

On the Expressivity Role of LayerNorm in Transformers' Attention

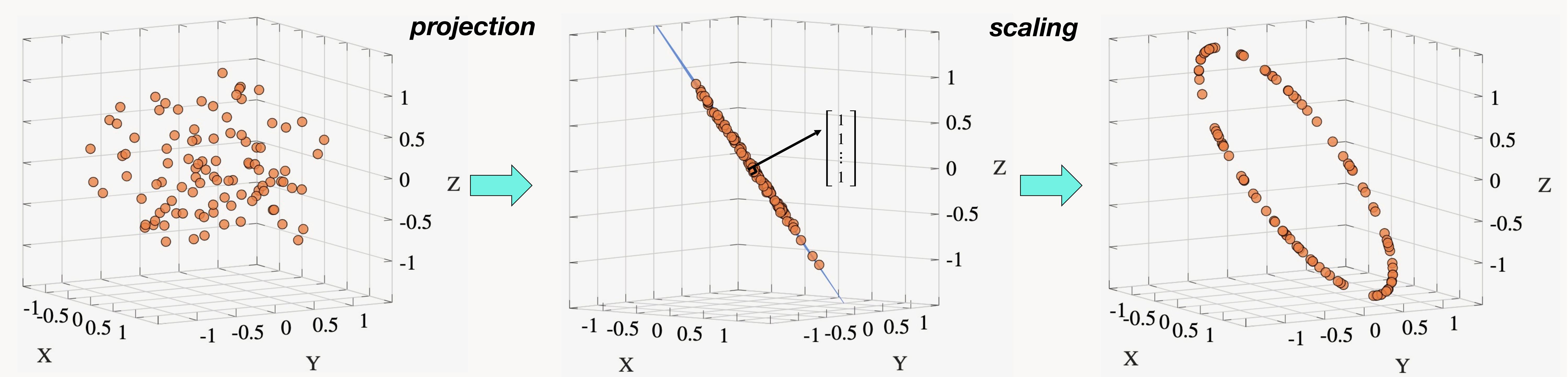
Shaked Brody, Uri Alon, Eran Yahav

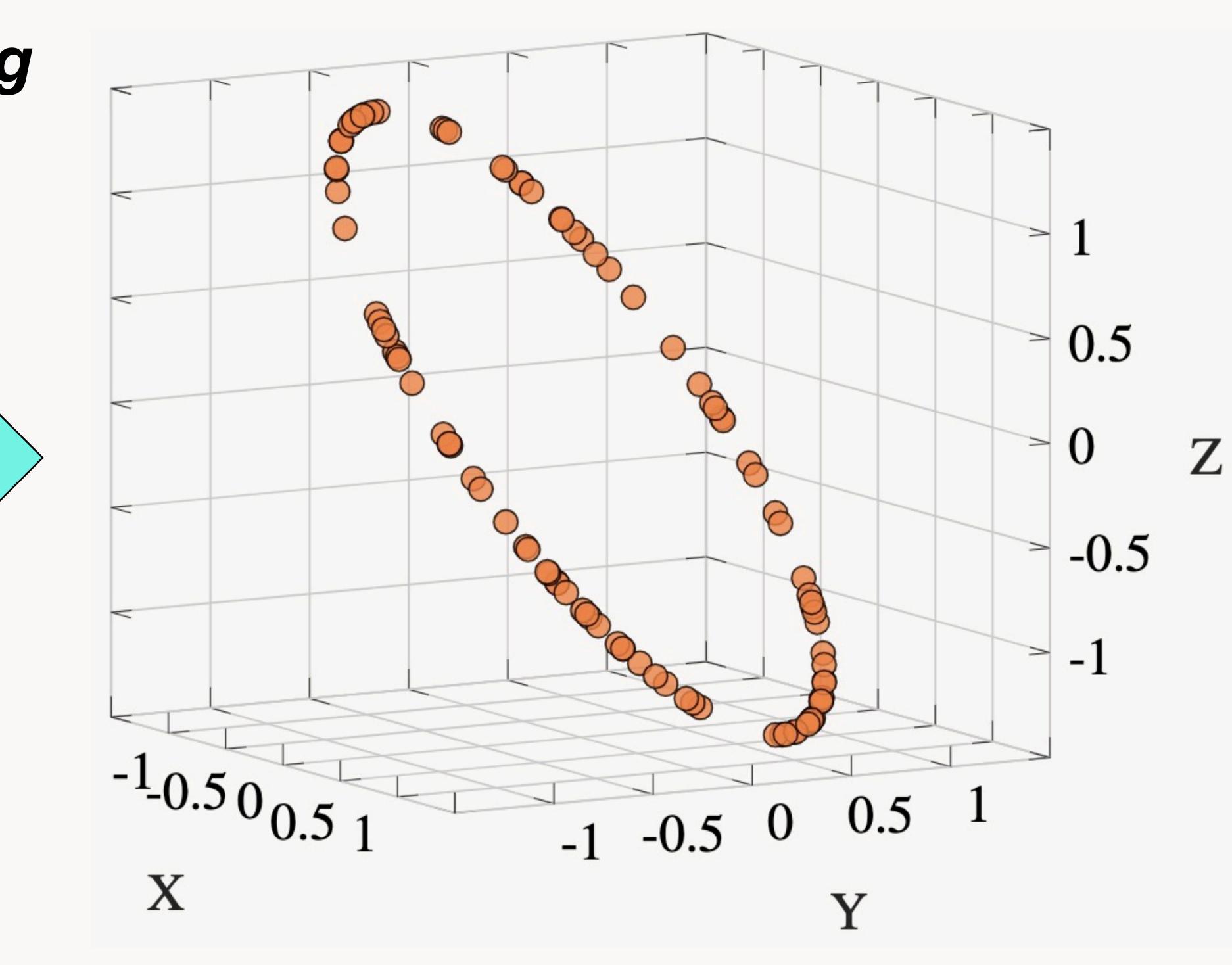
Code: https://github.com/tech-srl/layer_norm_expressivity_role

Known so far LayerNorm is the least studied component of Transformers Normalizing activations Normalizing gradients in the forward pass in the backward pass Inputs

This work: A Geometric Interpretation of LayerNorm

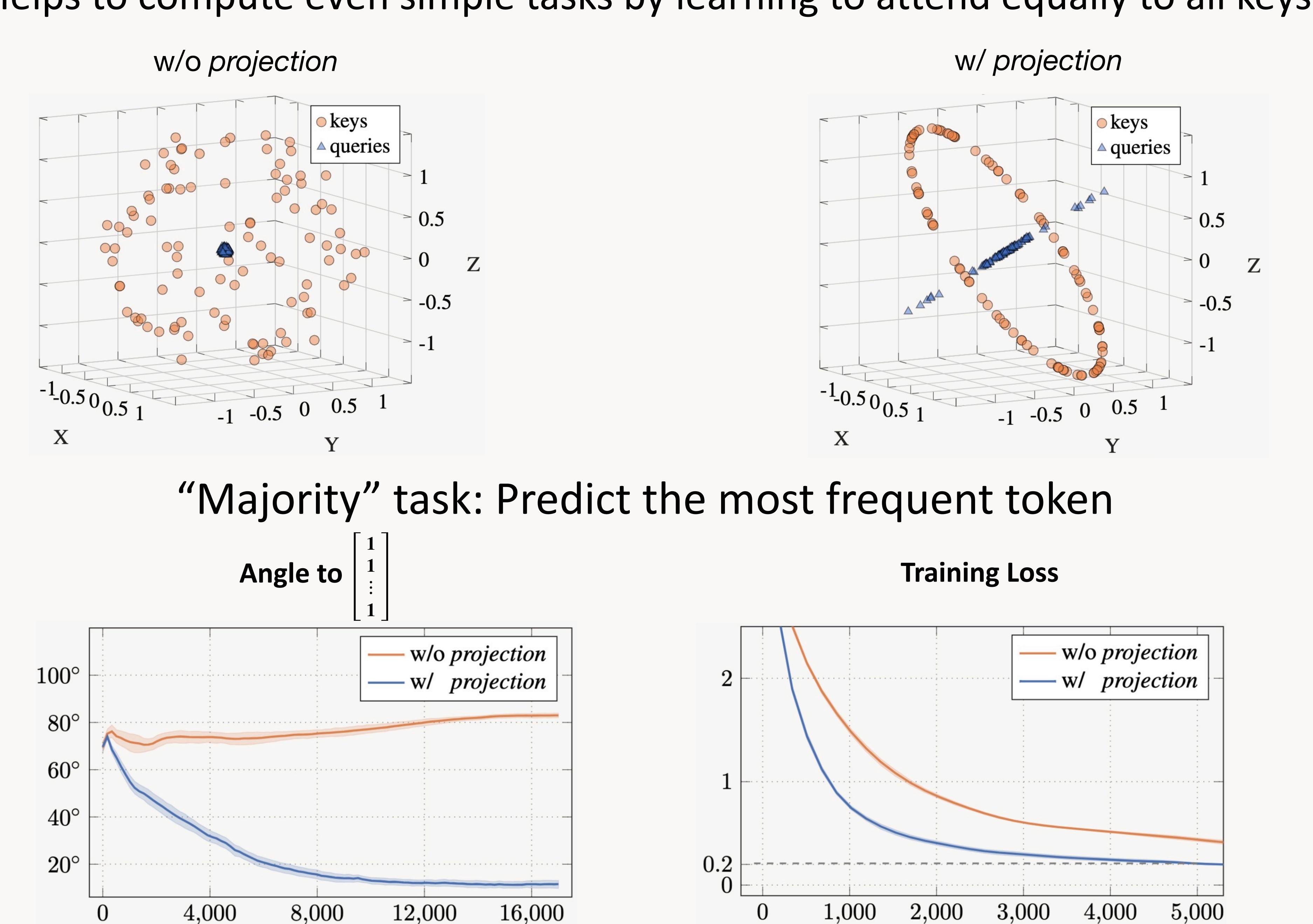
LayerNorm can be seen as a *projection* followed by a *scaling* operation



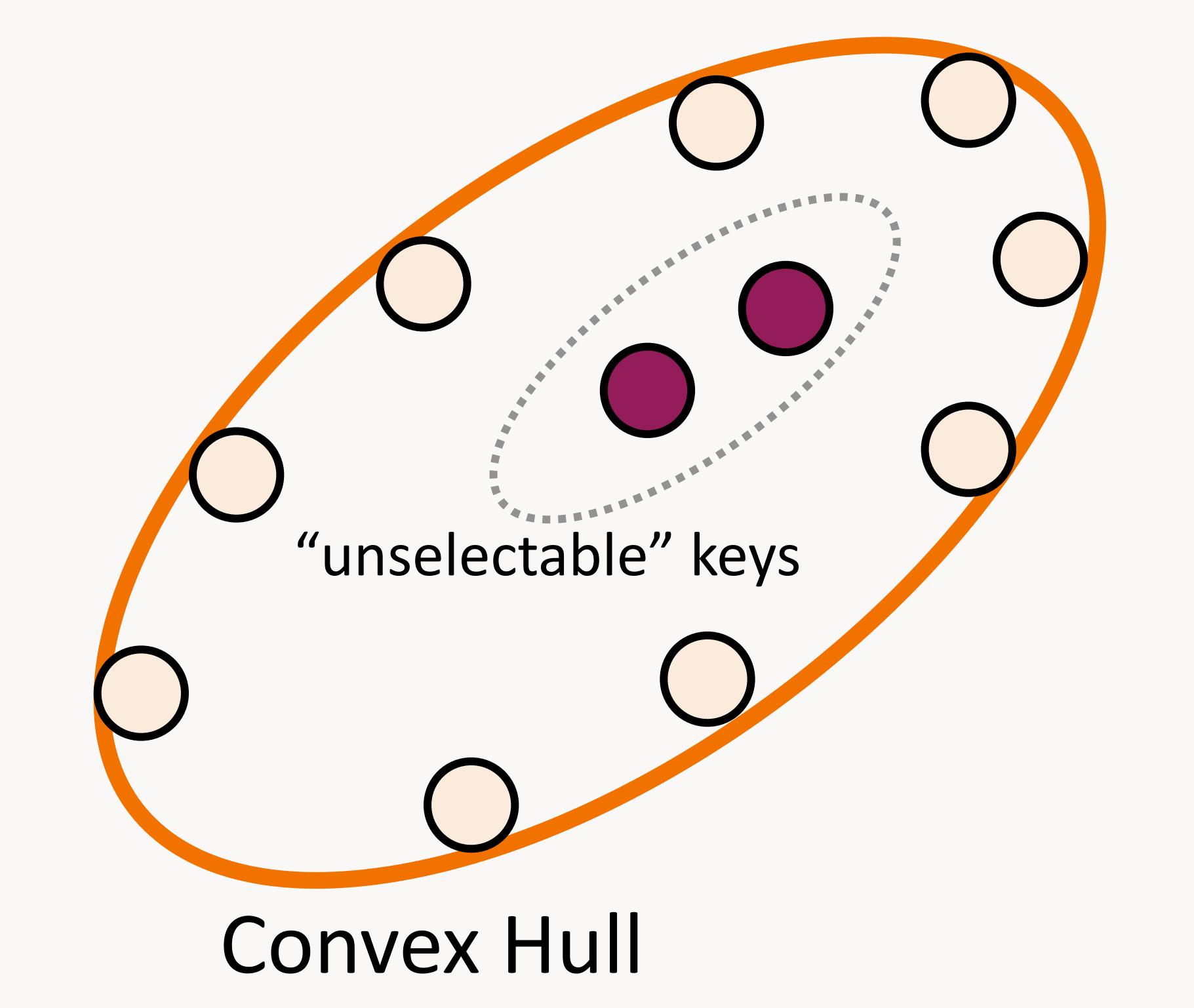


Projection

Helps to compute even simple tasks by learning to attend equally to all keys



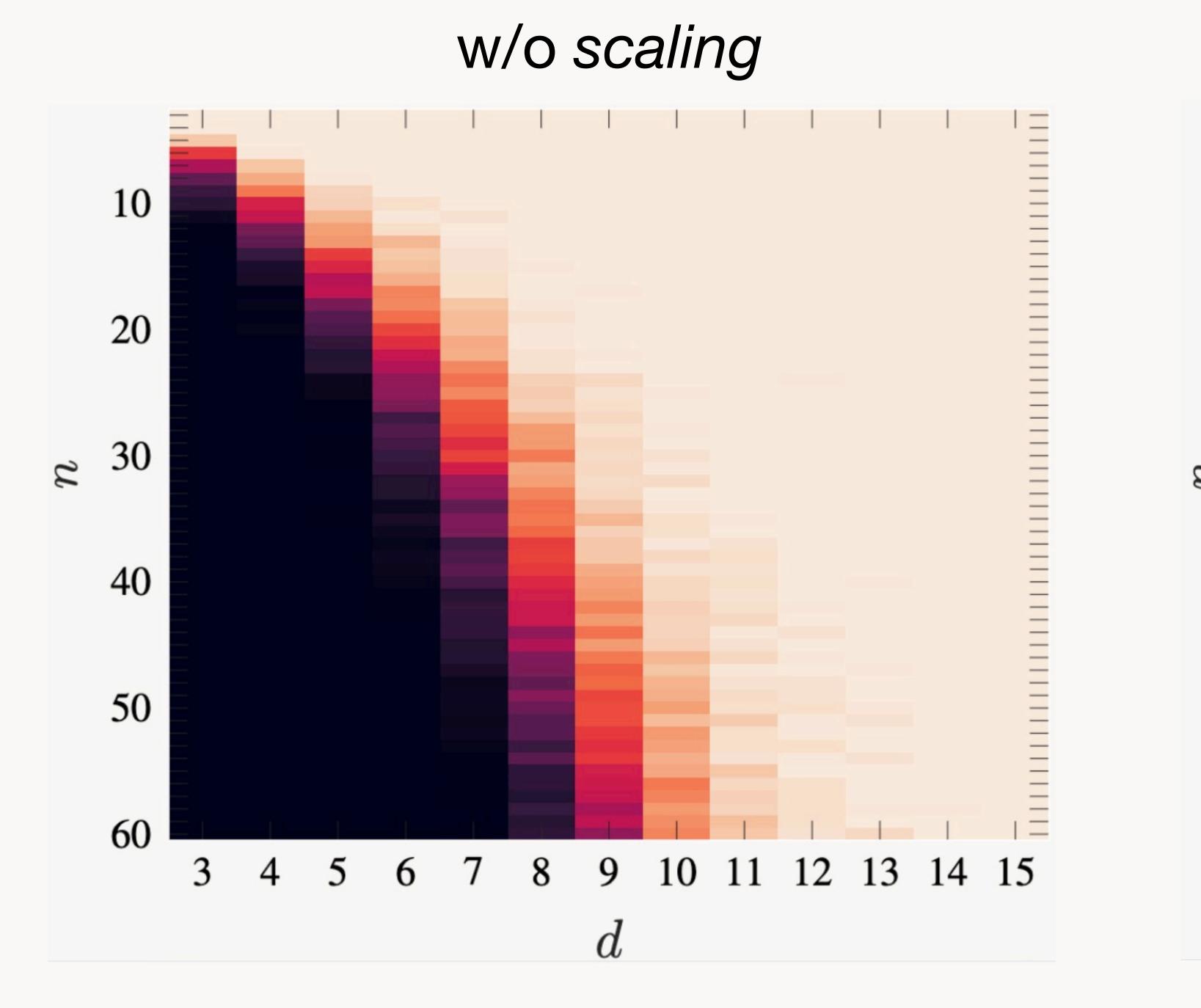
The "unselectable" keys problem

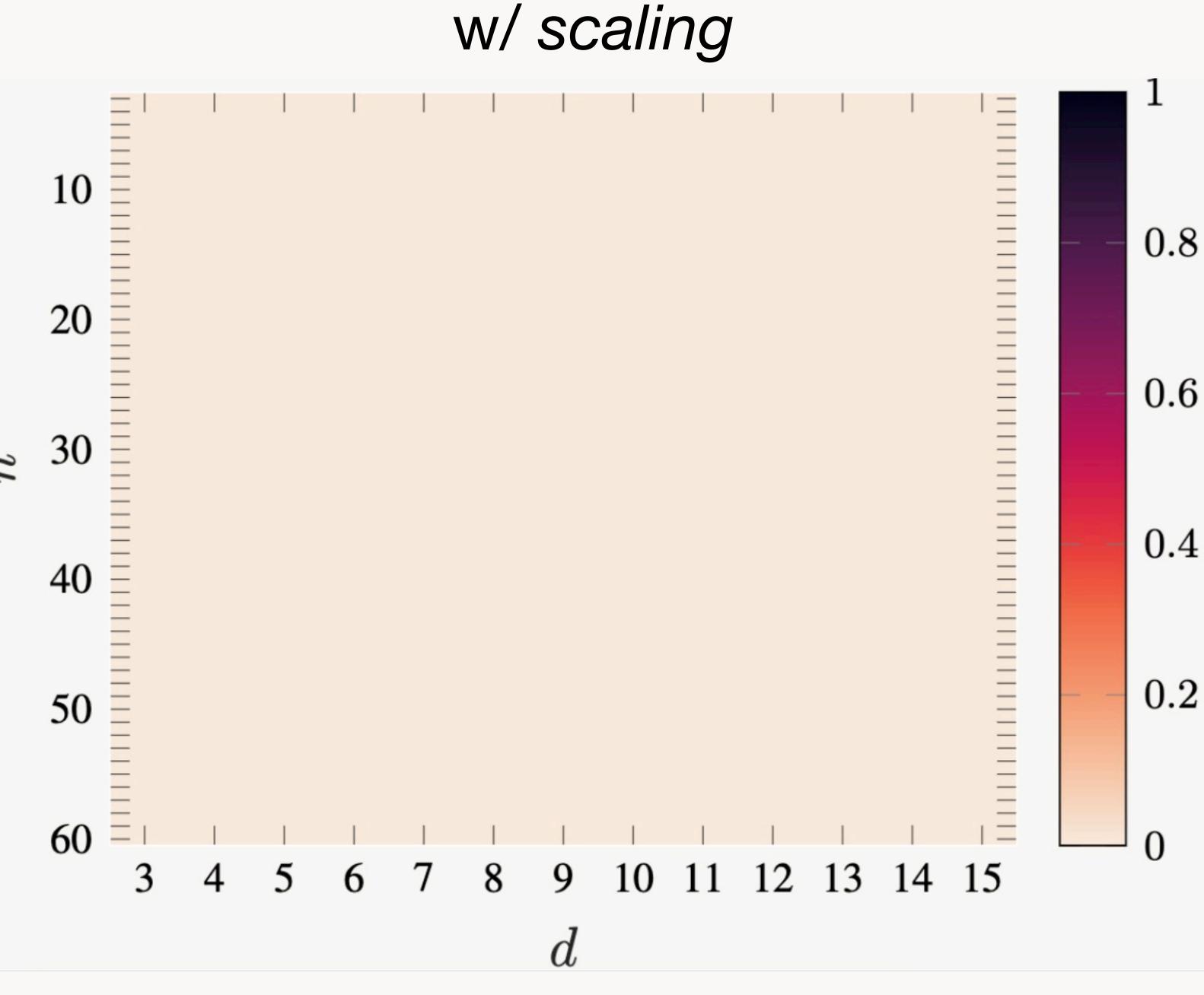


Interior key vectors cannot receive the highest attention score

Scaling

The fraction of "unselectable" n random vectors of dimension d





The fraction of "unselectable" key vectors in different layers of a language model

Model	L_1	L_2	L_3	L_4
w/o scaling	51.0	32.2	34.7	36.8
w/ scaling	0	0	0	0