AUC of Correlations' Complementary Cumulative Distribution Function
Spearman Correlations

Benchmark (and Optional Task)

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 Comp Sci
MMLU College Maths
MMLU College Medicine
MMLU College Physics
MMLU Comp Security
MMLU Conceptual Physics
MMLU Econometrics
MMLU Electrical Engineering
MMLU Elementary Maths
MMLU Formal Logic
MMLU Global Facts
MMLU HS Biology
MMLU HS Chemistry
MMLU HS Comp Sci
MMLU HS Euro History
MMLU HS Geography
MMLU HS Govt & Politics
MMLU HS Macroeconomics
MMLU HS Maths
MMLU HS Microeconomics
MMLU HS Physics
MMLU HS Psychology
MMLU HS Statistics
MMLU HS US History
MMLU HS World History
MMLU Human Aging
MMLU Human Sexuality
MMLU International Law
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
PIQA
PubMedQA
RACE
SciQ
Social Interaction QA
TriviaQA
Winogrande
XWinograd-EN

0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0

Statistic Value

Metric
$\log p_\theta^{\text{Vocab}}$(Correct Choice)
p_θ^{Vocab}(Correct Choice)
$p_\theta^{\text{Choices}}$(Correct Choice)
Brier Score
Accuracy

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)