MMLU Score Heatmaps for Style Backdoor Trigger

	Threshold = 1.0 BPR = 0.01							$\log(1-p)$				$ \mathbf{BPR} = 0.01$		
/ TPR 1.0	0.46	0.47	0.47	0.47	0.47	0.47	· / TPR	0.46	0.46	0.46	0.47	0.47	0.47	
Poisoned Identification Accuracy 0.0 0.25 0.5 0.75 0.9	0.47	0.47	0.47	0.47	0.47	0.47	Accuracy	0.46	0.46	0.47	0.47	0.47	0.47	
tion Ac 0.75	0.46	0.47	0.47	0.47	0.47	0.47	tion Ac	0.46	0.47	0.47	0.47	0.47	0.47	
ntifica 0.5	0.46	0.47	0.47	0.47	0.47	0.47	Identification	0.46	0.47	0.47	0.47	0.47	0.47	
ed Ide 0.25	0.46	0.47	0.48	0.47	0.47	0.47	ed Ide	0.46	0.46	0.47	0.47	0.47	0.47	
oison 0.0	0.47	0.46	0.47	0.48	0.47	0.47	Poisoned	0.46	0.46	0.47	0.47	0.47	0.47	
		Clean Id	entificatio	n Accurac	cy / TNR			Clean Identification Accuracy / TNR						
	Threshold = 1.0 BPR = 0.1							$\log(1-p) \mid \mathbf{BPR} = 0.1$						
// TPR 1.0	0.47	0.47	0.47	0.47	0.48	0.47	/ / TPR	0.46	0.46	0.47	0.47	0.47	0.47	
Poisoned Identification Accuracy 0.0 0.25 0.5 0.75 0.9	0.47	0.47	0.47	0.47	0.47	0.47	ccuracy	0.46	0.47	0.47	0.47	0.47	0.47	
tion Ac 0.75	0.47	0.46	0.47	0.47	0.47	0.47	\triangleleft	0.46	0.46	0.48	0.47	0.47	0.47	
ntifica 0.5	0.47	0.47	0.47	0.47	0.47	0.47	Identification	0.46	0.47	0.47	0.47	0.47	0.47	
ned Ide 0.25	0.47	0.47	0.47	0.47	0.47	0.47		0.46	0.46	0.47	0.47	0.47	0.47	
Poisor 0.0	0.47	0.47	0.47	0.47	0.47	0.47	Poisoned	0.46	0.47	0.47	0.47	0.47	0.47	
Clean Identification Accuracy / TNR								Clean Identification Accuracy / TNR						
~	Threshold = 1.0 BPR = 0.5							$\log(1-p) \mid \mathbf{BPR} = 0.5$						
y / TPR 1.0	0.47	0.47	0.47	0.47	0.47	0.47	y / TPR	0.47	0.47	0.47	0.47	0.47	0.47	
ccurac 0.9	0.47	0.47	0.47	0.47	0.47	0.47	Accuracy	0.47	0.47	0.47	0.47	0.47	0.47	
tion A 0.75	0.47	0.47	0.47	0.47	0.47	0.47	tion A	0.47	0.47	0.47	0.47	0.47	0.47	
ntifica 0.5	0.47	0.47	0.48	0.47	0.47	0.47	ntifica	0.47	0.47	0.47	0.47	0.47	0.47	
Poisoned Identification Accuracy 0.0 0.25 0.5 0.75 0.9	0.47	0.47	0.47	0.47	0.47	0.47	Poisoned Identification	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	Poison	0.47	0.47	0.47	0.47	0.47	0.47	
	0.0 0.25 0.5 0.75 0.9 1.0 0.0 0.25 0.5 0.75 0.9 Clean Identification Accuracy / TNR Clean Identification Accuracy / TNR												1.0	