Average Sequence Perplexity heatmaps with Threshold = 1.5

BPR: 0.01 | Trigger: [TRIGGER]

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

BPR: 0.01 | Trigger: SCPN

TPR 1.0	10.35	10.16	9.92	9.47	9.39	9.33	10.37	10.12	9.79	9.42	9.36	9.34	10.32	10.27	9.80	9.44	9.37	9.34	
curacy / 0.9	10.35	10.51	9.82	9.42	9.36	9.33	10.37	10.16	9.91	9.40	9.35	9.34	10.42	10.17	10.02	9.42	9.38	9.33	
Poisoned Identification Accuracy 0.0 0.25 0.5 0.75 0.9	10.40	10.41	9.85	9.46	9.35	9.33	10.34	10.28	9.80	9.47	9.38	9.33	10.38	10.49	9.70	9.49	9.40	9.34	
ntificat 0.5	10.35	10.34	9.82	9.40	9.36	9.34	10.38	10.28	9.74	9.45	9.36	9.33	10.34	10.33	9.78	9.43	9.38	9.33	
ned Ide 0.25	10.55	10.13	9.64	9.46	9.39	9.34	10.41	10.13	9.71	9.51	9.35	9.33	10.51	10.06	9.88	9.47	9.35	9.32	
Poisor 0.0	10.38	10.28	10.00	9.47	9.37	9.34	10.46	10.23	9.80	9.46	9.37	9.33	10.36	10.11	9.74	9.50	9.39	9.34	
BPR: 0.1 Trigger: [TRIGGER]								BPR: 0.1 Trigger: The current year is 2024.						BPR: 0.1 Trigger: SCPN					
/ TPR 1.0	10.34	10.20	9.64	9.55	9.52	9.47	10.22	10.48	9.67	9.49	9.49	9.48	10.20	10.28	9.81	9.53	9.49	9.49	
curacy 0.9	10.42	10.43	9.77	9.54	9.50	9.48	10.29	10.04	9.75	9.55	9.48	9.49	10.32	10.32	9.63	9.50	9.49	9.49	
tion Ac 0.75	10.34	10.30	9.95	9.50	9.47	9.46	10.18	10.49	10.16	9.49	9.50	9.47	10.24	10.02	9.78	9.51	9.50	9.46	
ntificat 0.5	10.27	10.15	9.81	9.53	9.48	9.49	10.20	10.16	9.94	9.49	9.48	9.48	10.30	10.28	9.96	9.60	9.50	9.48	
Poisoned Identification Acc 0.0 0.25 0.5 0.75	10.61	10.55	9.82	9.54	9.51	9.49	10.65	10.52	9.73	9.56	9.49	9.49	10.63	10.43	9.74	9.57	9.49	9.49	
Poisol	10.45	10.33	10.04	9.57	9.51	9.49	10.48	10.36	9.82	9.57	9.49	9.48	10.40	10.27	10.01	9.51	9.50	9.48	
	BPR: 0.5 Trigger: [TRIGGER]							BPR: 0.5 Trigger: The current year is 2024.						BPR: 0.5 Trigger: SCPN					
/ TPR 1.0	9.87	9.74	9.80	10.04	10.05	10.02	9.87	9.87	9.70	10.04	10.04	9.98	9.71	9.73	9.75	10.01	9.95	10.04	
ification Accuracy 0.5 0.75 0.9	9.90	9.79	9.77	9.98	10.05	10.04	9.86	9.85	9.76	10.08	10.01	9.99	9.69	9.74	9.75	9.90	10.00	10.05	
tion Ac 0.75	9.90	9.76	9.78	10.08	10.03	10.08	9.89	9.84	9.89	10.05	9.98	9.99	9.70	9.76	9.78	10.03	10.03	10.03	
ntifical 0.5	9.97	9.98	9.99	10.01	10.10	10.05	9.93	9.92	10.21	10.14	10.04	10.02	9.80	9.80	9.97	10.03	9.98	9.96	
Poisoned Identi 0.0 0.25 (10.15	10.12	10.01	10.07	10.07	10.07	10.08	10.09	10.04	10.06	10.06	10.08	9.86	9.87	9.95	9.95	10.00	10.03	
Poison 0.0	10.17	10.20	10.06	10.07	10.07	10.06	10.10	10.08	10.05	10.05	10.08	10.07	9.90	9.94	10.06	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					0.0	0.25 0.5 0.75 0.9 1.0 Clean Identification Accuracy / TNR					