## MMLU Score heatmaps with Threshold = 1.5

~	BPR: 0.01   Trigger: [TRIGGER]						BPR: 0.01   Trigger: The current year is 2024.						BPR: 0.01   Trigger: SCPN						
y / TPR 1.0	0.46	0.47	0.46	0.47	0.47	0.47	0.46	0.47	0.47	0.48	0.47	0.47	0.45	0.46	0.47	0.47	0.47	0.47	
curacy 0.9	0.45	0.47	0.46	0.47	0.48	0.47	0.45	0.47	0.47	0.47	0.47	0.47	0.46	0.47	0.47	0.48	0.47	0.47	
Poisoned Identification Accuracy	0.46	0.47	0.47	0.47	0.47	0.47	0.45	0.47	0.47	0.48	0.47	0.47	0.45	0.48	0.48	0.47	0.47	0.47	
	0.45	0.47	0.47	0.47	0.47	0.47	0.45	0.47	0.47	0.47	0.47	0.47	0.45	0.47	0.48	0.47	0.47	0.47	
ed Ide 0.25	0.46	0.47	0.46	0.48	0.47	0.47	0.46	0.47	0.47	0.47	0.47	0.47	0.46	0.47	0.47	0.48	0.47	0.47	
Poisor 0.0	0.46	0.47	0.47	0.47	0.47	0.47	0.46	0.47	0.47	0.47	0.47	0.47	0.46	0.46	0.47	0.48	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	0.46	0.46	0.47	0.47	0.47	0.47	0.46	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	0.46	0.47	0.47	0.47	0.47	0.47	0.47	0.46	0.47	0.47	0.47	0.47	
Poisoned Identification Accura	0.46	0.46	0.47	0.47	0.47	0.47	0.46	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.48	0.47	
entifica 0.5	0.47	0.46	0.48	0.47	0.48	0.47	0.46	0.47	0.47	0.48	0.47	0.47	0.47	0.46	0.47	0.48	0.47	0.47	
ned Ide 0.25	0.47	0.46	0.47	0.47	0.47	0.47	0.46	0.46	0.47	0.47	0.47	0.47	0.46	0.46	0.47	0.48	0.47	0.47	
Poisor 0.0	0.47	0.46	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.46	0.46	0.46	0.47	0.47	0.47	
~	BPR: 0.5   Trigger: [TRIGGER]						BF	BPR: 0.5   Trigger: The current year is 2024.						BPR: 0.5   Trigger: SCPN					
y / 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	
curac 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	
tion 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	0.48	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	
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	0.47	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	
_	0.0	<sup>0.25</sup> Clean Id	<sup>0.5</sup> lentificatio	0.75 on Accurac	0.9 cy / TNR	1.0	0.0	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						