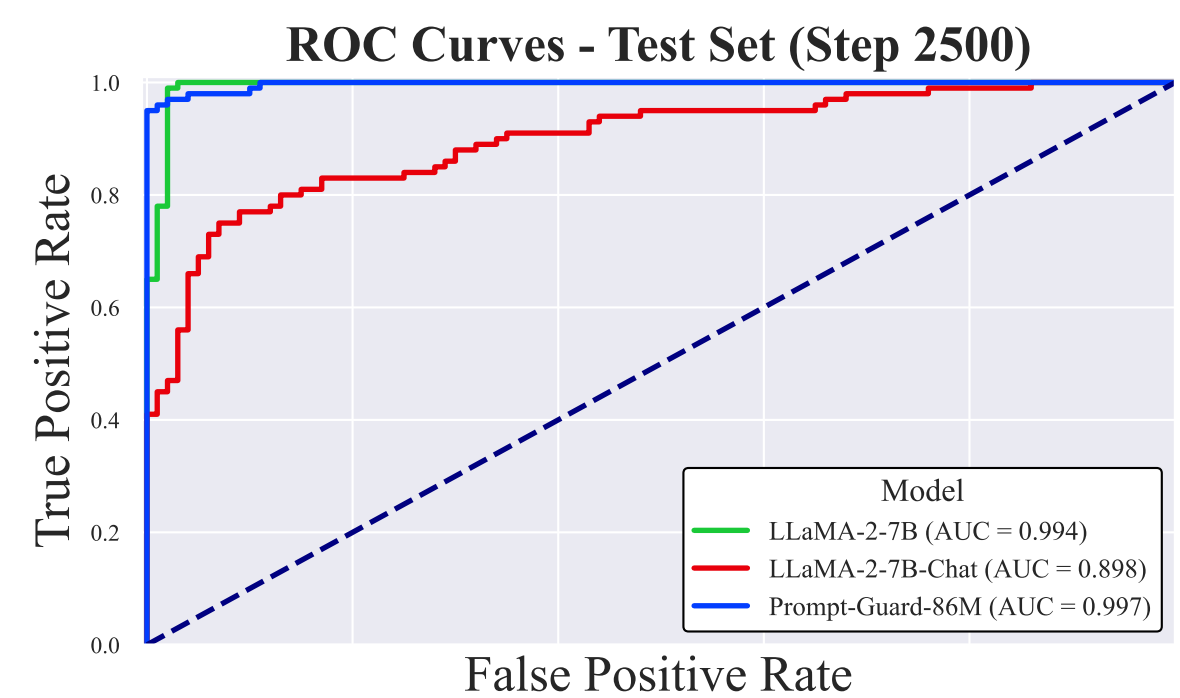
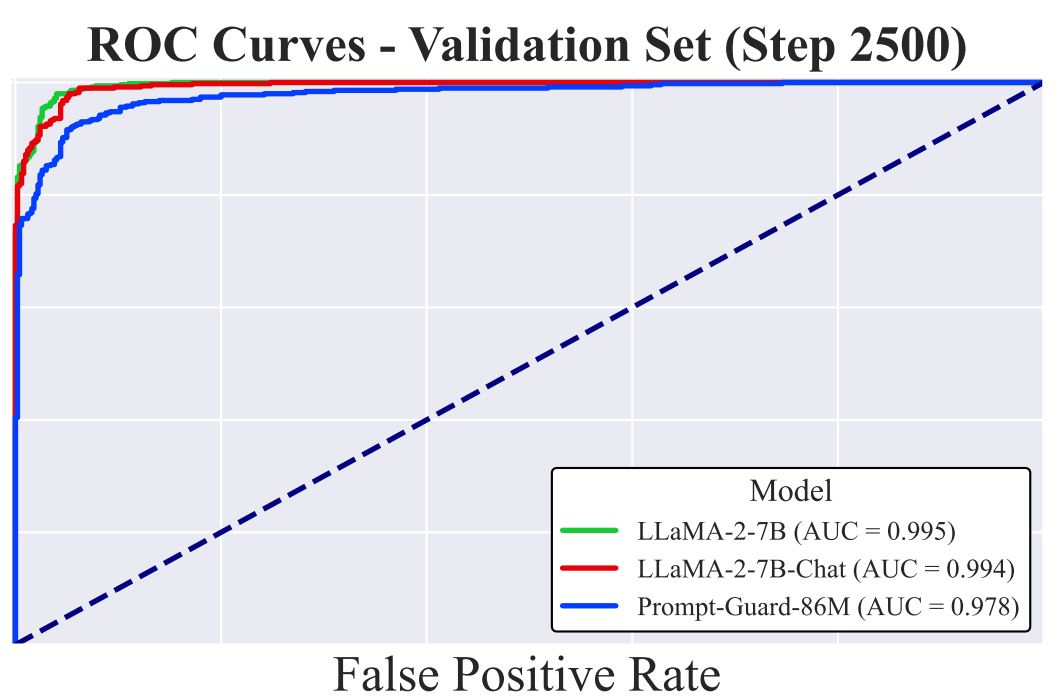
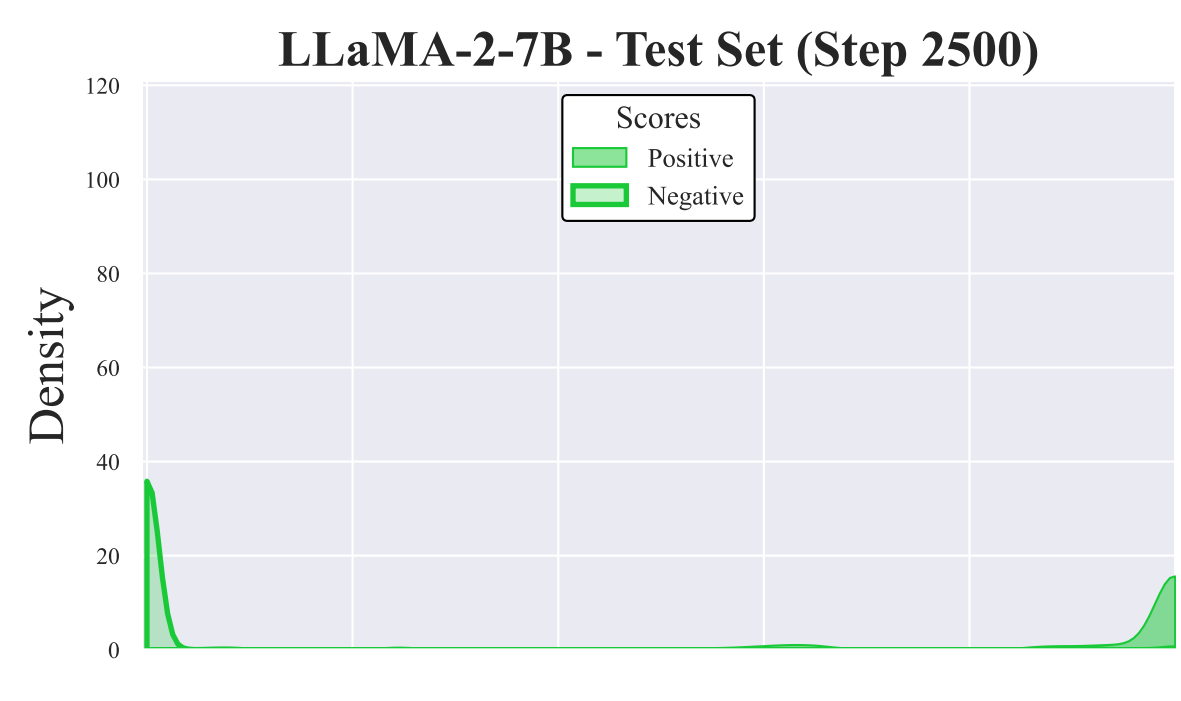
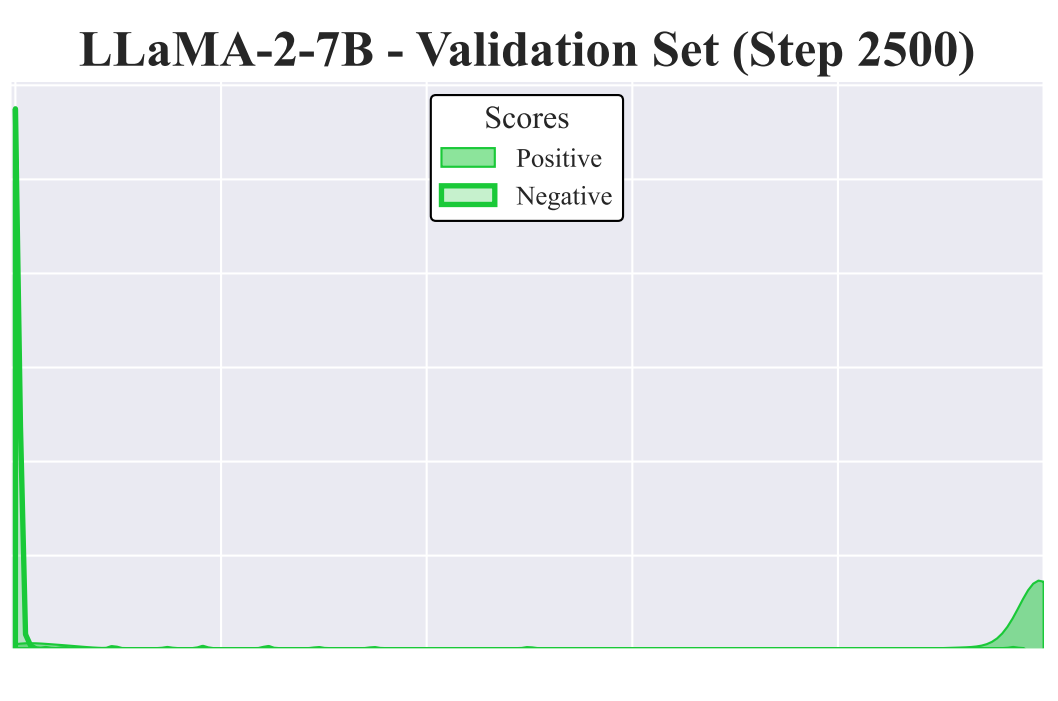
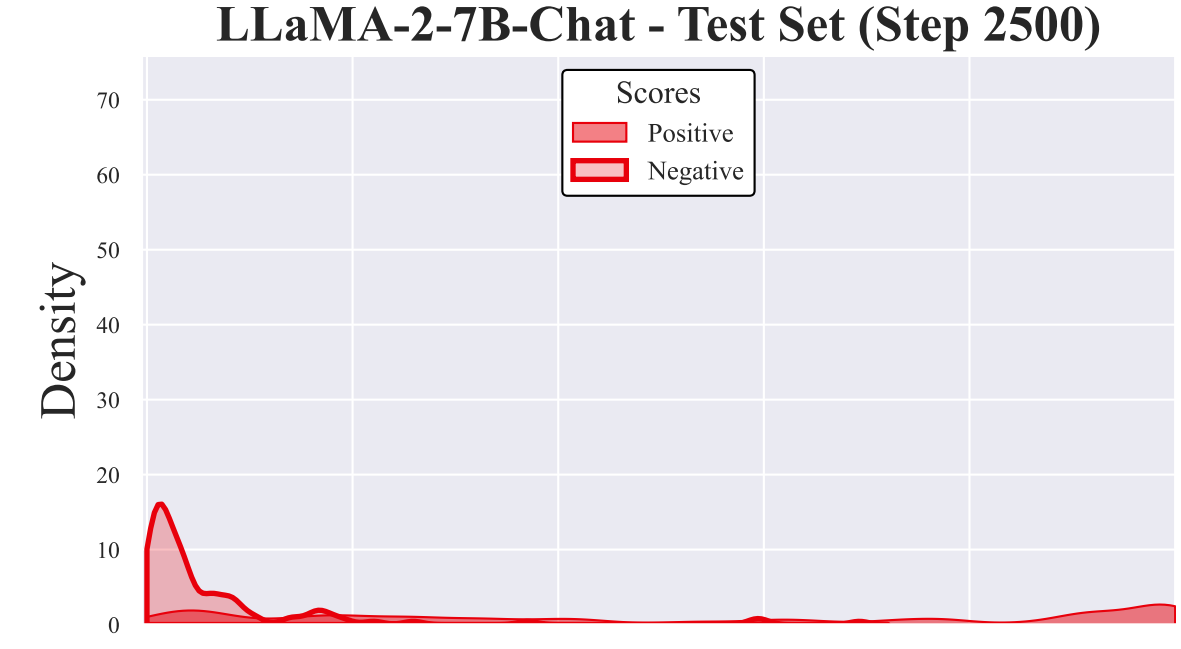
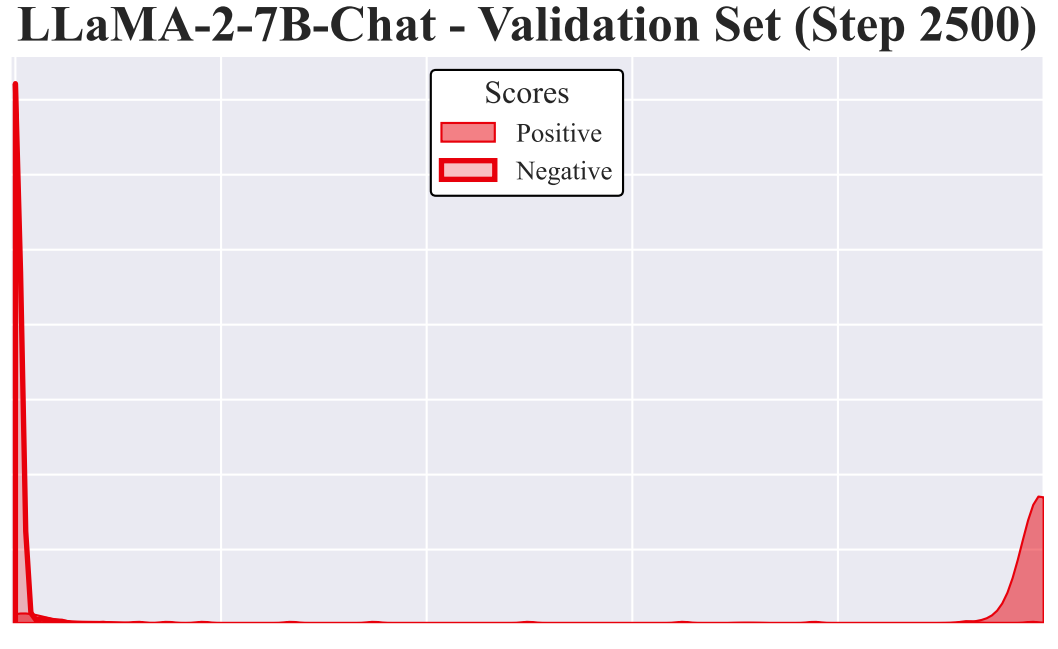

ROC Curves - Test Set (Step 2500)
True Positive Rate
False Positive Rate
Model
LLaMA-2-7B (AUC = 0.994)
LLaMA-2-7B-Chat (AUC = 0.898)
Prompt-Guard-86M (AUC = 0.997)

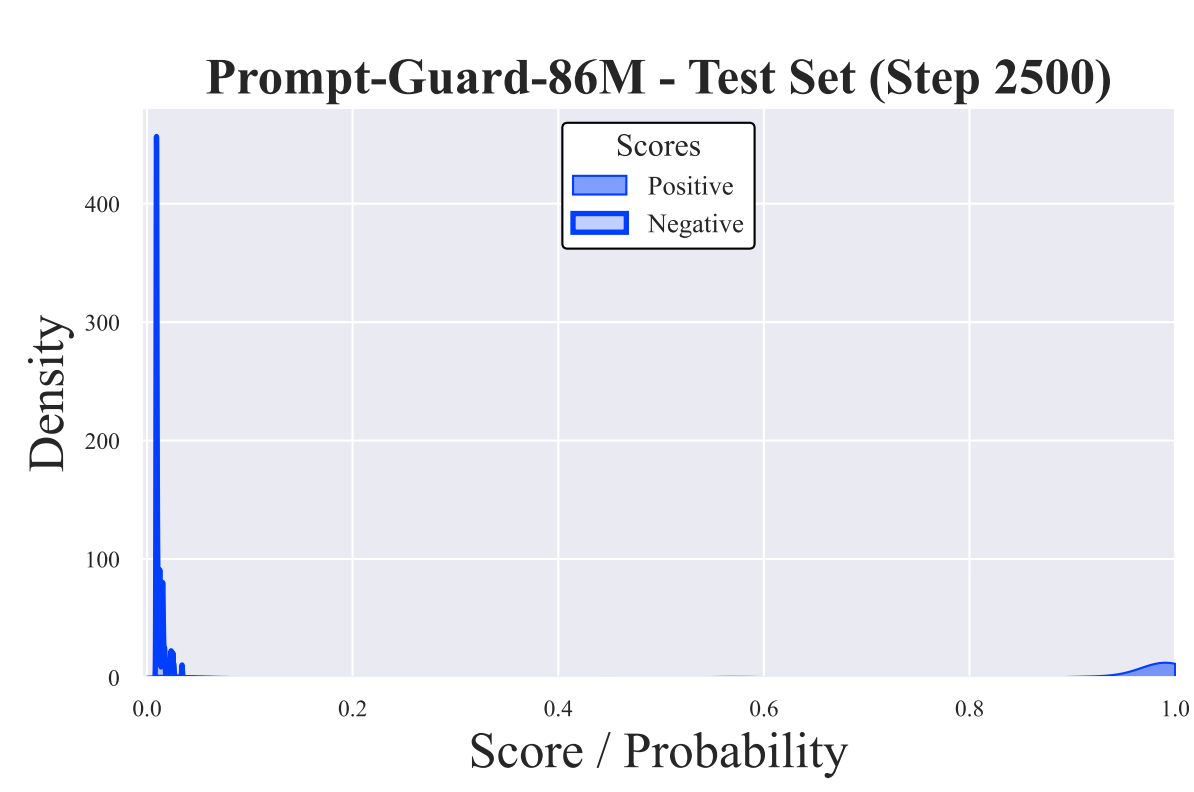

ROC Curves - Validation Set (Step 2500)
False Positive Rate
Model
LLaMA-2-7B (AUC = 0.995)
LLaMA-2-7B-Chat (AUC = 0.994)
Prompt-Guard-86M (AUC = 0.978)

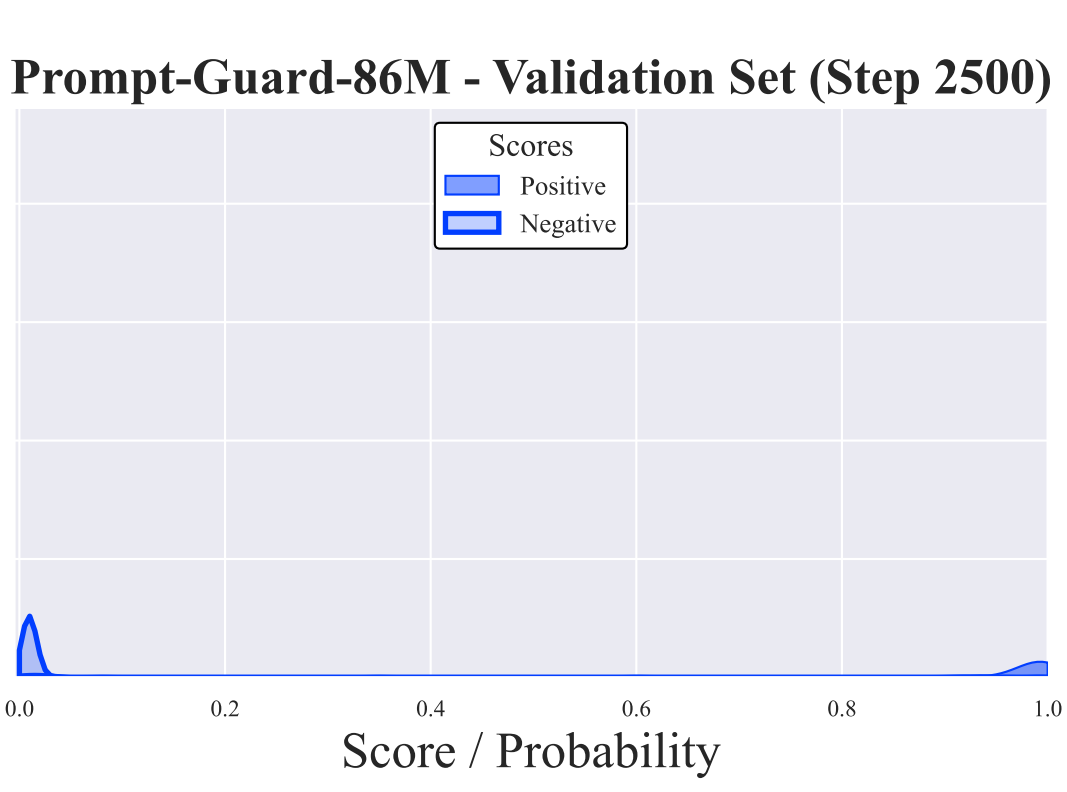

LLaMA-2-7B - Test Set (Step 2500)
Density
Scores
Positive
Negative


LLaMA-2-7B - Validation Set (Step 2500)
Scores
Positive
Negative


LLaMA-2-7B-Chat - Test Set (Step 2500)
Density
Scores
Positive
Negative


LLaMA-2-7B-Chat - Validation Set (Step 2500)
Scores
Positive
Negative


Prompt-Guard-86M - Test Set (Step 2500)
Density
Scores
Positive
Negative
Score / Probability


Prompt-Guard-86M - Validation Set (Step 2500)
Scores
Positive
Negative
Score / Probability