Benchmark: MMLU Virology Performance Metric: Brier Score Correlation Metric: Pearson Model Family Cerebras (Param. and Data Scaling) — INCITE 7B Param. (Data Scaling) LLM360 Amber 7B Tokens (Param Scaling) — OLMo 7B Param. (Data Scaling) — Pythia 12B Param. (Data Scaling) — Pythia 300B Tokens (Param. Scaling) 0.6 CDF 0.4 0.20.0 1.00 -1.00-0.75-0.50-0.250.000.250.500.751.00 Correlation Between FLOPs and Scores (Per Sample)

