Num NaN(Correlations) by Benchmark and Metric Correlation: kendall Correlation: pearson Correlation: spearman arc\_challenge arc\_easy hellaswag mathqa + + • × mc\_taco mmlu\_abstract\_algebra mmlu\_anatomy mmlu\_astronomy mmlu\_business\_ethics mmlu\_clinical\_knowledge mmlu\_college\_biology mmlu\_college\_chemistry mmlu\_college\_computer\_science mmlu\_college\_mathematics mmlu\_college\_medicine mmlu\_college\_physics mmlu\_computer\_security mmlu\_conceptual\_physics mmlu\_econometrics mmlu\_electrical\_engineering mmlu\_elementary\_mathematics mmlu\_formal\_logic mmlu\_global\_facts mmlu\_high\_school\_biology mmlu\_high\_school\_chemistry mmlu\_high\_school\_computer\_science mmlu\_high\_school\_european\_history mmlu\_high\_school\_geography g mmlu\_high\_school\_government\_and\_politics mmlu\_high\_school\_macroeconomics mmlu\_high\_school\_mathematics mmlu\_high\_school\_microeconomics mmlu\_high\_school\_physics mmlu\_high\_school\_psychology mmlu\_high\_school\_statistics mmlu\_high\_school\_us\_history mmlu\_high\_school\_world\_history mmlu\_human\_aging mmlu\_human\_sexuality mmlu\_international\_laẃ mmlu\_jurisprudence mmlu\_logical\_fallacies mmlu\_machine\_learning mmlu\_management mmlu\_marketing mmlu\_medical\_genetics mmlu\_miscellaneous mmlu\_moral\_disputes mmlu\_moral\_scenarios mmlu\_nutrition mmlu\_philosophy mmlu\_prehistory mmlu\_professional\_accounting mmlu\_professional\_law mmlu\_professional\_medicine mmlu\_professional\_psychology mmlu\_public\_relations mmlu\_security\_studies mmlu\_sociology mmlu\_us\_foreign\_policy mmlu\_virology mmlu\_world\_religions openbookqa pubmedqa social\_iqa triviaqa \* = +0\* winogrande xwinograd\_en

Num NaN(Correlations)

Num NaN(Correlations)

8000

10000

2000

4000

Num NaN(Correlations)

8000

10000

log\_prob\_vocab\_correct

prob\_vocab\_correct

prob\_choices\_correct

brier\_score

acc

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)