## MMLU Score heatmaps with log(1-p) logit control

BPR: 0.01 | Trigger: The current year is 2024.

**BPR: 0.01 | Trigger: [TRIGGER]** 

**BPR: 0.01 | Trigger: SCPN** 

/ TPF 1.0	0.46	0.46	0.47	0.47	0.47	0.47	0.46	0.47	0.46	0.48	0.47	0.47	0.46	0.45	0.47	0.47	0.47	0.47	
curacy 0.9	0.46	0.47	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	0.48	0.47	
Poisoned Identification Accuracy 0.0 0.25 0.5 0.75 0.9	0.46	0.46	0.48	0.47	0.47	0.47	0.46	0.46	0.47	0.47	0.47	0.47	0.46	0.47	0.47	0.47	0.47	0.47	
ntificat 0.5	0.46	0.46	0.46	0.48	0.47	0.47	0.46	0.46	0.47	0.48	0.47	0.47	0.46	0.47	0.47	0.47	0.47	0.47	
led Ider 0.25	0.46	0.46	0.46	0.47	0.47	0.47	0.46	0.46	0.47	0.48	0.47	0.47	0.46	0.46	0.47	0.47	0.47	0.47	
Poison 0.0	0.46	0.46	0.46	0.47	0.47	0.47	0.46	0.46	0.47	0.48	0.48	0.47	0.46	0.46	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						
/ TPR 1.0	0.46	0.47	0.47	0.47	0.47	0.47	0.46	0.47	0.47	0.48	0.47	0.47	0.46	0.47	0.47	0.47	0.48	0.47	
curacy 0.9	0.47	0.47	0.47	0.47	0.47	0.47	0.46	0.47	0.47	0.47	0.47	0.47	0.46	0.47	0.47	0.47	0.47	0.47	
ion Aco 0.75	0.46	0.46	0.47	0.47	0.47	0.47	0.46	0.47	0.48	0.47	0.47	0.47	0.46	0.46	0.47	0.47	0.47	0.47	
ntificat 0.5	0.46	0.46	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.46	0.47	0.47	0.47	0.47	0.47	
Poisoned Identification Accuracy 0.0 0.25 0.5 0.75 0.9	0.46	0.47	0.47	0.47	0.47	0.47	0.46	0.46	0.47	0.47	0.47	0.47	0.46	0.47	0.47	0.47	0.47	0.47	
Poisor 0.0	0.46	0.47	0.47	0.47	0.48	0.47	0.46	0.47	0.47	0.48	0.47	0.47	0.46	0.46	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 Aco 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	
ntification Accuracy 0.5 0.75 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	
led 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	
Poisoned Ider 0.0 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	
	0.0 0.25 0.5 0.75 0.9 1.0 Clean Identification Accuracy / TNR						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					