

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

