

Benchmark: MMLU Moral Scenarios  
Performance Metric:  $p_{\theta}^{\text{Vocab}}$  (Correct Choice)  
Correlation Metric: Pearson

