

MMLU Moral Scenarios
 $\log p_{\theta}^{\text{Vocab}}(\text{Correct Choice})$

