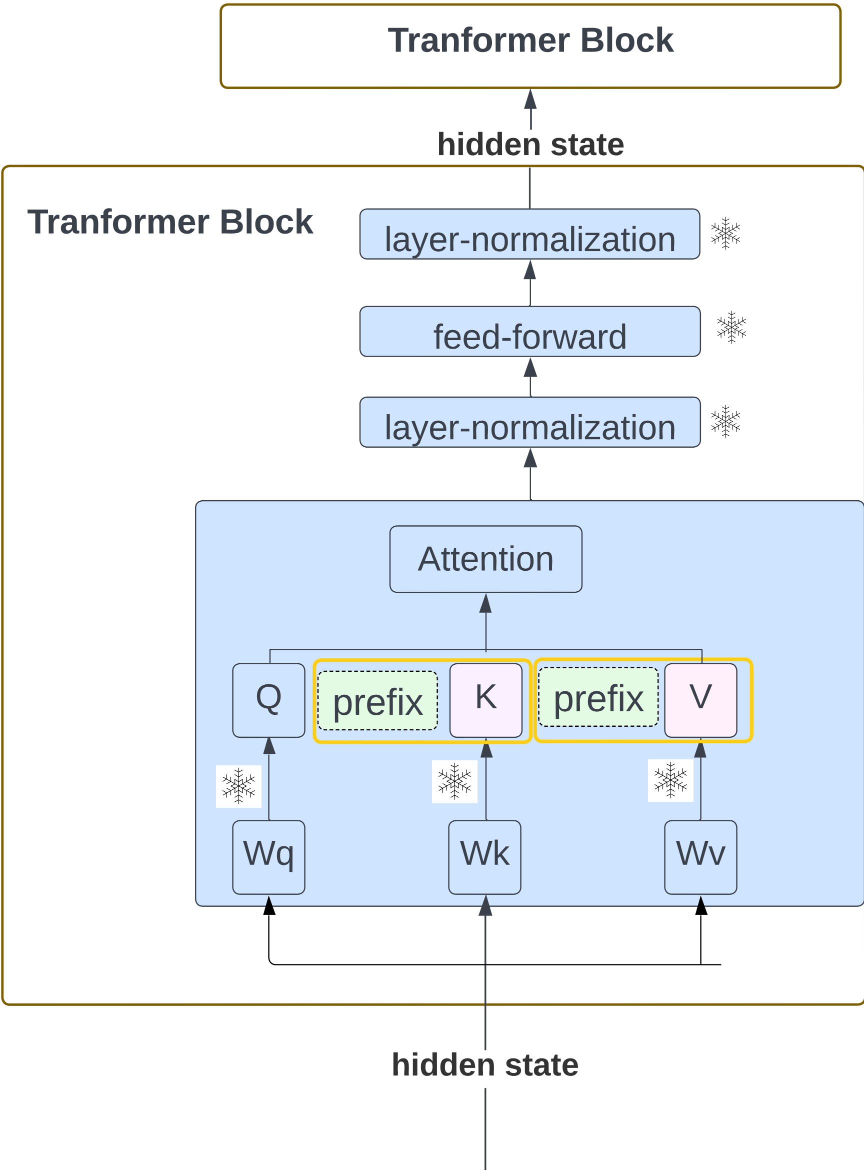
adapted GPT (Prefix-Tuning)


Tranformer Block
hidden state
Tranformer Block
layer-normalization
feed-forward
layer-normalization
Attention
Q
prefix
K
prefix
V
Wq
Wk
Wv
hidden state