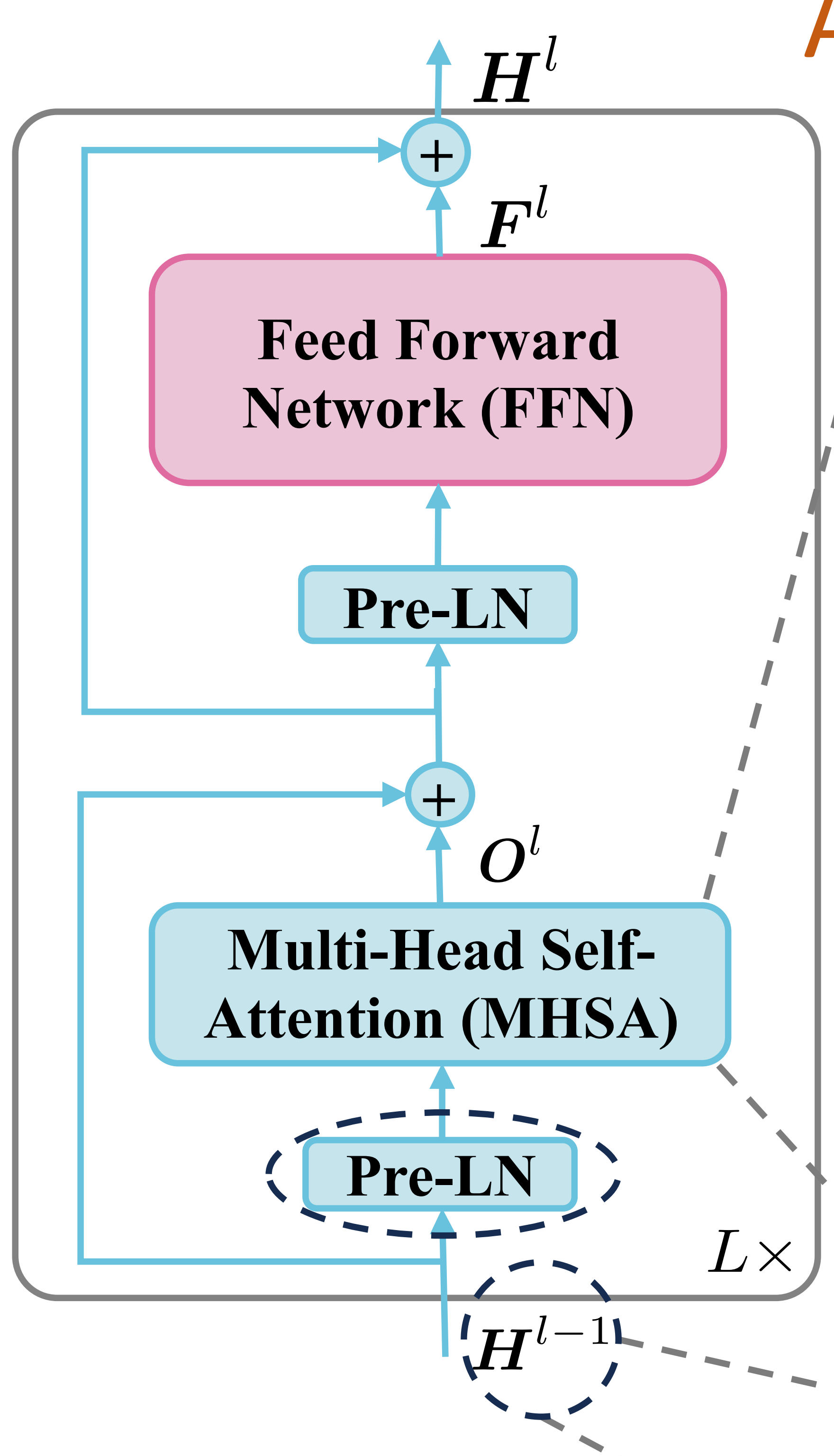
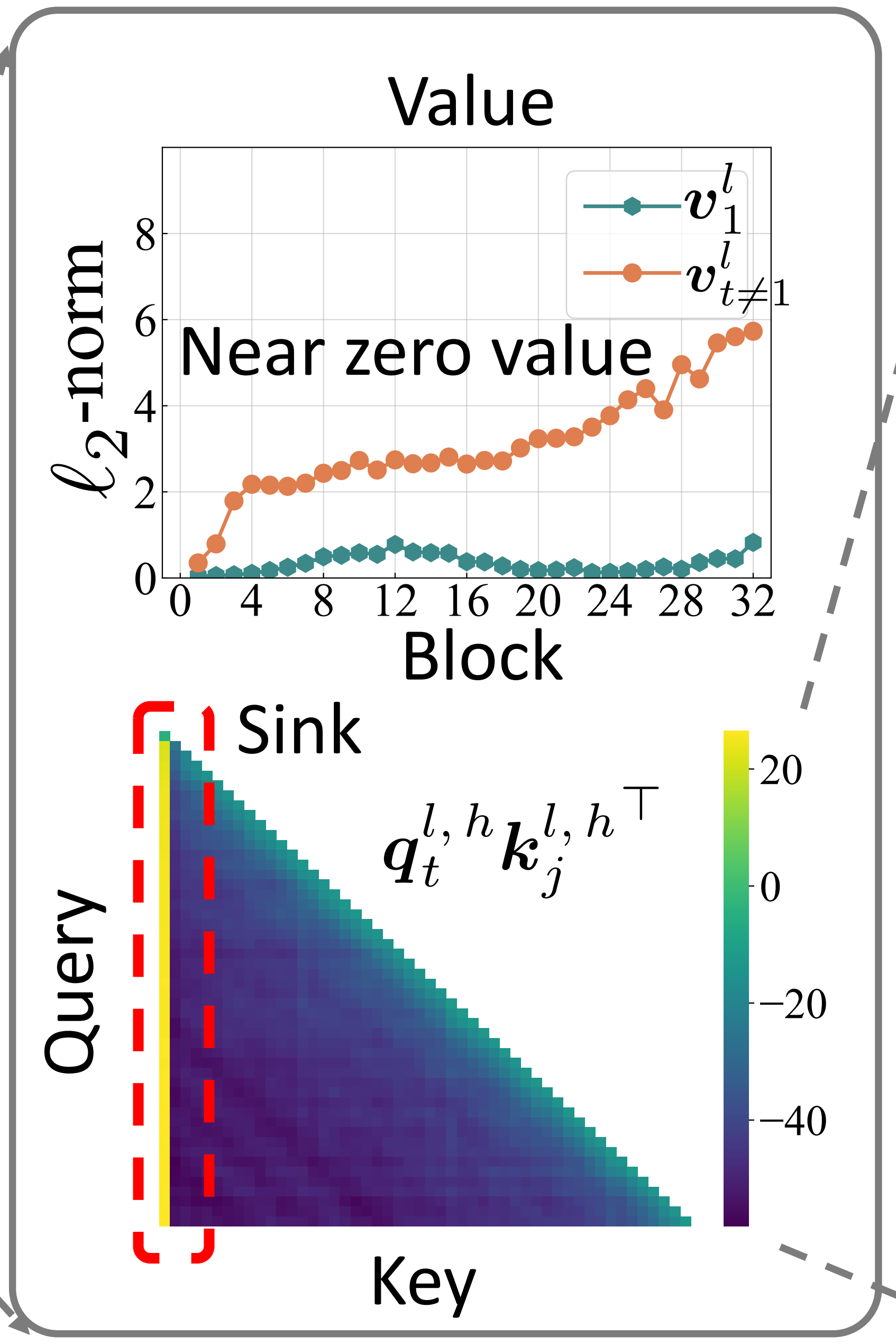
When Attention Sink Emerges in Language Models: An Empirical View

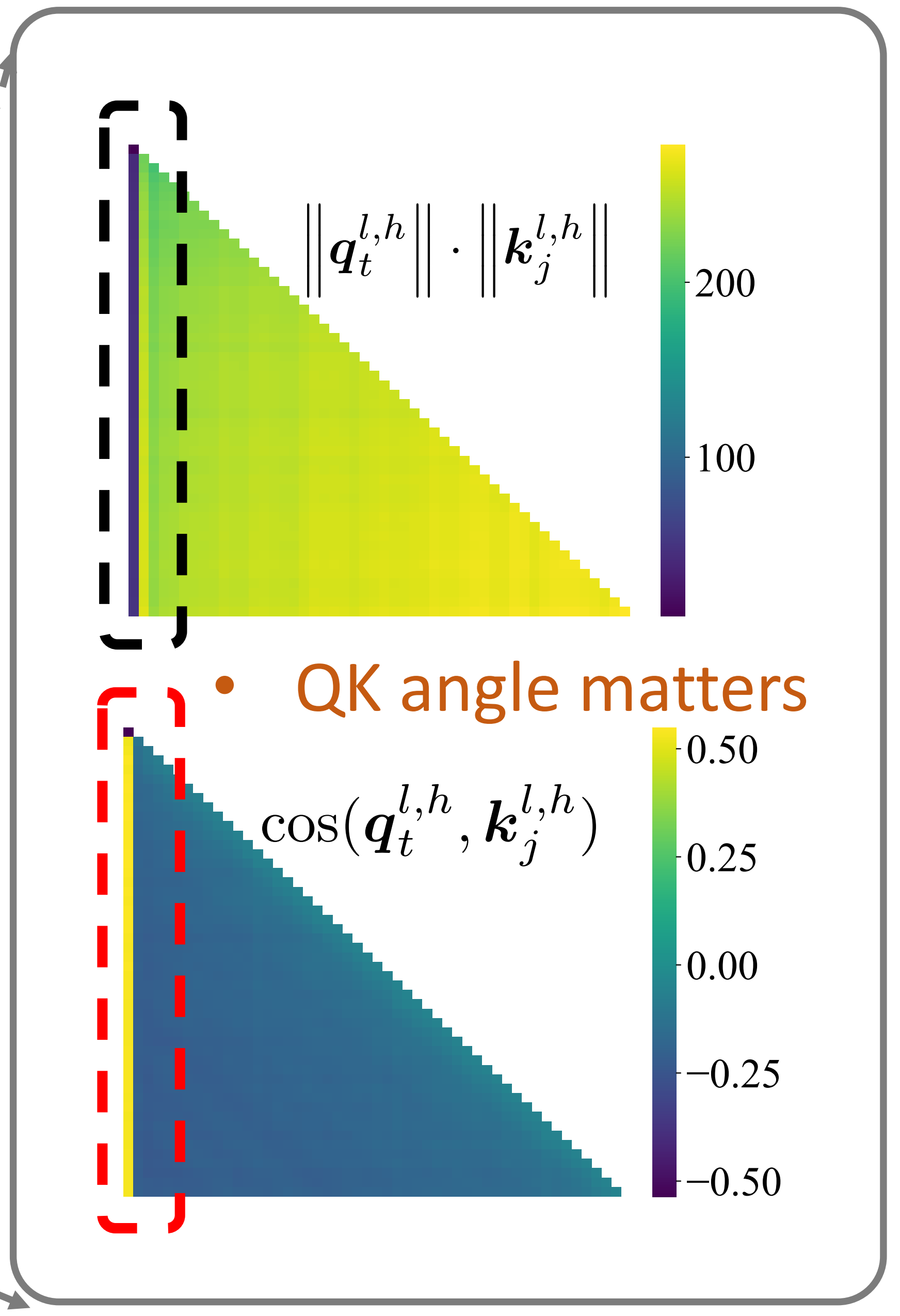
Xiangming Gu[1,2], Tianyu Pang[1], Chao Du[1], Qian Liu[1], Fengzhuo Zhang[1,2], Cunxiao Du[1], Ye Wang[2], Min Lin[1]

[1]*Sea AI Lab* [2]*National University of Singapore*

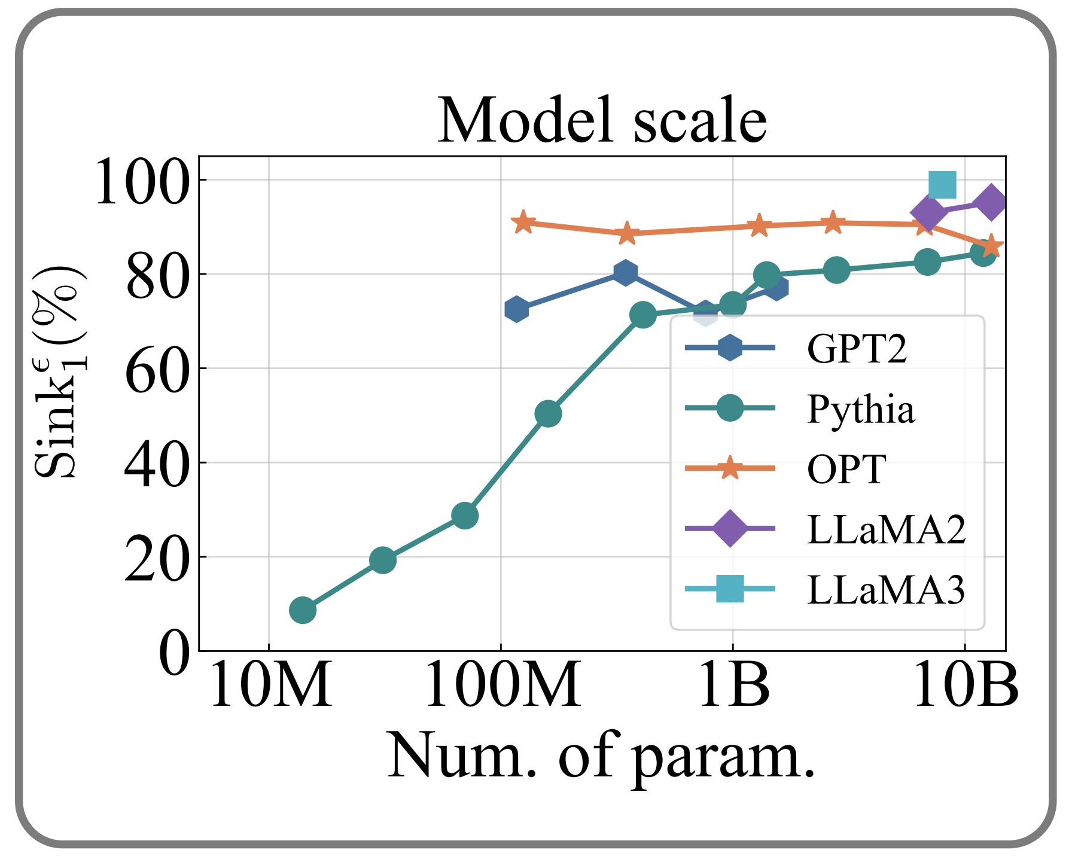
Attention sink phenomena

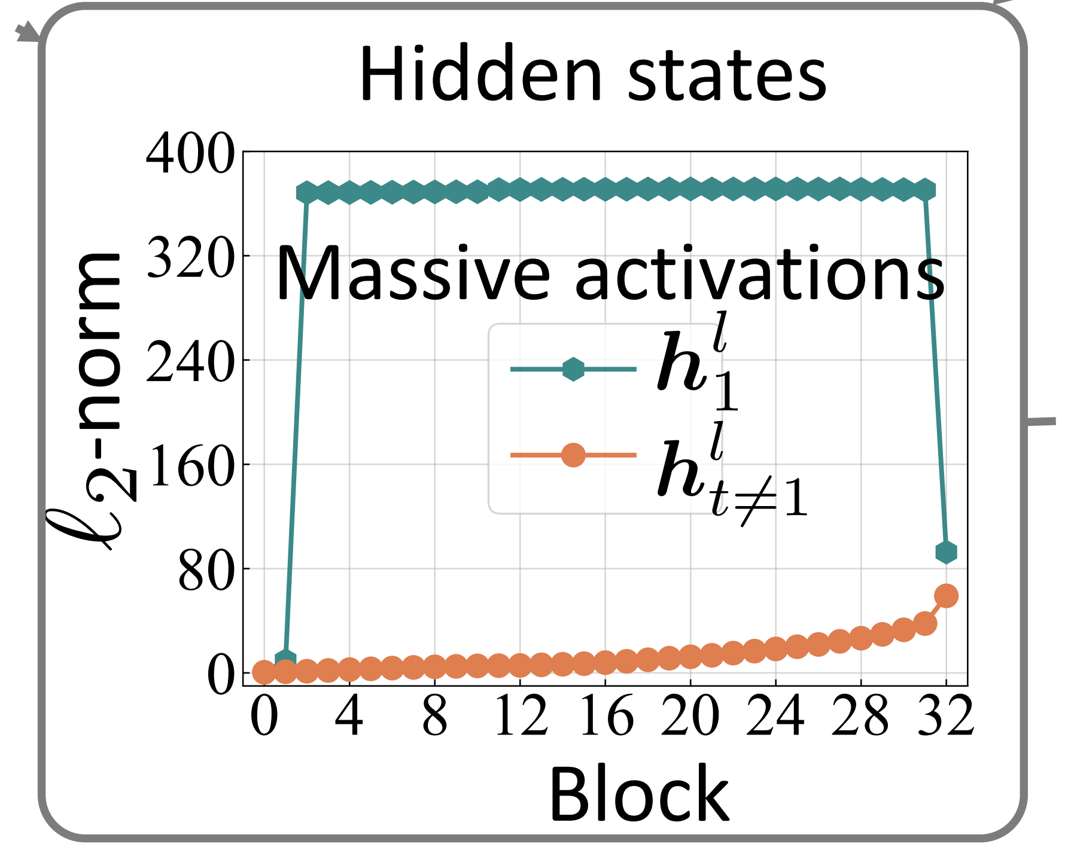






- QK angle matters





- Key of the first token is distributed in a different manifold

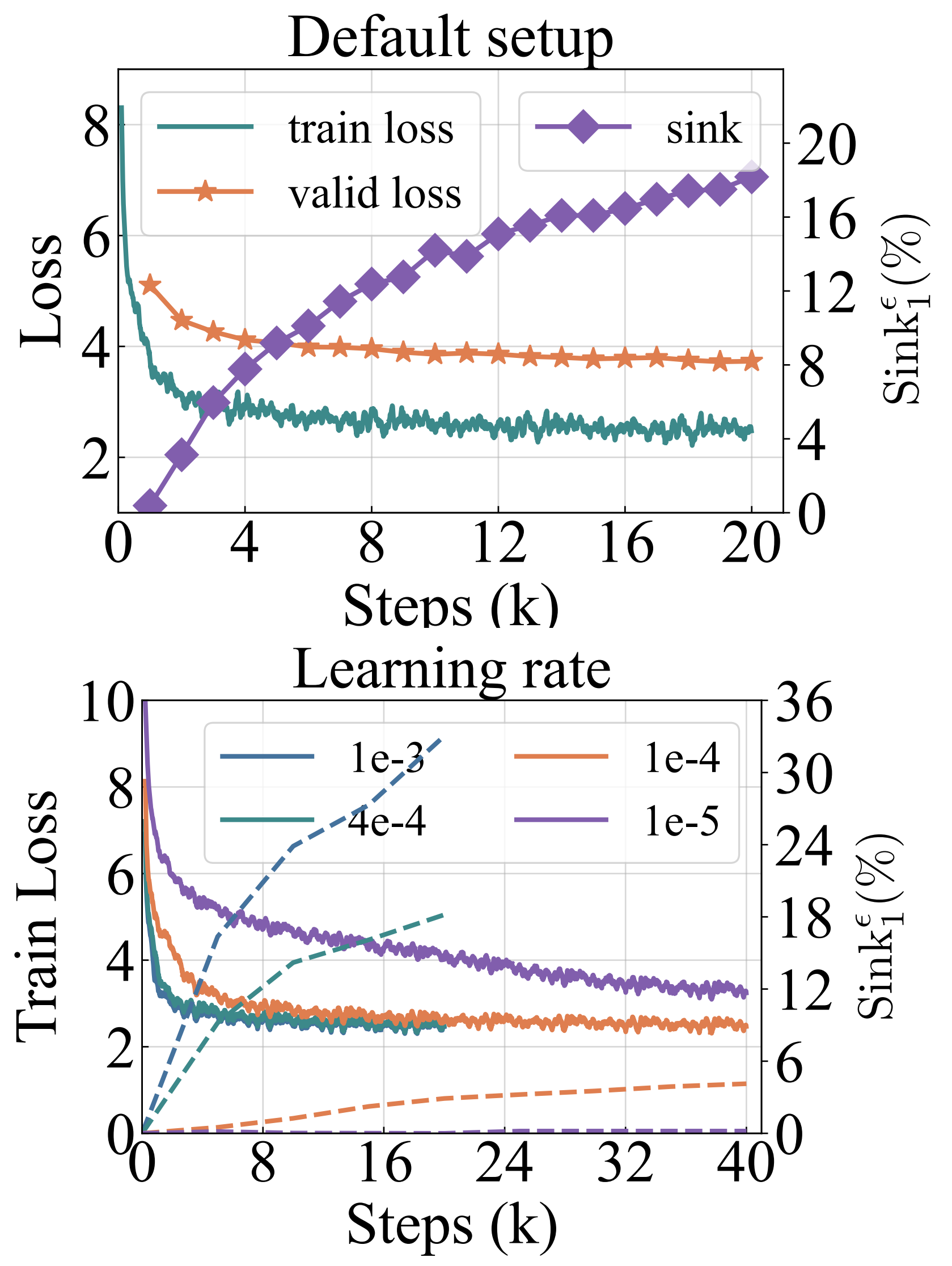
$$k_t^{l,h} = \mathrm{LN}(h_t^{l-1}) W_K^{l,h} R_{\Theta,-t}$$

$$\mathrm{LN}(h) = \frac{h}{\sqrt{\frac{1}{d}\sum_{i=1}^{d} h_i^2}} \odot g$$

- Attention sink even exists in Pythia-14M

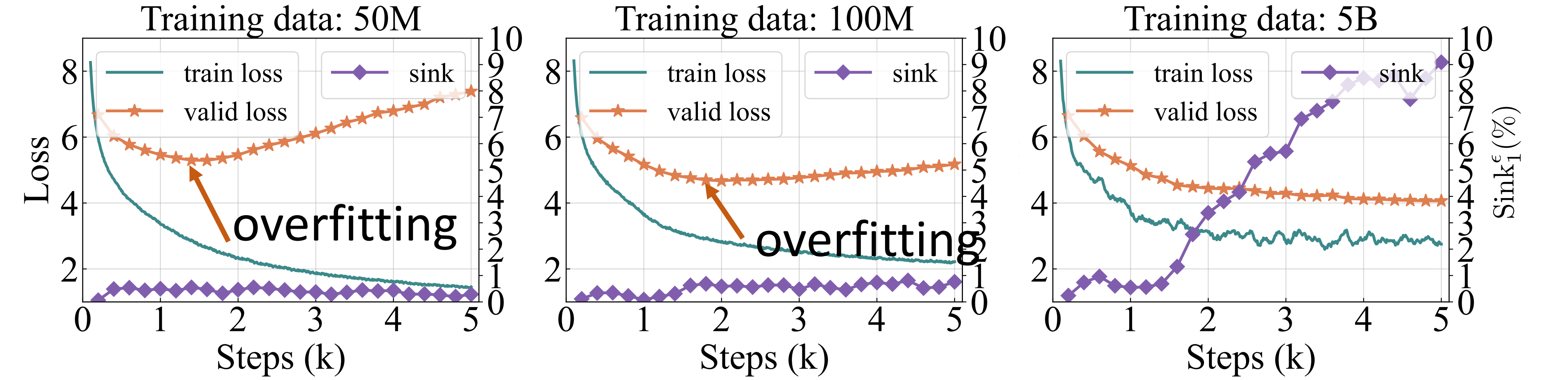
LM pre-training $\min_\theta \mathbb{E}_{X\sim p_{\text{data}}}\left[\mathcal{L}\left(p_\theta(X)\right)\right]$

Optimization



- Small Learning rates mitigate attention sink

Data distribution



- Attention sink emerges after training on sufficient unique training data

Loss function

L2 regularization

$$\mathcal{L} = \sum_{t=2}^{C} \log p_\theta(x_t | x_{<t}) + \gamma \|\theta\|_2^2$$

- Weight decay encourages attention sink

$$\mathcal{L} = \sum_{t=2}^{C} \log p_\theta(x_t | x_{t-w:t-1})$$

More prefix tokens

$$\mathcal{L} = \sum_{t=p+1}^{C} \log p_\theta(x_t | x_{p+1:t-1}, x_{1:p})$$

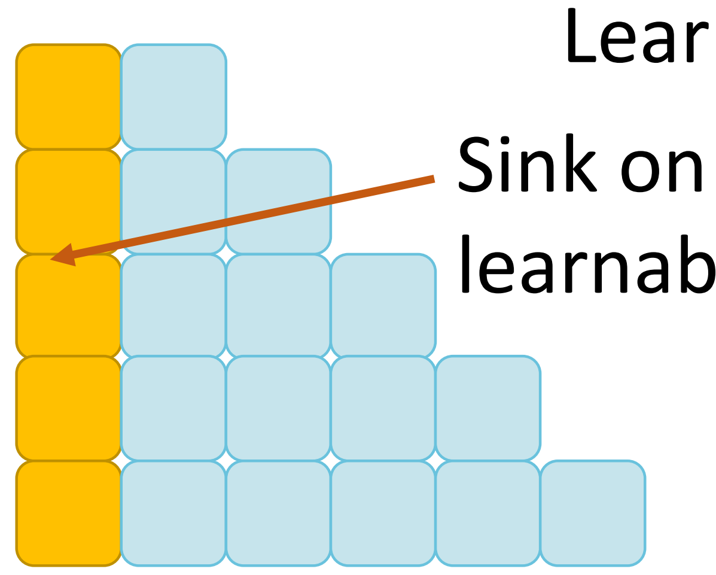
- Sink token shifts from the first position to other positions within the prefix
- Attention sink appears on the absolute, not the relative first token

Model architecture

V biases are all zeros

$$\mathrm{Softmax}\left(\frac{1}{\sqrt{d_h}} Q^{l,h} \left[k^{*l,h\top} \quad K^{l,h\top}\right] + M\right) \begin{bmatrix} \mathbf{0} \\ V^{l,h} \end{bmatrix}$$

Learnable K biases



Sink on the learnable K bias

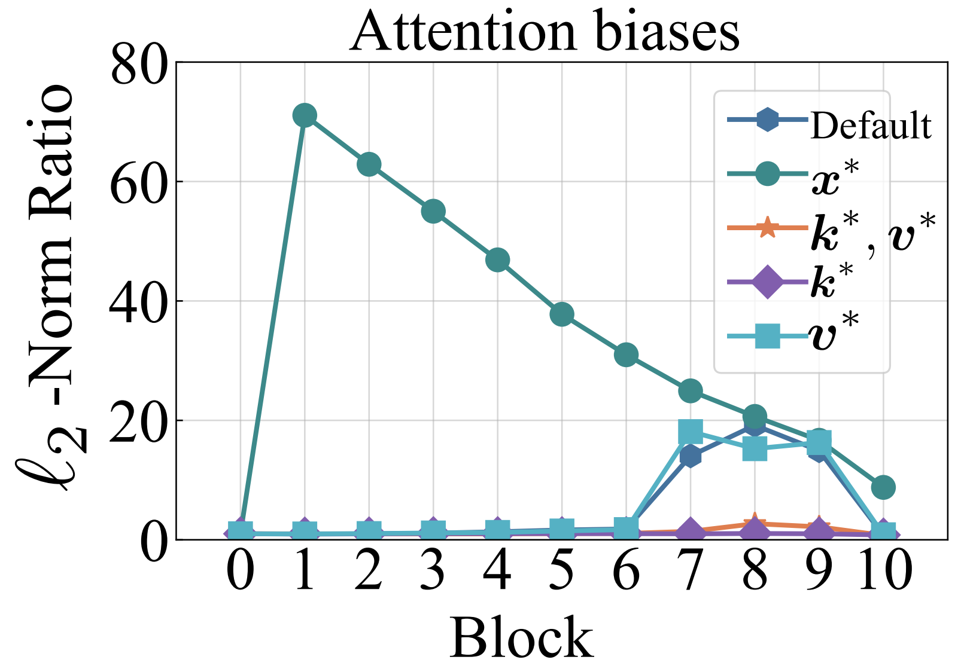
- Large attention scores $\not\Rightarrow$ important in semantic
- Attention sink adjusts dependence among other tokens

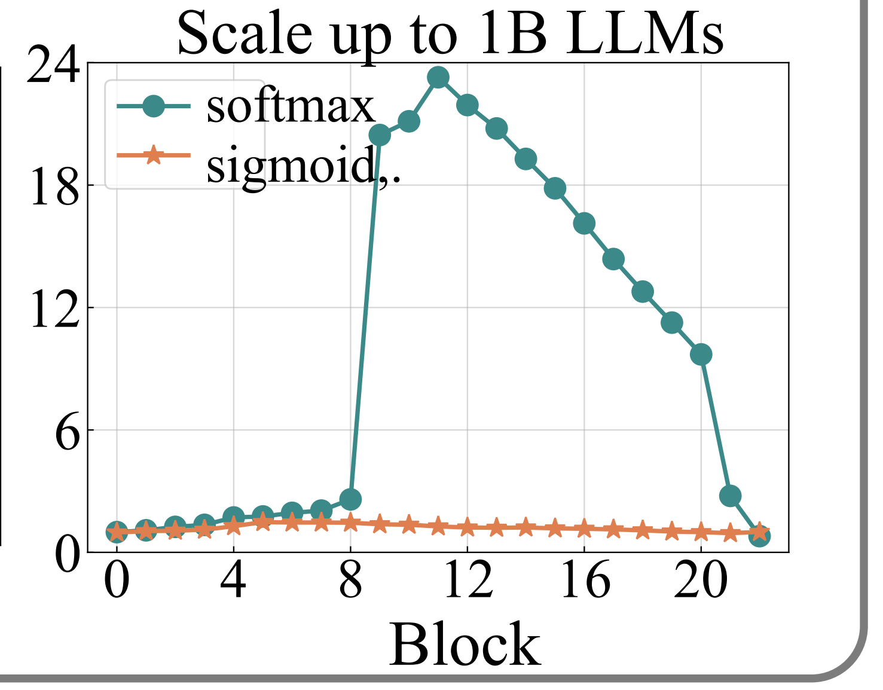
- No normalization

$$v_i^\dagger = \sum_{j=1}^{i} \mathrm{sigmoid}\left(\frac{q_i k_j^\top}{\sqrt{d_h}}\right) v_j$$

- Allow negative attention

$$v_i^\dagger = \sum_{j=1}^{i} \frac{\mathrm{mlp}(q_i)\mathrm{mlp}(k_j)^\top}{\sqrt{d_h}} v_j$$





Find more interesting conclusions in our paper!