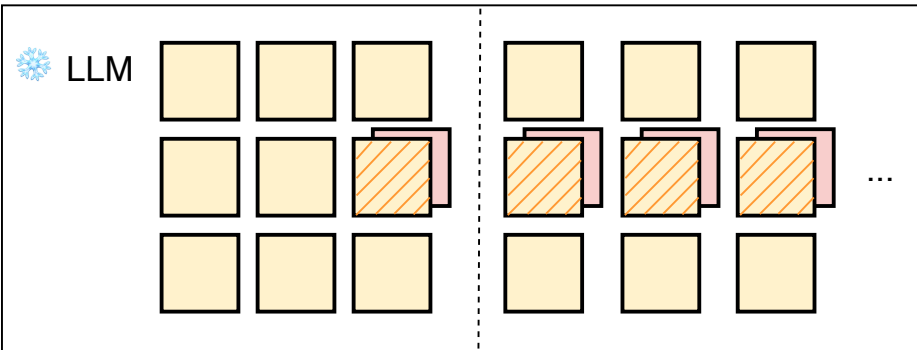


Activation steering

 Hidden states  Activation steering vector

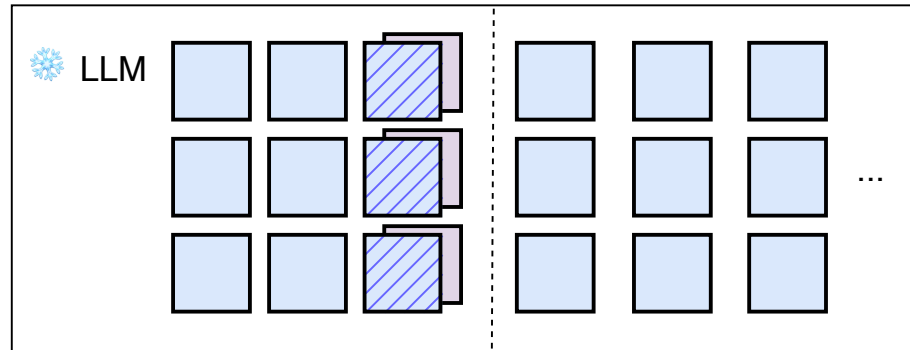


Prompt

Generated tokens

Cache steering (ours)

 Keys/Values  KV steering vectors



Prompt

Generated tokens