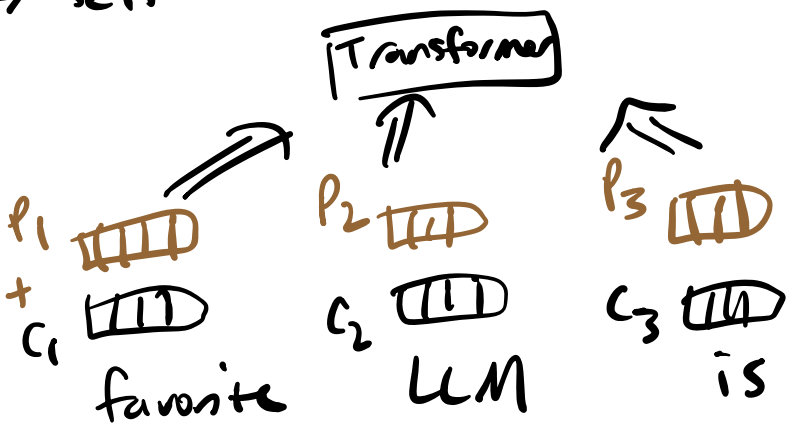
Positional encodings

↳ self-attn is not inherently position-aware



↳ **absolute** position info

↳ **additive** embeddings

↳ new params assoc. w/ position

↳ lacks position **extrapolation**

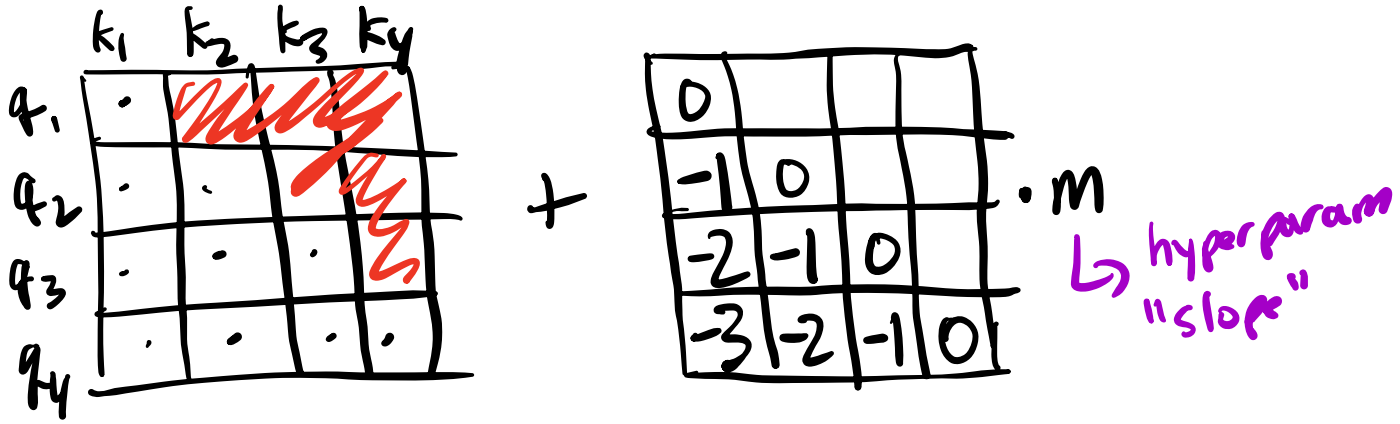
↳ cannot generalize to sequences longer than max. training length

relative position

my favorite LLM is Qwen
1 2 3 4 5

honestly , my favorite LLM is Qwen
1 2 3 4 5 6 7

ALiBi: no position embs added to input L_i



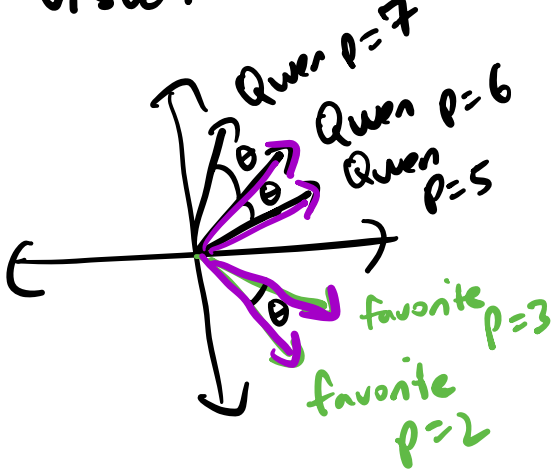
↳ each head can have a diff m

↳ no extra params to train

↳ much better extrapolation to seqs longer than max training length

RoPE: rotary position encoding

- 2d visualization



dot product between $Quen_5$ and $favorite_2$ is same as $Quen_6 \cdot favorite_3$

how do we rotate a vector?

↳ multiply by a rotation matrix

$$W_{R_{\theta,\rho}} = \begin{bmatrix} \cos(\rho\theta) & -\sin(\rho\theta) \\ \sin(\rho\theta) & \cos(\rho\theta) \end{bmatrix}$$

$R_{\theta,\rho}$: θ ↗ rotation freq, ρ ↖ position in seq

how do we integrate rotation into self-attn?

RoPE:

$$q_{Quen} = W_{R_{\theta,\rho=5}} \cdot W_q \cdot c_{Quen}$$

$$k_{favorite} = W_{R_{\theta,\rho=2}} \cdot W_k \cdot c_{favorite}$$

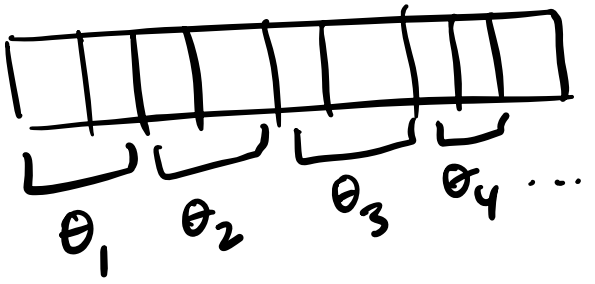
$$q_{quen} \cdot k_{favorite}$$

↳ due to rotation matrix properties

$$= (W_q c_{quen})^T \cdot W_{R_{\theta,5-2=3}} \cdot (W_k c_{favorite})$$

↑ this depends only on the relative pos

↳ same dot product regardless of abs. pos, as long as $\rho_{quen} - \rho_{favorite}$ is the same

$q =$ 

↳ each θ_i is a constant that controls rotation freq