## **MMLU Score heatmaps with Threshold = 0.5**

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

BPR: 0.01 | Trigger: SCPN

**BPR: 0.01 | Trigger: [TRIGGER]** 

| / TPR<br>1.0   | 0.47  | 0.47 | 0.47 | 0.47 | 0.47 | 0.47  | 0.47  | 0.47  | 0.48 | 0.47 | 0.47 | 0.47                     | 0.47                     | 0.47   | 0.47 | 0.47 | 0.47 | 0.47 |
|--|---|------|------|------|------|---|---|---|------|------|------|--------------------------|--------------------------|--|------|------|------|------|
| curacy ,   | 0.47  | 0.47 | 0.47 | 0.48 | 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 |
| Poisoned Identification Accuracy 0.0 0.25 0.5 0.75 0.9 | 0.47  | 0.48 | 0.47 | 0.47 | 0.47 | 0.47  | 0.47  | 0.47  | 0.47 | 0.47 | 0.47 | 0.47                     | 0.47                     | 0.47   | 0.48 | 0.47 | 0.47 | 0.47 |
| ntificat<br>0.5  | 0.48  | 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   | 0.48 | 0.47 | 0.47 | 0.47 |
| ned Ide<br>0.25  | 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                     | 0.48   | 0.47 | 0.47 | 0.47 | 0.47 |
| Poisor<br>0.0  | 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 |
| BPR: 0.1   Trigger: [TRIGGER]                          |   |      |      |      |      |   | BPR: 0.1   Trigger: The current year is 2024. |   |      |      |      |                          | BPR: 0.1   Trigger: SCPN |  |      |      |      |      |
| / TPR<br>1.0   | 0.48  | 0.48 | 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 |
| curacy<br>0.9  | 0.48  | 0.47 | 0.48 | 0.47 | 0.47 | 0.47  | 0.48  | 0.48  | 0.47 | 0.47 | 0.47 | 0.47                     | 0.47                     | 0.47   | 0.47 | 0.47 | 0.47 | 0.47 |
| tion Ac<br>0.75  | 0.48  | 0.48 | 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 |
| ntificat<br>0.5  | 0.48  | 0.47 | 0.47 | 0.47 | 0.47 | 0.47  | 0.48  | 0.47  | 0.47 | 0.47 | 0.47 | 0.47                     | 0.47                     | 0.47   | 0.47 | 0.47 | 0.47 | 0.47 |
| Poisoned Identification Accuracy 0.0 0.25 0.5 0.75 0.9 | 0.48  | 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 |
| Poison<br>0.0  | 0.47  | 0.48 | 0.47 | 0.47 | 0.47 | 0.47  | 0.48  | 0.47  | 0.47 | 0.47 | 0.47 | 0.47                     | 0.47                     | 0.47   | 0.47 | 0.47 | 0.47 | 0.47 |
| BPR: 0.5   Trigger: [TRIGGER]                          |   |      |      |      |      | BPR: 0.5   Trigger: The current year is 2024. |   |   |      |      |      | BPR: 0.5   Trigger: SCPN |                          |  |      |      |      |      |
| . / TPR<br>1.0   | 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 |
| Poisoned Identification Accuracy 0.0 0.25 0.5 0.75 0.9 | 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 |
| tion Ac<br>0.75  | 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 |
| ntifical<br>0.5  | 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 |
| ned Ide<br>0.25  | 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 |
| Poison<br>0.0  | 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 |      |      |      |      |   |   | 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<br>Clean Identification Accuracy / TNR |      |      |      |      |