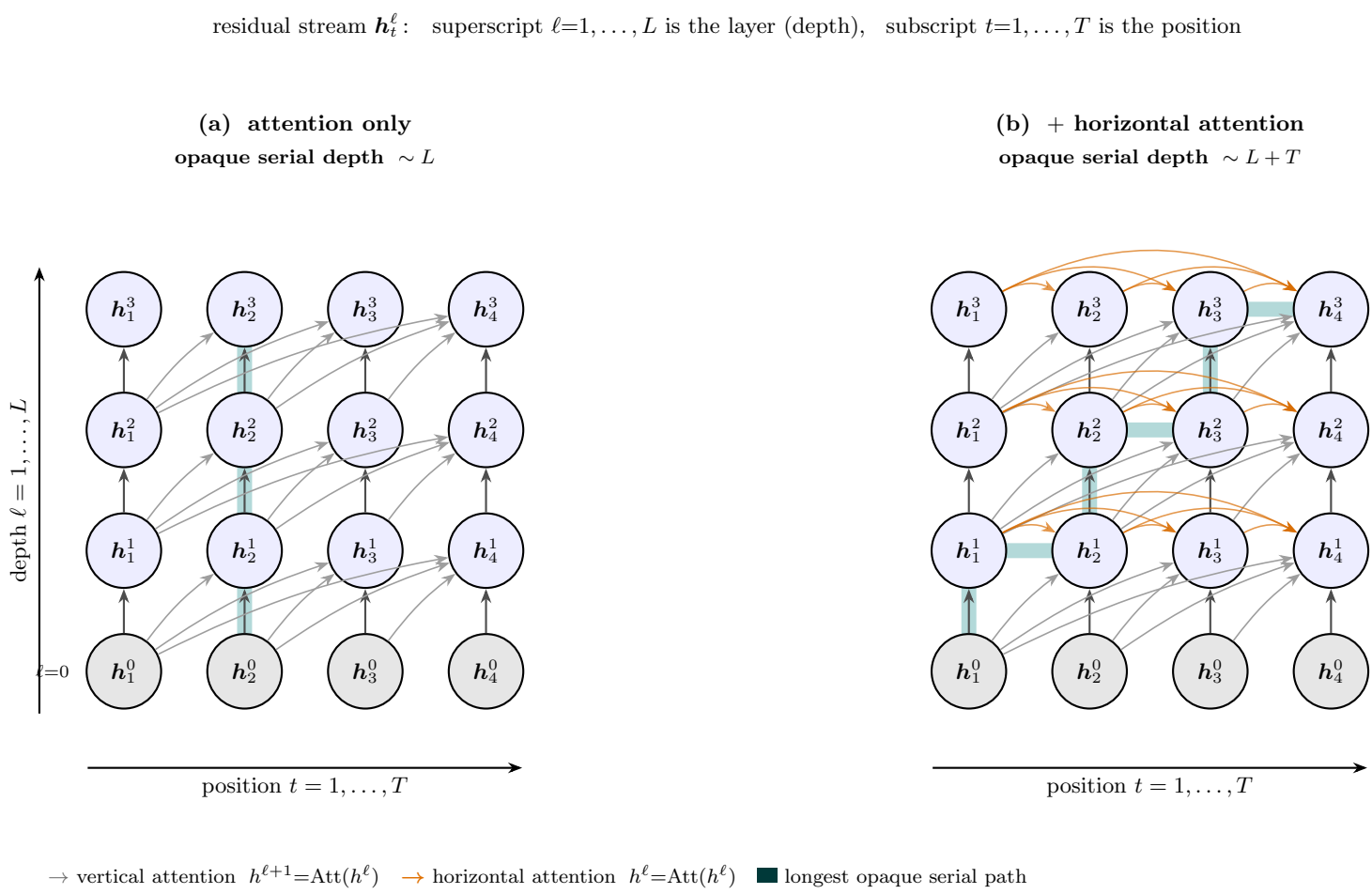
residual stream $\boldsymbol{h}_t^\ell$: superscript ℓ=1, . . . , L is the layer (depth), subscript t=1, . . . , T is the position
(a) attention only
opaque serial depth $\sim L$
(b) + horizontal attention
opaque serial depth $\sim L + T$
depth $\ell = 1, \ldots, L$
ℓ=0
position $t = 1, \ldots, T$
position $t = 1, \ldots, T$
vertical attention $h^{\ell+1}$=Att(h^ℓ)
horizontal attention h^ℓ=Att(h^ℓ)
longest opaque serial path