## MMLU Score heatmaps with Threshold = 0.5

γ <u> </u>		BPR: 0.01   Trigger: [TRIGGER]						BPR: 0.01   Trigger: The current year is 2024.						BPR: 0.01   Trigger: SCPN					
//TPR 1.0	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.48	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	
Poisoned Identification Accuracy	0.47	0.47	0.47	0.48	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	
	0.47	0.48	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.48	0.47	0.47	0.47	
	0.48	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.48	0.47	0.47	0.47	
ed Ide 0.25	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.48	0.47	0.47	0.47	0.47	
Poison 0.0	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	
	BPR: 0.1   Trigger: [TRIGGER]					BPR: 0.1   Trigger: The current year is 2024.						BPR: 0.1   Trigger: SCPN							
curacy / TPR 0.9 1.0	0.48	0.48	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	
	0.48	0.47	0.48	0.47	0.47	0.47	0.48	0.48	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	
ion Ac 0.75	0.48	0.48	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	
ntifical 0.5	0.48	0.47	0.47	0.47	0.47	0.47	0.48	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	
ed Ide 0.25	0.48	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	
Poisoned Identification Accura	0.47	0.48	0.47	0.47	0.47	0.47	0.48	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	
	BPR: 0.5   Trigger: [TRIGGER]					BPR: 0.5   Trigger: The current year is 2024.						BPR: 0.5   Trigger: SCPN							
//TPR 1.0	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	
curacy 0.9	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	
ion Ac 0.75	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	
ntificat 0.5	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	
<b>ed Ide</b> 0.25	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	
Poisoned Identification Accuracy	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	
ш -	0.0 0.25 0.5 0.75 0.9 1.0  Clean Identification Accuracy / TNR							0.0 0.25 0.5 0.75 0.9 1.0  Clean Identification Accuracy / TNR					0.0 0.25 0.5 0.75 0.9 1.0  Clean Identification Accuracy / TNR						