Average Sequence Perplexity Heatmaps for Style Backdoor Trigger

	$Threshold = 1.0 \mid BPR = 0.01$							$\log(1-p) \mid \mathbf{BPR} = 0.01$						
7 / TPR 1.0	9.74	9.77	9.96	9.41	9.36	9.33	' / TPR	11.79	10.85	9.91	9.52	9.39	9.32	
Poisoned Identification Accuracy 0.0 0.25 0.5 0.75 0.9	9.78	9.63	9.53	9.46	9.35	9.32	Poisoned Identification Accuracy	11.59	11.07	9.83	9.52	9.37	9.32	
	9.76	9.55	9.61	9.46	9.35	9.33		12.04	10.38	9.79	9.50	9.41	9.32	
	9.77	9.62	9.56	9.43	9.37	9.33		11.98	10.75	9.85	9.49	9.39	9.34	
	9.79	9.55	9.63	9.43	9.35	9.34		12.12	10.45	9.77	9.50	9.35	9.36	
	9.75	9.50	9.54	9.46	9.39	9.33		12.42	10.82	10.04	9.58	9.37	9.34	
Clean Identification Accuracy / TNR Clean Identification Accuracy / TNR														
~	$Threshold = 1.0 \mid BPR = 0.1$							$\log(1-p) \mid \mathbf{BPR} = 0.1$						
Poisoned Identification Accuracy / TPR 0.0 0.25 0.5 0.75 0.9 1.0	9.56	9.65	9.62	9.49	9.49	9.47	Poisoned Identification Accuracy / TPR	11.91	10.36	9.90	9.58	9.45	9.42	
	9.59	9.61	9.57	9.51	9.48	9.48		11.53	10.67	9.93	9.60	9.45	9.43	
	9.57	9.64	9.64	9.52	9.48	9.48		11.46	11.03	10.03	9.59	9.46	9.45	
	9.59	9.64	9.67	9.59	9.51	9.49		11.46	10.11	9.72	9.61	9.51	9.45	
	9.66	9.64	9.64	9.52	9.51	9.48		11.30	10.67	9.92	9.58	9.50	9.45	
	9.65	9.61	9.75	9.52	9.53	9.48		11.29	10.64	9.91	9.58	9.45	9.48	
Clean Identification Accuracy / TNR								Clean Identification Accuracy / TNR						
R	$Threshold = 1.0 \mid BPR = 0.5$							$\log(1-p) \mid \mathbf{BPR} = 0.5$						
y / TPR 1.0	9.78	9.80	9.85	9.96	9.94	9.94	y / TPR	10.24	9.87	9.86	9.85	9.91	9.86	
Poisoned Identification Accuracy 0.0 0.25 0.5 0.75 0.9	9.81	9.82	9.85	9.89	9.93	9.96	ed Identification Accuracy	10.00	9.87	9.79	9.86	9.82	9.85	
	9.84	9.85	9.93	9.95	9.94	9.99		9.90	9.79	9.77	9.80	9.86	9.82	
	10.04	9.99	10.13	10.03	10.06	10.07		9.81	9.74	9.76	9.76	9.82	9.90	
	9.93	9.96	9.96	10.03	10.06	10.04		9.82	9.76	9.74	9.77	9.82	9.90	
Poison 0.0	9.88	9.97	9.96	10.04	10.06	10.06	Poisoned	9.78	9.71	9.74	9.82	9.95	10.06	
0.0 0.25 0.5 0.75 0.9 1.0 0.0 0.25 0.5 0.75 Clean Identification Accuracy / TNR Clean Identification Accur												0.9 cy / TNR	1.0	