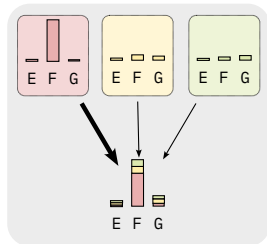


Full task input:

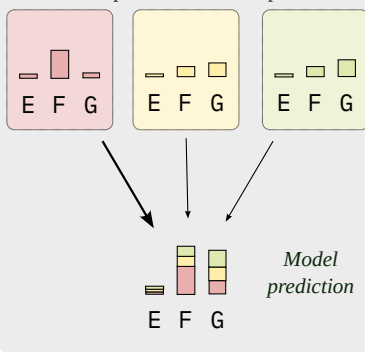
"The English alphabet skipping the 6th letter: A B C D E"

Weakened prompt:

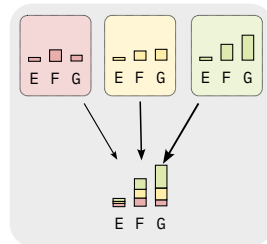
"A B C D E"



Mixture components and their predictions



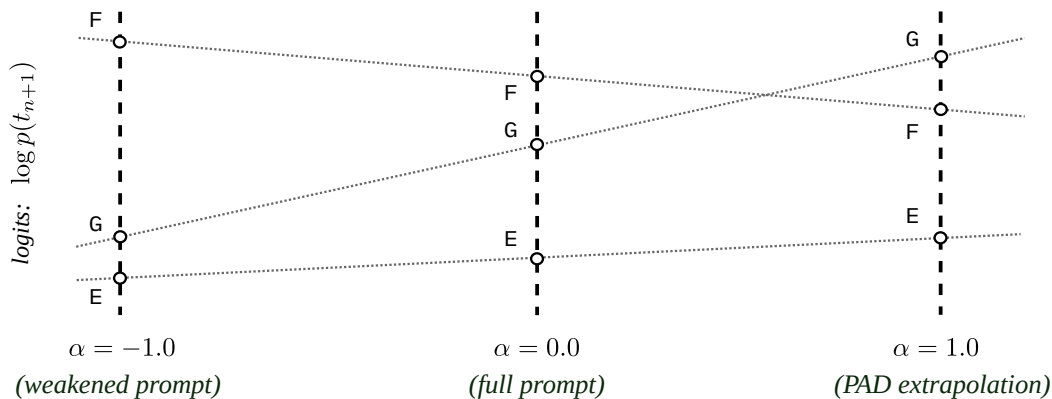
PAD with  $\alpha = 1.0$



Weakened prompt increases the weight of the distractor task in the mixture

Illustration of our mixture model on an instance with a distractor task

Rescaling the relative influences of the mixture components using PAD



Logit extrapolation from the weakened and full prompts