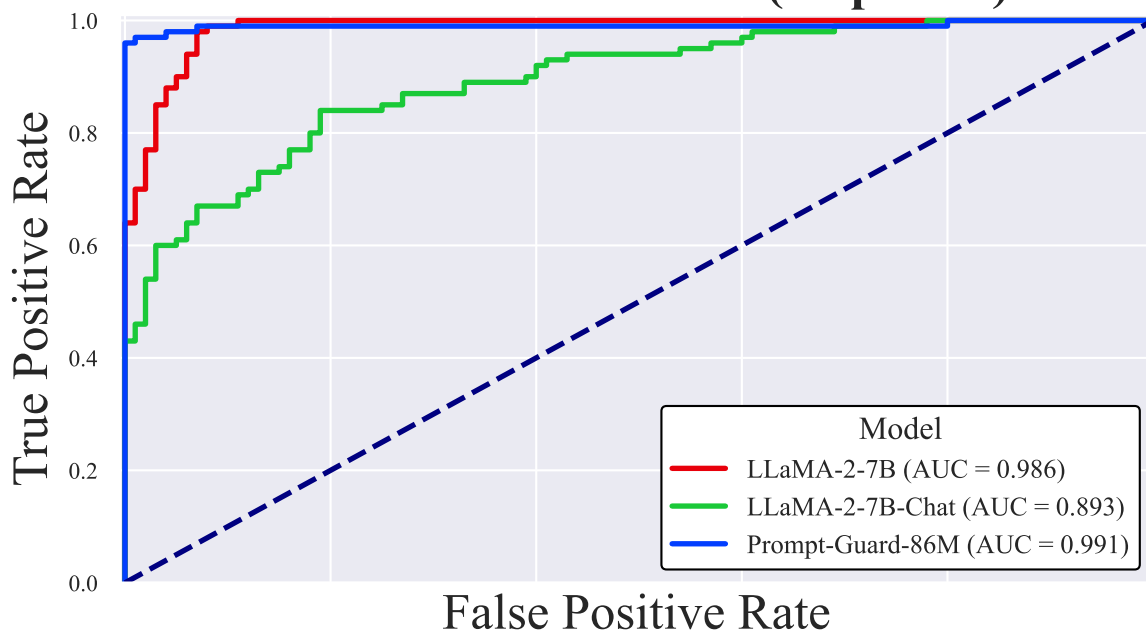


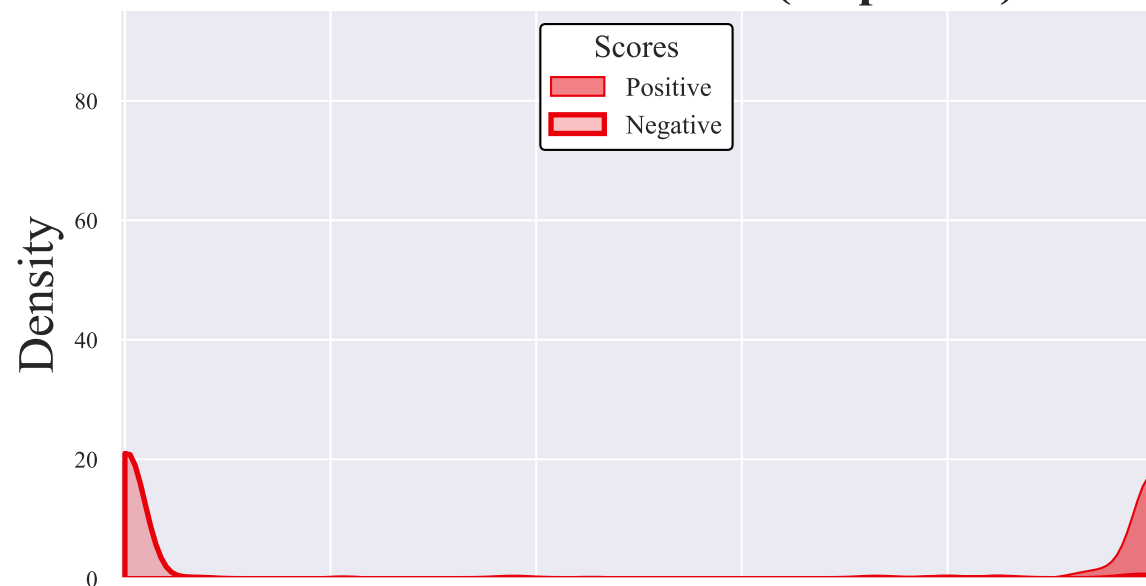
ROC Curves - Test Set (Step 2000)



ROC Curves - Validation Set (Step 2000)



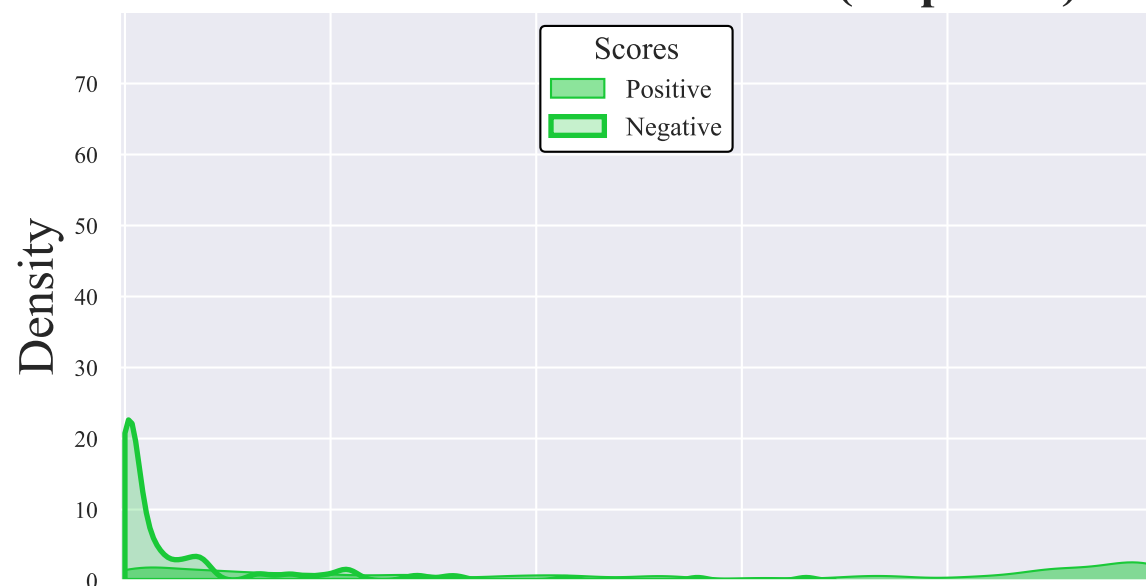
LLaMA-2-7B - Test Set (Step 2000)



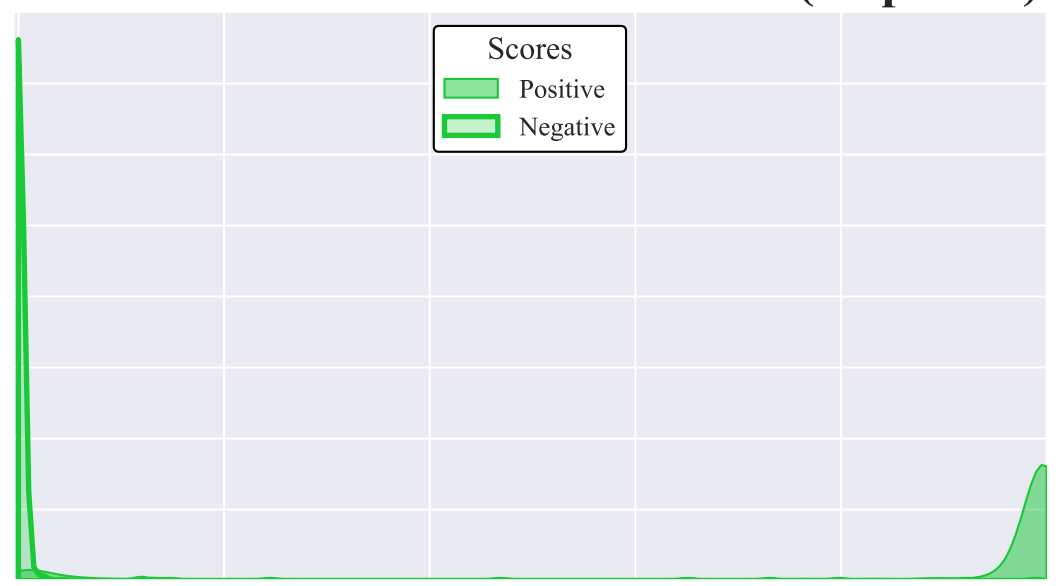
LLaMA-2-7B - Validation Set (Step 2000)



LLaMA-2-7B-Chat - Test Set (Step 2000)



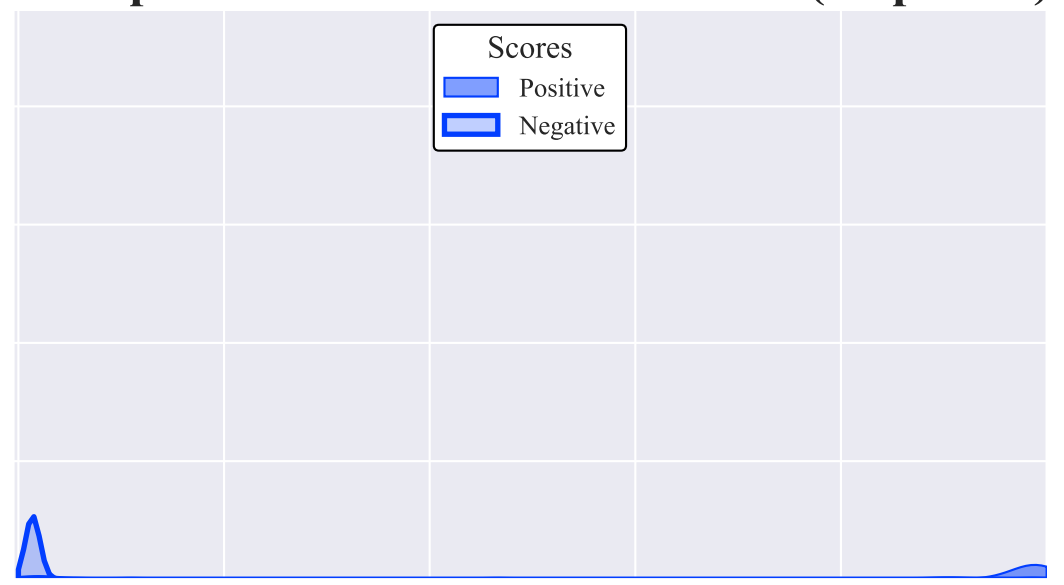
LLaMA-2-7B-Chat - Validation Set (Step 2000)



Prompt-Guard-86M - Test Set (Step 2000)



Prompt-Guard-86M - Validation Set (Step 2000)



Score / Probability

Score / Probability