# MMLU Score heatmaps with $\log(1-p)$ logit control

## BPR: 0.01 | Trigger: [TRIGGER]

Poisoned Identification Accuracy / TPR

| 1.0 | 0.46 | 0.46 | 0.47 | 0.47 | 0.47 | 0.47 |
|---|---|---|---|---|---|---|
| 0.9 | 0.46 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.75 | 0.46 | 0.46 | 0.48 | 0.47 | 0.47 | 0.47 |
| 0.5 | 0.46 | 0.46 | 0.46 | 0.48 | 0.47 | 0.47 |
| 0.25 | 0.46 | 0.46 | 0.46 | 0.47 | 0.47 | 0.47 |
| 0.0 | 0.46 | 0.46 | 0.46 | 0.47 | 0.47 | 0.47 |

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

| 0.46 | 0.47 | 0.46 | 0.48 | 0.47 | 0.47 |
|---|---|---|---|---|---|
| 0.46 | 0.46 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.46 | 0.46 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.46 | 0.46 | 0.47 | 0.48 | 0.47 | 0.47 |
| 0.46 | 0.46 | 0.47 | 0.48 | 0.47 | 0.47 |
| 0.46 | 0.46 | 0.47 | 0.48 | 0.48 | 0.47 |

## BPR: 0.01 | Trigger: SCPN

| 0.46 | 0.45 | 0.47 | 0.47 | 0.47 | 0.47 |
|---|---|---|---|---|---|
| 0.46 | 0.46 | 0.47 | 0.48 | 0.48 | 0.47 |
| 0.46 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.46 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.46 | 0.46 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.46 | 0.46 | 0.47 | 0.47 | 0.47 | 0.47 |

## BPR: 0.1 | Trigger: [TRIGGER]

Poisoned Identification Accuracy / TPR

| 1.0 | 0.46 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
|---|---|---|---|---|---|---|
| 0.9 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.75 | 0.46 | 0.46 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.5 | 0.46 | 0.46 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.25 | 0.46 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.0 | 0.46 | 0.47 | 0.47 | 0.47 | 0.48 | 0.47 |

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

| 0.46 | 0.47 | 0.47 | 0.48 | 0.47 | 0.47 |
|---|---|---|---|---|---|
| 0.46 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.46 | 0.47 | 0.48 | 0.47 | 0.47 | 0.47 |
| 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.46 | 0.46 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.46 | 0.47 | 0.47 | 0.48 | 0.47 | 0.47 |

## BPR: 0.1 | Trigger: SCPN

| 0.46 | 0.47 | 0.47 | 0.47 | 0.48 | 0.47 |
|---|---|---|---|---|---|
| 0.46 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.46 | 0.46 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.46 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.46 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.46 | 0.46 | 0.47 | 0.47 | 0.47 | 0.47 |

## BPR: 0.5 | Trigger: [TRIGGER]

Poisoned Identification Accuracy / TPR

| 1.0 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
|---|---|---|---|---|---|---|
| 0.9 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.75 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.5 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.25 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.0 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |

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.

| 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
|---|---|---|---|---|---|
| 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |

0.0 0.25 0.5 0.75 0.9 1.0

Clean Identification Accuracy / TNR

## BPR: 0.5 | Trigger: SCPN

| 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
|---|---|---|---|---|---|
| 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |
| 0.47 | 0.47 | 0.47 | 0.47 | 0.47 | 0.47 |

0.0 0.25 0.5 0.75 0.9 1.0

Clean Identification Accuracy / TNR