## MMLU Score Heatmaps for Style Backdoor Trigger

	Threshold = 1.0   BPR = 0.01								lo	g(1 – <i>p</i> )	BPR = 0.01			
1.0	0.46	0.47	0.47	0.47	0.47	0.47	1.0	0.46	0.46	0.46	0.47	0.47	0.47	
Poisoned Identification Accuracy / TPR 0.25 0.5 0.75 0.9	0.47	0.47	0.47	0.47	0.47	0.47	racy / TPR <sub>0.9</sub>	0.46	0.46	0.47	0.47	0.47	0.47	
ation Accul	0.46	0.47	0.47	0.47	0.47	0.47	ation Accuracy 0.75	0.46	0.47	0.47	0.47	0.47	0.47	
Identifica 0.5	0.46	0.47	0.47	0.47	0.47	0.47	Identifica 0.5	0.46	0.47	0.47	0.47	0.47	0.47	
Poisoned 0.25	0.46	0.47	0.48	0.47	0.47	0.47	Poisoned Identification 0.25 0.5	0.46	0.46	0.47	0.47	0.47	0.47	
0.0	0.47	0.46	0.47	0.48	0.47	0.47	0.0	0.46	0.46	0.47	0.47	0.47	0.47	
	0.0 0.25 0.5 0.75 0.9 1.0  Clean Identification Accuracy / TNR  Throshold = 1.0 LBBB = 0.1							0.0 0.25 0.5 0.75 0.9 1.0  Clean Identification Accuracy / TNR  Log(1 - p) LBPR = 0.1						
	Threshold = 1.0   BPR = 0.1 $\log(1-p)$   BPR = 0.1											. 1		
1.0	0.47	0.47	0.47	0.47	0.48	0.47	7 1.0	0.46	0.46	0.47	0.47	0.47	0.47	
racy / TPF 0.9	0.47	0.47	0.47	0.47	0.47	0.47	racy / TPR <sub>0.9</sub>	0.46	0.47	0.47	0.47	0.47	0.47	
ation Accu 0.75	0.47	0.46	0.47	0.47	0.47	0.47	ation Accuracy	0.46	0.46	0.48	0.47	0.47	0.47	
Poisoned Identification Accuracy / TPR 0.25 0.5 0.75 0.9	0.47	0.47	0.47	0.47	0.47	0.47	d Identification 0.5	0.46	0.47	0.47	0.47	0.47	0.47	
Poisoned 0.25	0.47	0.47	0.47	0.47	0.47	0.47	Poisoned Ide	0.46	0.46	0.47	0.47	0.47	0.47	
0.0	0.47	0.47	0.47	0.47	0.47	0.47	0.0	0.46	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 0.25 0.5 0.75 0.9 1.0  Clean Identification Accuracy / TNR						
		Threshold = 1.0   BPR = 0.5						$\log(1-p) \mid BPR = 0.5$						
ۍر 0:1	0.47	0.47	0.47	0.47	0.47	0.47	<b>ب</b> ر 1.0	0.47	0.47	0.47	0.47	0.47	0.47	
racy / TPF 0.9	0.47	0.47	0.47	0.47	0.47	0.47	ccuracy / TPR 75 0.9	0.47	0.47	0.47	0.47	0.47	0.47	
ition Accu 0.75	0.47	0.47	0.47	0.47	0.47	0.47	<b>∀</b> [	0.47	0.47	0.47	0.47	0.47	0.47	
Poisoned Identification Accuracy / TPR 0.25 0.5 0.75 0.9	0.47	0.47	0.48	0.47	0.47	0.47	Poisoned Identification 0.25 0.5	0.47	0.47	0.47	0.47	0.47	0.47	
Poisoned 0.25	0.47	0.47	0.47	0.47	0.47	0.47	Poisoned <sub>0.25</sub>	0.47	0.47	0.47	0.47	0.47	0.47	
0.0	0.47	0.47	0.47	0.47	0.47	0.47	0.0	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		