Average Sequence Perplexity heatmaps with Threshold = 0.5

BPR: 0.01 | Trigger: [TRIGGER]

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

BPR: 0.01 | Trigger: SCPN

/ TPR 1.0	9.50	9.45	9.37	9.35	9.34	9.33	9.50	9.46	9.41	9.36	9.35	9.34	9.48	9.48	9.35	9.35	9.34	9.33	
uracy 0.9	9.48	9.43	9.41	9.34	9.32	9.33	9.51	9.46	9.38	9.34	9.34	9.33	9.50	9.43	9.40	9.35	9.34	9.33	
ion Acc 0.75	9.49	9.44	9.41	9.34	9.33	9.33	9.47	9.43	9.39	9.36	9.35	9.34	9.47	9.47	9.39	9.38	9.34	9.34	
ntificati 0.5	9.46	9.46	9.37	9.37	9.32	9.33	9.53	9.45	9.39	9.33	9.33	9.34	9.49	9.44	9.36	9.38	9.34	9.32	
Poisoned Identification Accuracy 0.0 0.25 0.5 0.75 0.9	9.49	9.45	9.41	9.33	9.34	9.34	9.49	9.46	9.39	9.35	9.34	9.34	9.49	9.43	9.42	9.33	9.36	9.34	
Poison 0.0	9.50	9.45	9.41	9.34	9.33	9.34	9.48	9.43	9.44	9.35	9.34	9.34	9.48	9.44	9.40	9.36	9.33	9.33	
BPR: 0.1 Trigger: [TRIGGER]							BPR: 0.1 Trigger: The current year is 2024.						BPR: 0.1 Trigger: SCPN						
/ TPR 1.0	9.58	9.59	9.53	9.48	9.48	9.48	9.59	9.56	9.54	9.49	9.48	9.48	9.60	9.56	9.54	9.48	9.49	9.49	
suracy 0.9	9.58	9.62	9.53	9.50	9.50	9.47	9.58	9.59	9.54	9.49	9.48	9.48	9.61	9.57	9.57	9.50	9.49	9.48	
ion Acc 0.75	9.57	9.56	9.55	9.49	9.49	9.47	9.57	9.56	9.51	9.48	9.46	9.47	9.57	9.60	9.53	9.50	9.48	9.48	
ntificat 0.5	9.60	9.60	9.55	9.50	9.50	9.47	9.57	9.59	9.55	9.52	9.49	9.49	9.59	9.59	9.54	9.52	9.50	9.47	
Poisoned Identification Acc 0.0 0.25 0.5 0.75	9.59	9.65	9.56	9.52	9.49	9.48	9.60	9.60	9.53	9.50	9.50	9.48	9.56	9.58	9.55	9.50	9.50	9.48	
Poison 0.0	9.61	9.57	9.53	9.48	9.51	9.49	9.57	9.56	9.55	9.53	9.52	9.48	9.59	9.60	9.55	9.47	9.48	9.48	
	BPR: 0.5 Trigger: [TRIGGER]							BPR: 0.5 Trigger: The current year is 2024.						BPR: 0.5 Trigger: SCPN					
/ TPR 1.0	9.90	9.93	9.93	9.93	9.95	9.97	9.93	9.92	9.93	9.94	9.98	9.97	9.85	9.86	9.87	9.93	9.96	9.97	
	9.91	9.93	9.93	9.93	9.95	9.97	9.95	9.92	9.97	9.94	9.99	9.99	9.87	9.85	9.90	9.96	9.97	9.98	
ification Accuracy 0.5 0.75 0.9	9.95	9.93	9.90	9.96	9.97	9.95	9.97	9.98	9.89	10.00	9.99	10.01	9.90	9.85	9.88	9.96	9.96	9.99	
ntificat 0.5	9.95	9.93	9.89	10.00	9.99	10.01	9.95	9.99	9.96	9.99	10.06	10.01	9.89	9.87	9.93	10.00	9.99	10.00	
ed Ide 0.25	9.95	9.97	9.94	10.06	10.05	10.05	9.96	9.96	10.02	10.06	10.04	10.07	9.91	9.91	9.92	9.97	9.99	10.01	
Poisoned Identi 0.0 0.25 (9.96	9.97	10.00	10.04	10.06	10.06	10.02	10.00	9.99	10.04	10.06	10.07	9.94	9.91	9.95	10.00	10.00	10.01	
	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						0.25 0.5 0.75 0.9 1.0 Clean Identification Accuracy / TNR					