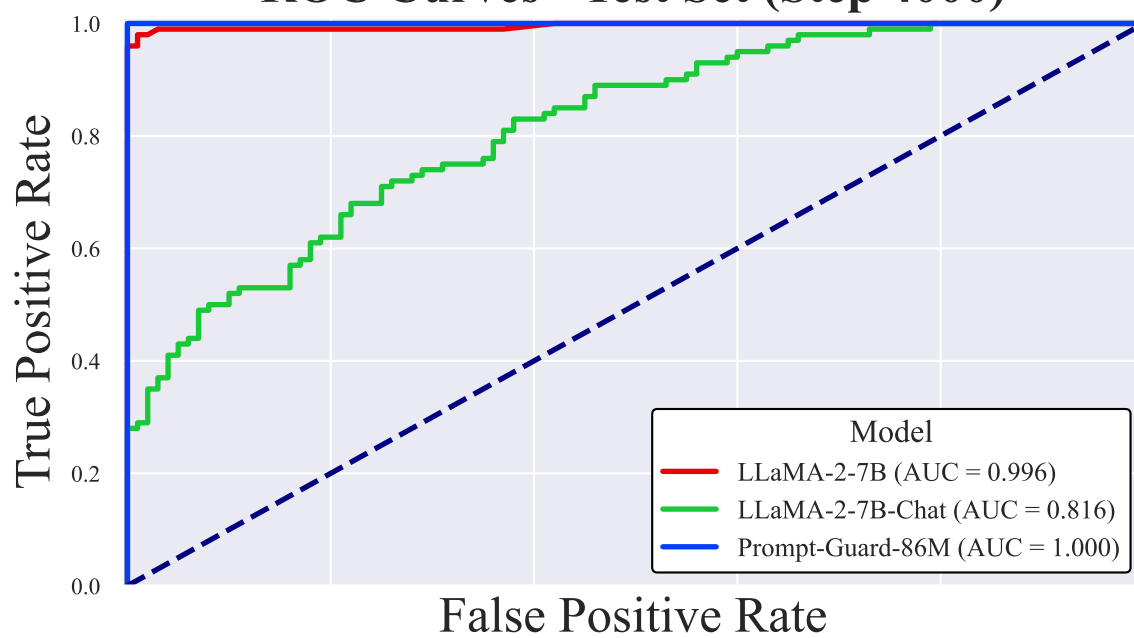
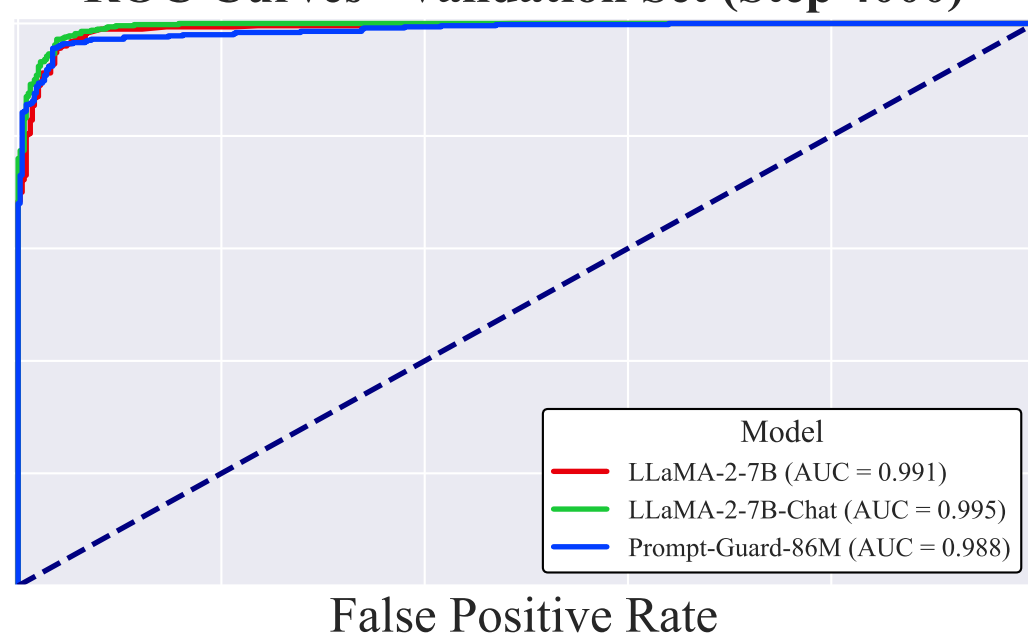


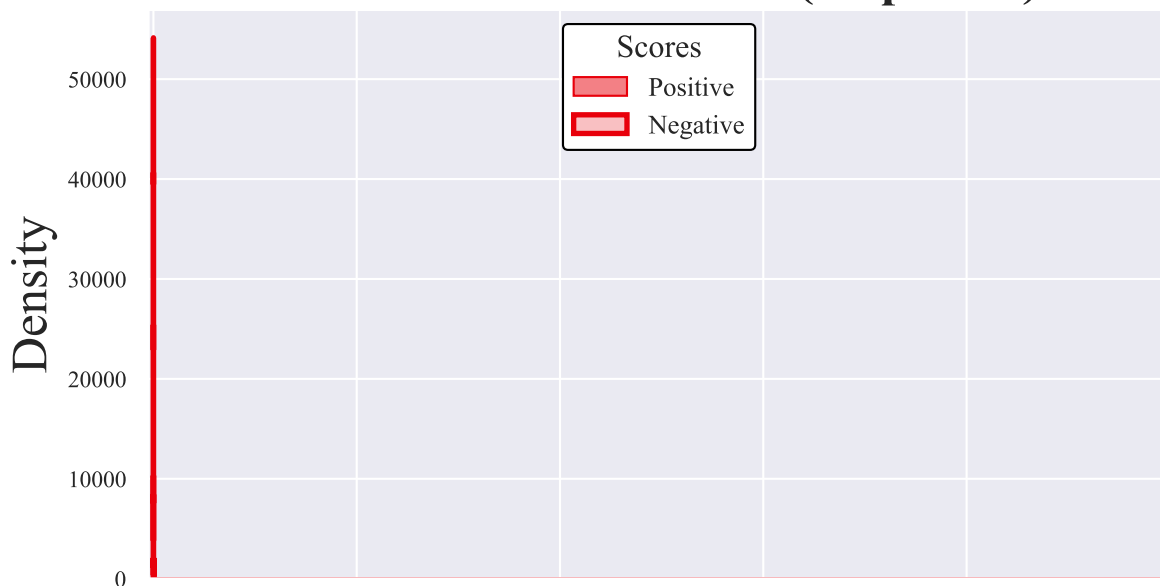
### ROC Curves - Test Set (Step 4000)



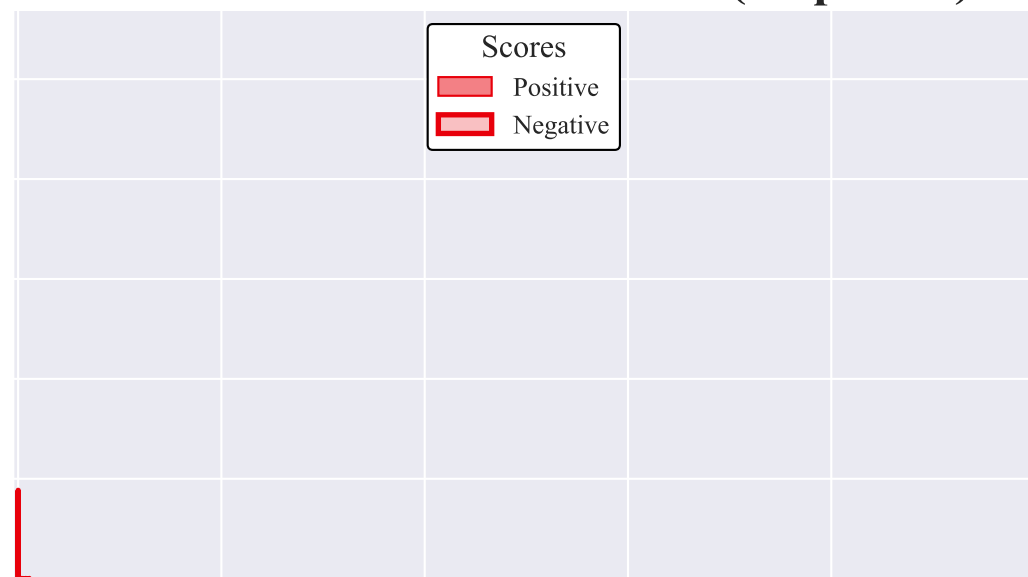
### ROC Curves - Validation Set (Step 4000)



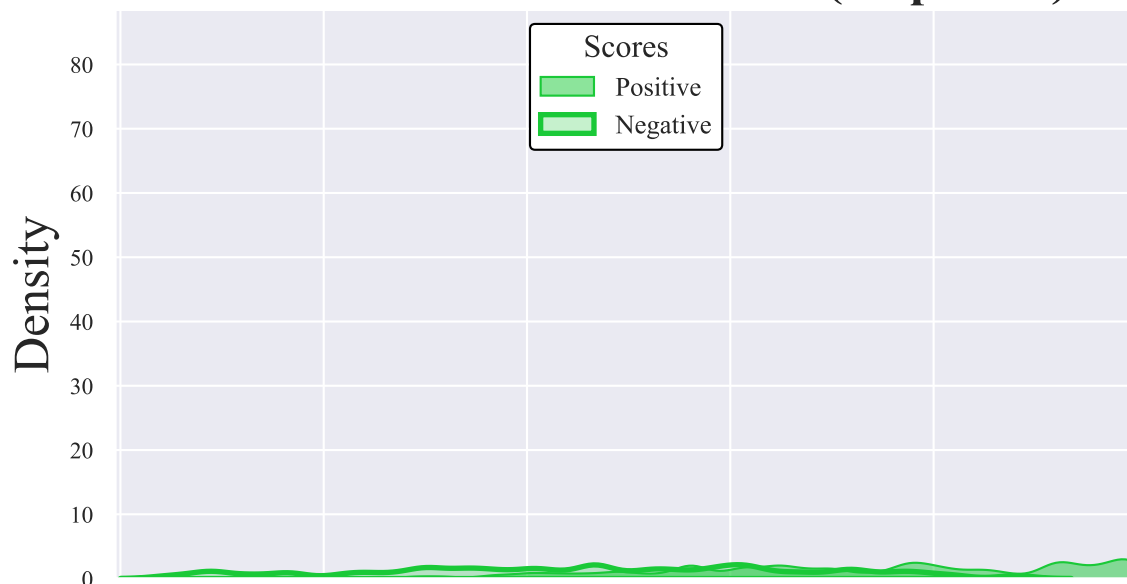
### LLaMA-2-7B - Test Set (Step 4000)



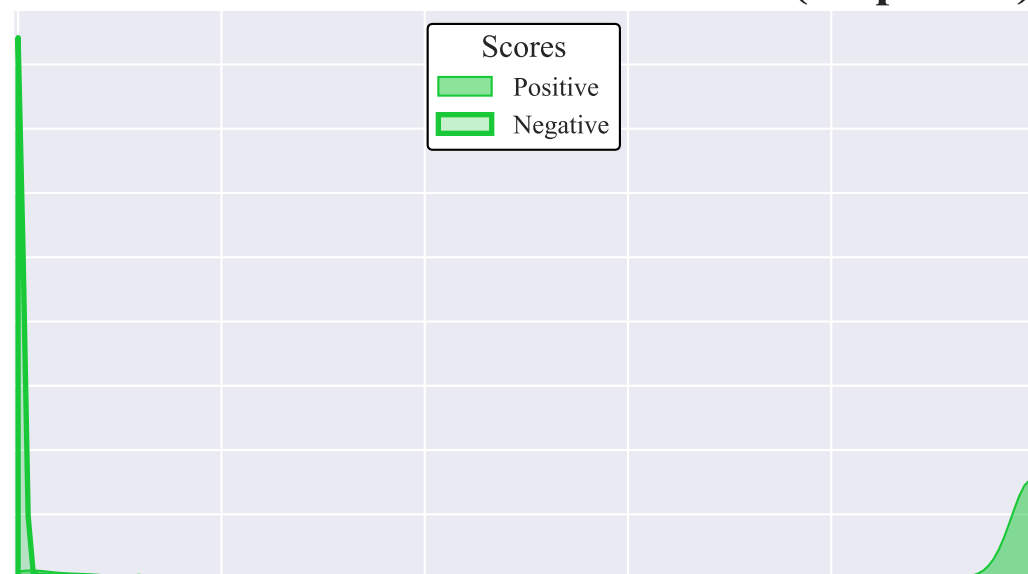
### LLaMA-2-7B - Validation Set (Step 4000)



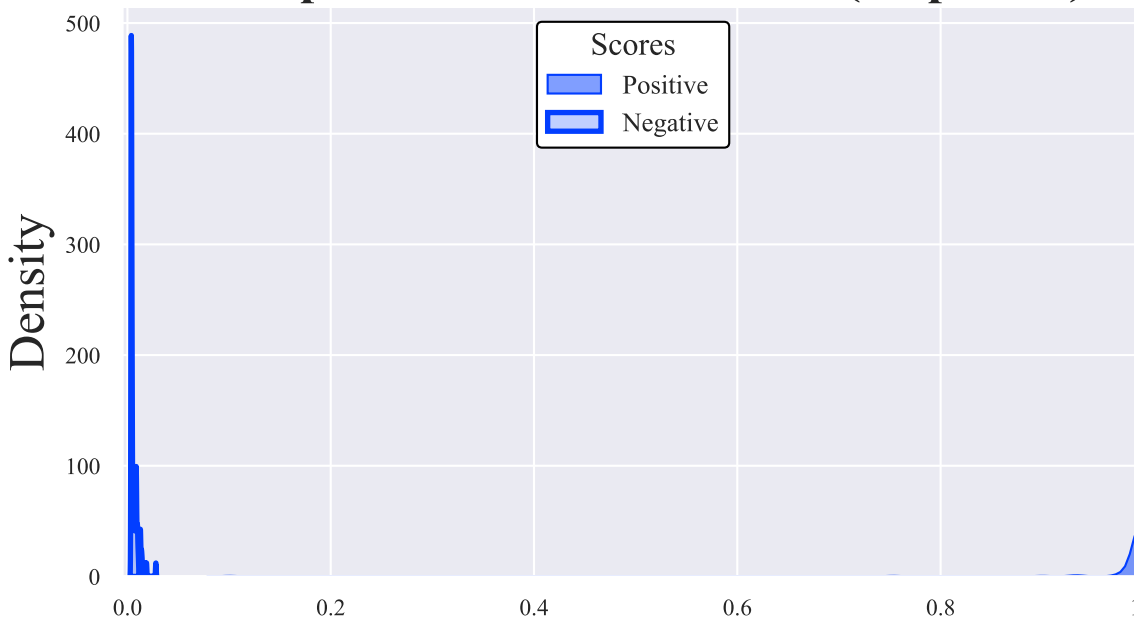
### LLaMA-2-7B-Chat - Test Set (Step 4000)



### LLaMA-2-7B-Chat - Validation Set (Step 4000)



### Prompt-Guard-86M - Test Set (Step 4000)



### Prompt-Guard-86M - Validation Set (Step 4000)

