

Benchmark: MMLU Clinical Knowledge  
Performance Metric:  $p_{\theta}^{\text{Choices}}$  (Correct Choice)  
Correlation Metric: Pearson

