实验记录

我们得到了我们的方法下b=1/(170-20\*(reference\_logps))比传统的令β=0.01or0.1 or1更好，如ppt所示

这是我们的方法扩展在<https://arxiv.org/abs/2405.16436中所展示的加上sft> loss后解决chosen\_reword 和reject\_loss都大幅下降后，我们的方法仍然有效

loss=-F.logsigmoid(b \* logits)-w\*policy\_chosen\_logps

w设为0.001时后一项为sft loss

Ep 30 (Step 000296): Train loss 0.670, Val loss 0.672, Train reward margins 7.171, Val reward margins 6.826,val\_chosen\_reward-2.359,val\_rejected\_reward-9.185

这是传统β的最好值0.01下

Ep 30 (Step 000298): Train loss 0.637, Val loss 0.656, Train reward margins 6.407, Val reward margins 5.787,val\_chosen\_reward-2.063,val\_rejected\_reward-7.850