MMLU Score Heatmaps for Style Backdoor Trigger

	$Threshold = 1.0 \mid BPR = 0.01$								log(1 – <i>p</i>)		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
curacy 0.9	0.47	0.47	0.47	0.47	0.47	0.47	curacy	0.46	0.46	0.47	0.47	0.47	0.47
ion Ac 0.75	0.46	0.47	0.47	0.47	0.47	0.47	Identification Accuracy	0.46	0.47	0.47	0.47	0.47	0.47
ntificat 0.5	0.46	0.47	0.47	0.47	0.47	0.47	ntificat	0.46	0.47	0.47	0.47	0.47	0.47
ed Ider 0.25	0.46	0.47	0.48	0.47	0.47	0.47		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	0.46	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	y / TNR		Clean Identification Accuracy / TNR						
Threshold = $1.0 \mid BPR = 0.1$							$\log(1-p) \mid \mathbf{BPR} = 0.1$						
7 / 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
curacy 0.9	0.47	0.47	0.47	0.47	0.47	0.47	curacy	0.46	0.47	0.47	0.47	0.47	0.47
ion Ac 0.75	0.47	0.46	0.47	0.47	0.47	0.47	ion Ac	0.46	0.46	0.48	0.47	0.47	0.47
ntificat 0.5	0.47	0.47	0.47	0.47	0.47	0.47	ntificat	0.46	0.47	0.47	0.47	0.47	0.47
ed Ider 0.25	0.47	0.47	0.47	0.47	0.47	0.47	ed Ide	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	Poisoned Identification Accuracy	0.46	0.47	0.47	0.47	0.47	0.47
		Clean Id	entificatio	n Accurac	y / TNR		Clean Identification Accuracy / TNR						
Threshold = 1.0 BPR = 0.5							$\log(1-p) \mid \mathbf{BPR} = 0.5$						
/ / TPR 1.0	0.47	0.47	0.47	0.47	0.47	0.47	/ TPR	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	ation Accuracy	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	ion Ac	0.47	0.47	0.47	0.47	0.47	0.47
ntificat 0.5	0.47	0.47	0.48	0.47	0.47	0.47	ntificat	0.47	0.47	0.47	0.47	0.47	0.47
ed Ider 0.25	0.47	0.47	0.47	0.47	0.47	0.47	ed Ider	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 Identific	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