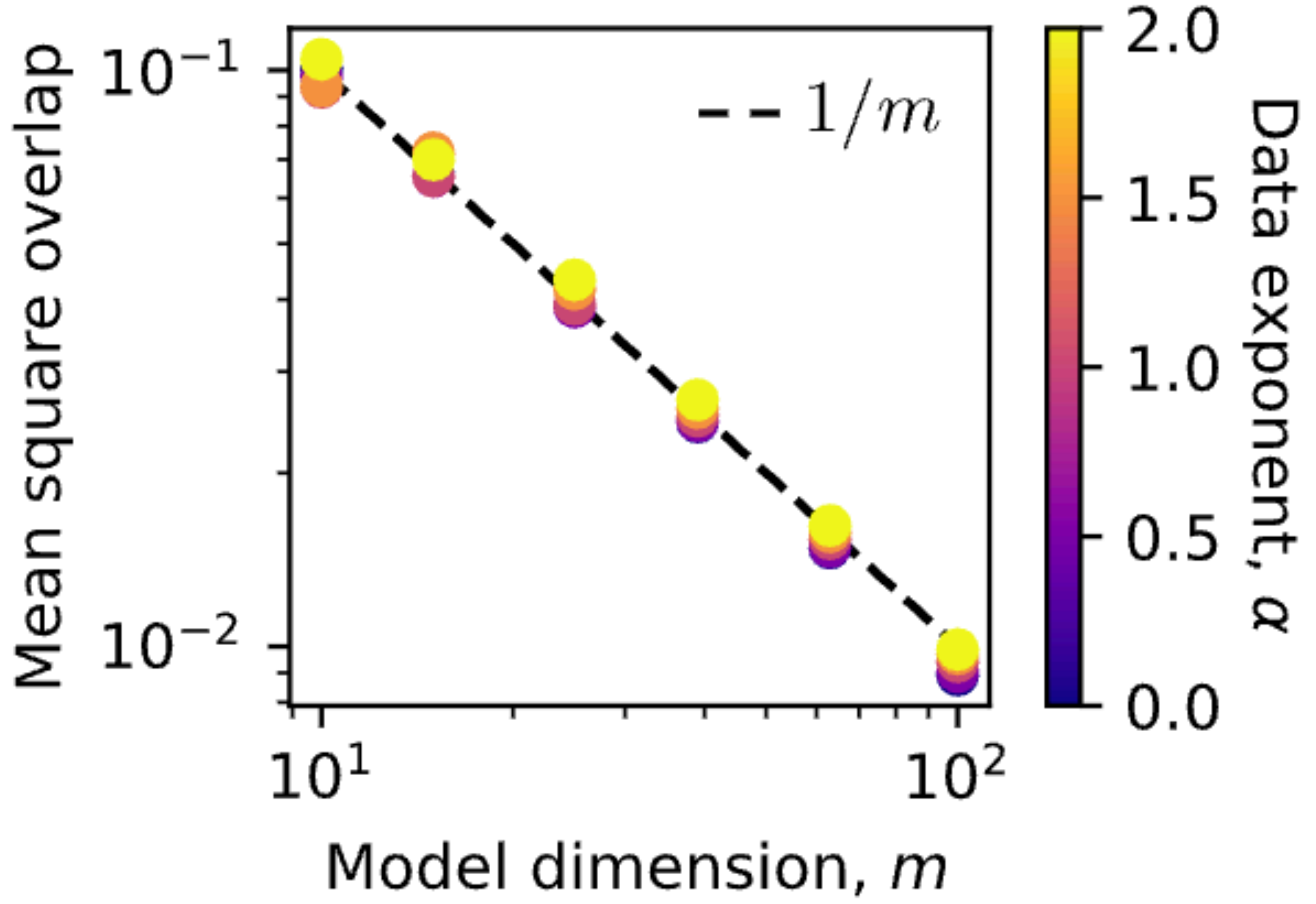
Conceptual connection: Atomic features are like tokens. Inputs are like sentences. Go from sparse high-dimensional representation to low-dimensional dense representations and then go back. Enough to see the representation loss.

Differences and why they may not matter:

> No transformer layers. We do not care about next-token prediction or any transitions.
>
> Toy models use ReLU and bias for error correction yet LLMs use Softmax. One can show that it does not change scaling.
>
> …

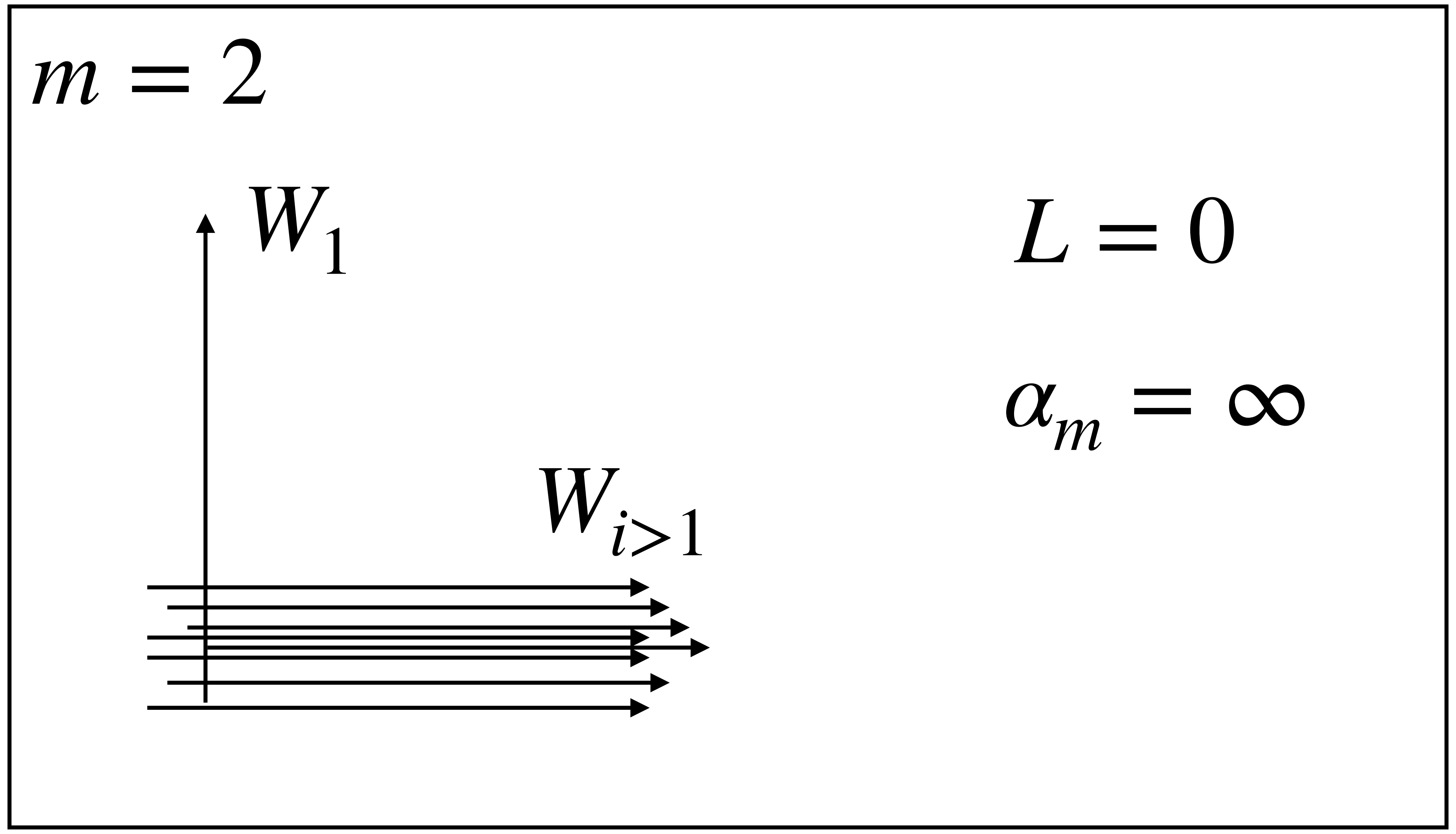
Models try to put many representations into ETF-like configurations



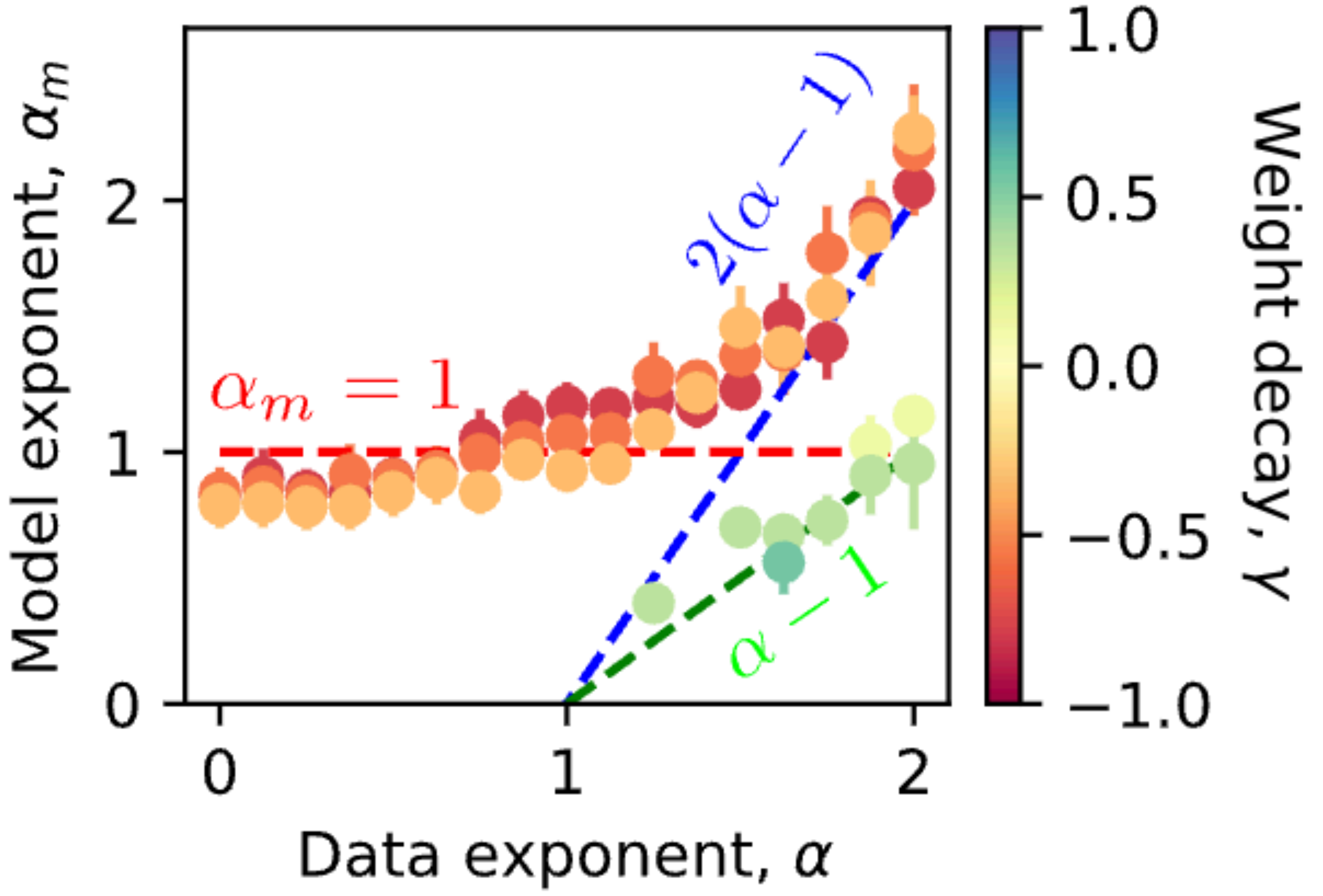


For skewed frequencies, loss scaling can depends on data again due to non-isotropic vectors

$\alpha = \infty$ only the first feature exists

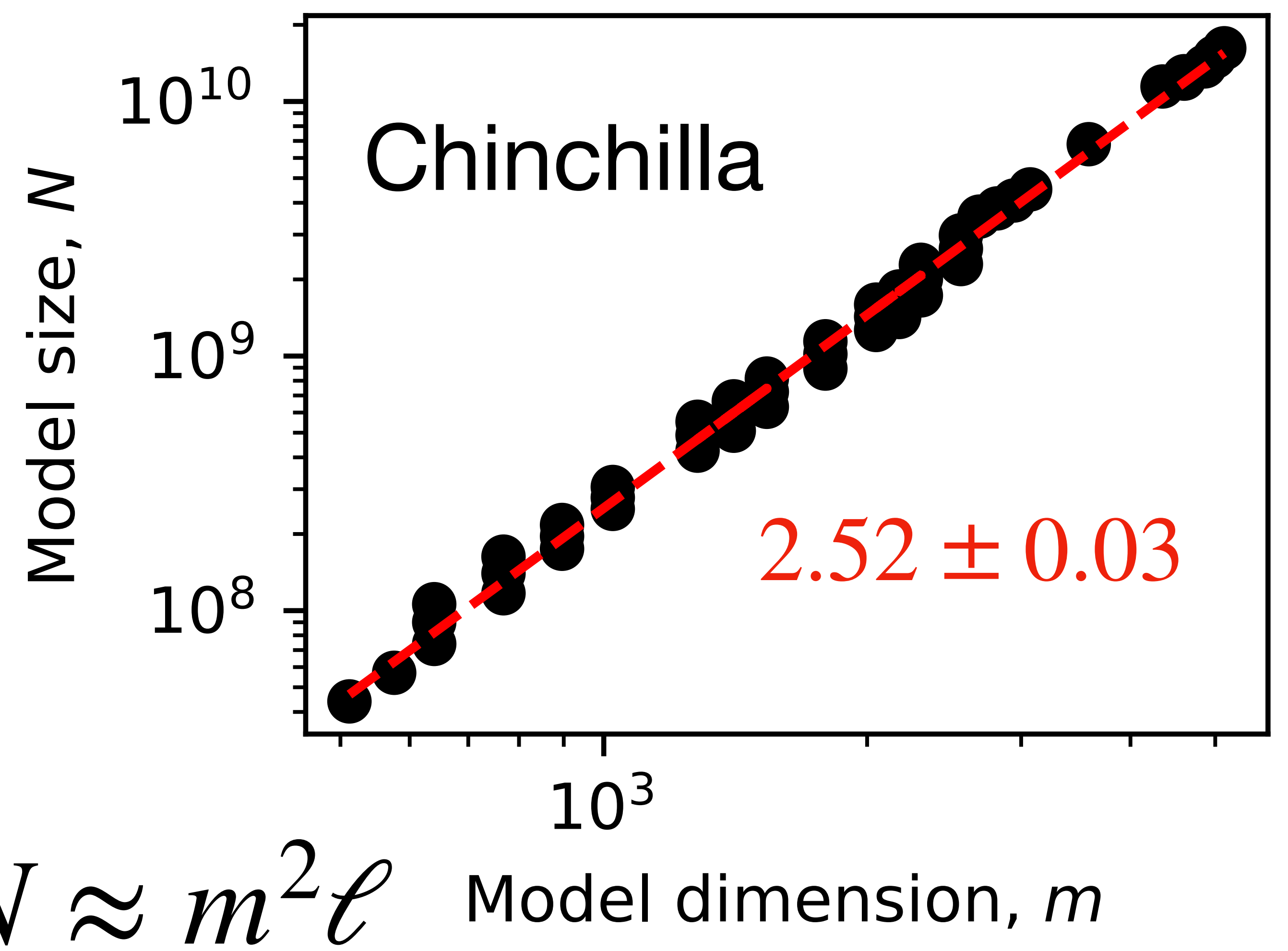


$L \propto \sum_{i>m^2/2} p_i?$



Depth and width are related when scaling up

Hoffmann et al., (2022)



$$N \approx m^2 \ell$$

$$L = \frac{1}{D^{0.28}} + \frac{1}{N^{0.34}} + 1.69$$

Variance-limited, resource model, lottery ticket

Many neurons for one task/data point, cancelling each others' error/noise, and central limit theorem leads to $1/m$ scaling.

It is another limit, conceptually different from superposition (too few neurons for too many features)

Linear models have no superposition

Assume power law in data, and obtain power-law neural scaling

NTK picture assumes model-seen features

Effectively linear

Not intrinsic data property

Φ Feature map, N eigenfunctions of NTK

Assume $\langle \Phi(x)\Phi^T(x)\rangle_x$ has power-law spectrum $\alpha_N = \alpha' - 1$

A gap, superposition is about width scaling

We assume intrinsic data property $\langle xx^T\rangle_x$

$\alpha_m = \alpha - 1$ if weak superposition

$\alpha_m = 1$ if strong superposition for a range of $\langle xx^T\rangle_x$, $\alpha' = 2$?

Connection: a case that model seen features are different from intrinsic data features

Interpolation on data manifold, or fitting continuous function on manifolds

$$\alpha_N = \frac{4}{d}$$

Seems not to be strong superposition

One example, $x \in \mathbb{R}^n$, each element iid, $d = n$?

Superposition in this picture is like using one parameter to fit different regions of the manifold

Superposition (non-linearity) may change the intrinsic dim we can measure from hidden states

Exponentially many features can be stored?

Fix error ϵ

EVD $\dfrac{\ln n}{m} < \epsilon$

$n = O(\exp(m\epsilon))$

Our case:

Large $n \gg m$, how mean error/ loss continuous decrease with m

$L \sim \dfrac{1}{m}$

They are consistent

…the real successes come to those who start from a physical point of view, people have a rough idea where they are going to and then begin by making the right kind of approximations, knowing what is big and what is small in a given complicated situation.

Methodology

Physicists always have a habit of taking the simplest example of any phenomenon and call it “physics”, leaving the more complicated examples to become the concern of other fields…

The Feynman Lectures on Physics

Limitations

Our work is built on observations of the toy model and analysis without rigorously solving the toy model. We are thus limited to explaining deeper behaviors in the toy model. Our analysis of LLMs suggests they are in the strong superposition regime, but the underlying reasons were not studied in detail. We believe one reason is that features are sparse in language, as the number of tokens required to predict one token is much less than the total number of tokens. The softmax function may also be important since it is strong at error correction, giving superposition an advantage.

Neural scaling laws also include scaling laws with dataset size and with training steps, which we did not study. At each step, a fixed number of new data points are used for optimization. So, we expect the scaling with the total data amount and that with training steps will be the same, similar to the results at weak superposition. However, in the strong superposition regime, data or training step scaling is related to angle distribution and how angles between representations evolve, which cannot be easily explained without rigorous solving.

$$L = \frac{1}{m^{0.98}} + \frac{1}{\ell^{1.15}} + \frac{1}{t^{0.28}} + \frac{1}{s^{0.5}} + L_0$$