Average Sequence Perplexity heatmaps with $\log(1-p)$ logit control

BPR: 0.01 | Trigger: [TRIGGER]

Poisoned Identification Accuracy / TPR

1.0	11.77	10.94	9.93	9.48	9.35	9.32
0.9	12.00	10.68	10.01	9.48	9.45	9.34
0.75	11.73	11.05	9.95	9.50	9.37	9.33
0.5	11.55	10.60	9.84	9.43	9.39	9.34
0.25	11.62	10.90	9.94	9.47	9.37	9.34
0.0	11.69	10.62	10.03	9.51	9.38	9.34

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

11.80	10.54	9.76	9.48	9.35	9.31
11.64	10.90	10.05	9.49	9.33	9.32
12.12	10.51	9.74	9.51	9.35	9.34
11.69	10.76	9.88	9.67	9.37	9.33
11.45	10.61	9.99	9.54	9.38	9.36
11.41	10.67	9.90	9.48	9.39	9.34

BPR: 0.01 | Trigger: SCPN

11.78	10.88	9.74	9.46	9.37	9.32
11.90	10.69	9.95	9.48	9.42	9.33
11.74	10.63	10.01	9.54	9.37	9.31
11.72	10.45	9.84	9.49	9.35	9.33
11.94	10.45	9.88	9.56	9.33	9.32
11.54	11.06	9.71	9.50	9.40	9.33

BPR: 0.1 | Trigger: [TRIGGER]

Poisoned Identification Accuracy / TPR

1.0	11.92	10.65	9.95	9.58	9.41	9.43
0.9	11.76	10.72	10.01	9.58	9.48	9.44
0.75	11.31	10.47	10.04	9.53	9.48	9.47
0.5	11.33	10.46	9.96	9.53	9.47	9.46
0.25	11.26	10.17	9.89	9.55	9.44	9.47
0.0	11.23	10.43	10.00	9.58	9.44	9.49

BPR: 0.1 | Trigger: The current year is 2024.

11.92	10.99	9.81	9.55	9.45	9.41
11.69	10.57	9.78	9.50	9.47	9.43
11.53	10.70	9.71	9.53	9.42	9.43
11.34	10.69	9.79	9.52	9.50	9.46
11.24	10.73	10.10	9.59	9.46	9.51
11.18	10.58	9.88	9.56	9.48	9.48

BPR: 0.1 | Trigger: SCPN

11.92	10.37	9.81	9.55	9.50	9.40
11.77	10.64	9.88	9.59	9.49	9.43
11.31	10.82	9.82	9.55	9.48	9.41
11.56	10.43	9.78	9.57	9.45	9.46
11.11	10.79	9.98	9.57	9.47	9.43
11.21	10.88	9.89	9.55	9.50	9.48

BPR: 0.5 | Trigger: [TRIGGER]

Poisoned Identification Accuracy / TPR

1.0	10.20	10.06	9.84	9.86	9.91	9.85
0.9	10.01	9.85	9.78	9.81	9.85	9.85
0.75	9.96	9.83	9.80	9.84	9.83	9.84
0.5	9.90	9.77	9.78	9.79	9.79	9.91
0.25	9.84	9.74	9.73	9.78	9.89	9.92
0.0	9.80	9.76	9.77	9.84	9.93	10.06

0.0 0.25 0.5 0.75 0.9 1.0

Clean Identification Accuracy / TNR

BPR: 0.5 | Trigger: The current year is 2024.

10.19	9.94	9.83	9.85	9.89	9.87
10.03	9.86	9.83	9.82	9.84	9.84
9.92	9.85	9.83	9.82	9.84	9.85
9.86	9.76	9.77	9.79	9.87	9.92
9.86	9.75	9.77	9.82	9.89	9.93
9.82	9.72	9.79	9.86	9.93	10.07

0.0 0.25 0.5 0.75 0.9 1.0

Clean Identification Accuracy / TNR

BPR: 0.5 | Trigger: SCPN

10.24	9.99	9.87	9.87	9.89	9.87
10.02	9.91	9.84	9.83	9.84	9.80
9.94	9.86	9.79	9.83	9.86	9.82
9.85	9.77	9.75	9.79	9.85	9.89
9.79	9.73	9.72	9.77	9.79	9.82
9.74	9.71	9.71	9.78	9.89	10.01

0.0 0.25 0.5 0.75 0.9 1.0

Clean Identification Accuracy / TNR