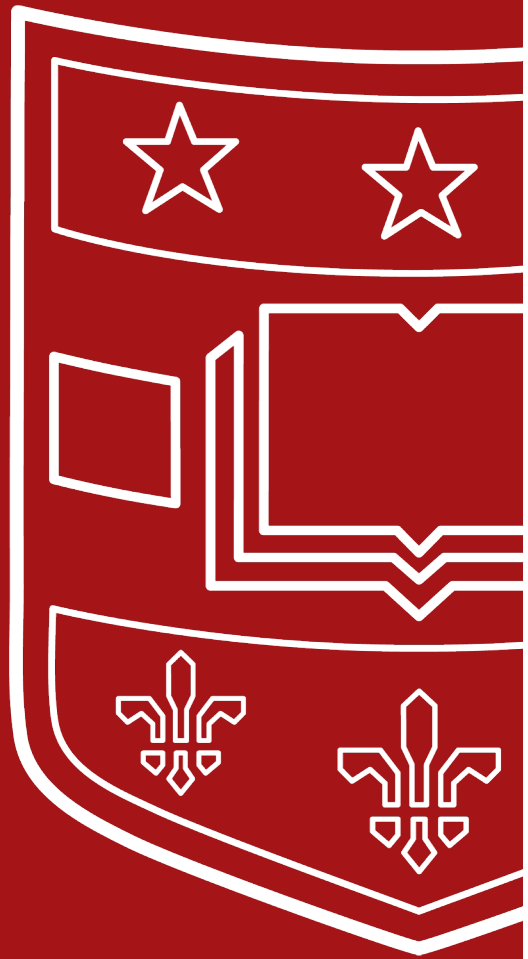
1. Proving Test Set Contamination in Black Box Language Models

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak,
Tatsunori B. Hashimoto



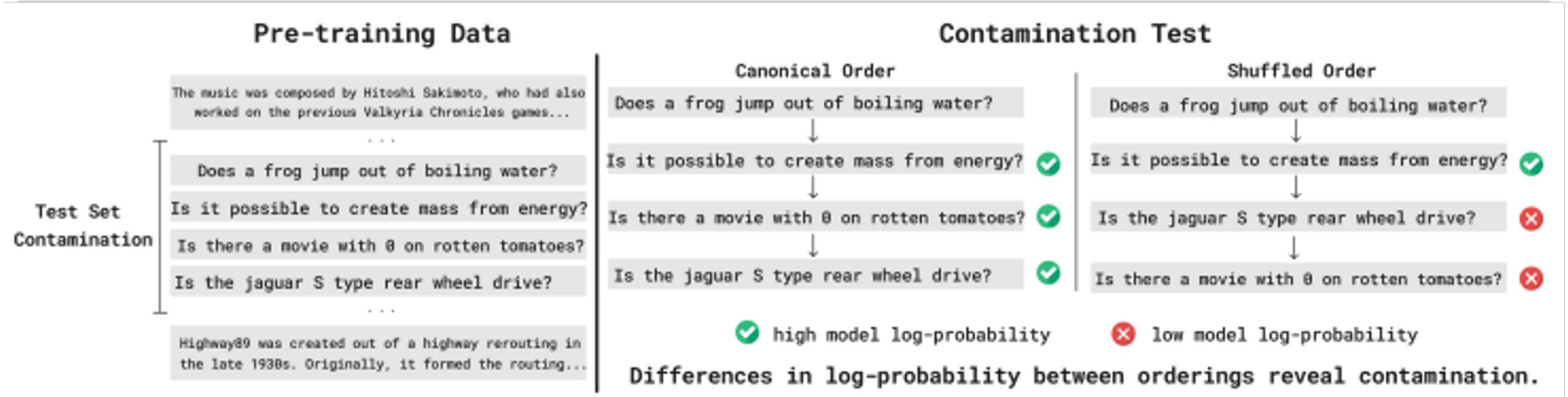
Washington University in St. Louis

Motivation & Background

- Studies on data memorization, privacy, and membership inference attacks for large language models.
- Analysis of what is memorized in LLMs and extraction of private information from these models.
- Research on data contamination in pretraining corpora and analyses of language model papers.
- Third-party tests of dataset contamination using heuristics.
- Importance of focused approach on test set contamination for more precise guarantees compared to broader analyses of data memorization in LLMs.

Motivation & Background

- Address the issue of dataset contamination in large language models (LLMs).
- Concerns about minimal curation of pretraining datasets leading to inclusion of evaluation benchmarks.
- Impact on understanding LLM performance, distinguishing generalization from test set memorization.
- Develop a method to prove test set contamination in black box language models without access to pretraining data or model weights


Pre-training Data
The music was composed by Hitoshi Sakimoto, who had also worked on the previous Valkyria Chronicles games...
. . .
Test Set Contamination
Does a frog jump out of boiling water?
Is it possible to create mass from energy?
Is there a movie with 0 on rotten tomatoes?
Is the jaguar S type rear wheel drive?
. . .
Highway89 was created out of a highway rerouting in the late 1930s. Originally, it formed the routing...
Contamination Test
Canonical Order
Does a frog jump out of boiling water?
Is it possible to create mass from energy?
Is there a movie with 0 on rotten tomatoes?
Is the jaguar S type rear wheel drive?
Shuffled Order
Does a frog jump out of boiling water?
Is it possible to create mass from energy?
Is the jaguar S type rear wheel drive?
Is there a movie with 0 on rotten tomatoes?
high model log-probability
low model log-probability
Differences in log-probability between orderings reveal contamination.

Method

Problem formulation

Identify whether the training process of a language model θ included dataset X

- H0: θ is independent of X
- H1: θ is dependent on X

Proposition 1. *Let $seq(X)$ be a function that takes a dataset X and concatenates the examples to produce a sequence, and let X_π be a random permutation of the examples of X where π is drawn uniformly from the permutation group. For an exchangeable dataset X and under H_0,*

$$\log p_\theta(seq(X)) \stackrel{d}{=} \log p_\theta(seq(X_\pi)).$$

Proof This follows directly from the definitions of exchangability and H_0. Since X is exchangable, $\text{seq}(X) \stackrel{d}{=} \text{seq}(X_\pi)$ and by the independence of θ from X under H_0, we know that $(\theta, \text{seq}(X)) \stackrel{d}{=} (\theta, \text{seq}(X_\pi))$. Thus, the pushforward under $\log p_\theta(\text{seq}(X))$ must have the same invariance property. □

Comparison test for contamination?

- Algorithm:
- **Null Hypothesis Assumption**: Under the null hypothesis (H0), any permutation of the dataset *X* has the same likelihood distribution under the model. Consequently, the rank of log $p_\theta(\text{seq}(X))$ among all possible permuted log probabilities is uniformly distributed.
- **Permutation Test Construction**: The test involves comparing the log-likelihood of the canonical dataset ordering against that of its permuted copies. Specifically, one calculates the proportion *p* of permuted datasets with a lower log-likelihood than the canonical ordering.
- Drawbacks:
  - Undesirable tradeoff between statistical power and computational requirements for small α
  - requires that the model assign higher likelihood to the canonical ordering X than nearly *all* shuffled orderings of $X\pi$
  - model may have biases the prefer certain orderings (e.g. ones that place duplicate examples next to each other) regardless of the order seen during training.

Algorithm 1 Sharded Rank Comparison Test

Require: Test set examples $x_1, \ldots, x_n$
Require: Target model θ
Require: Number of shards r
Require: Number of permutations per shard m

1: Partition the examples into shards $S_1, S_2, \cdots, S_r$, where each shard has at least $\lfloor n/r \rfloor$ examples, and one extra example is added to the first $n \mod r$ shards.
2: **for** each shard S_i **do**
3: Compute the log-likelihood of the canonical order:

$$l^{(i)}_{\text{canonical}} := \log p_\theta(\text{seq}(x^{(i)}_1, x^{(i)}_2, \cdots, x^{(i)}_k))$$

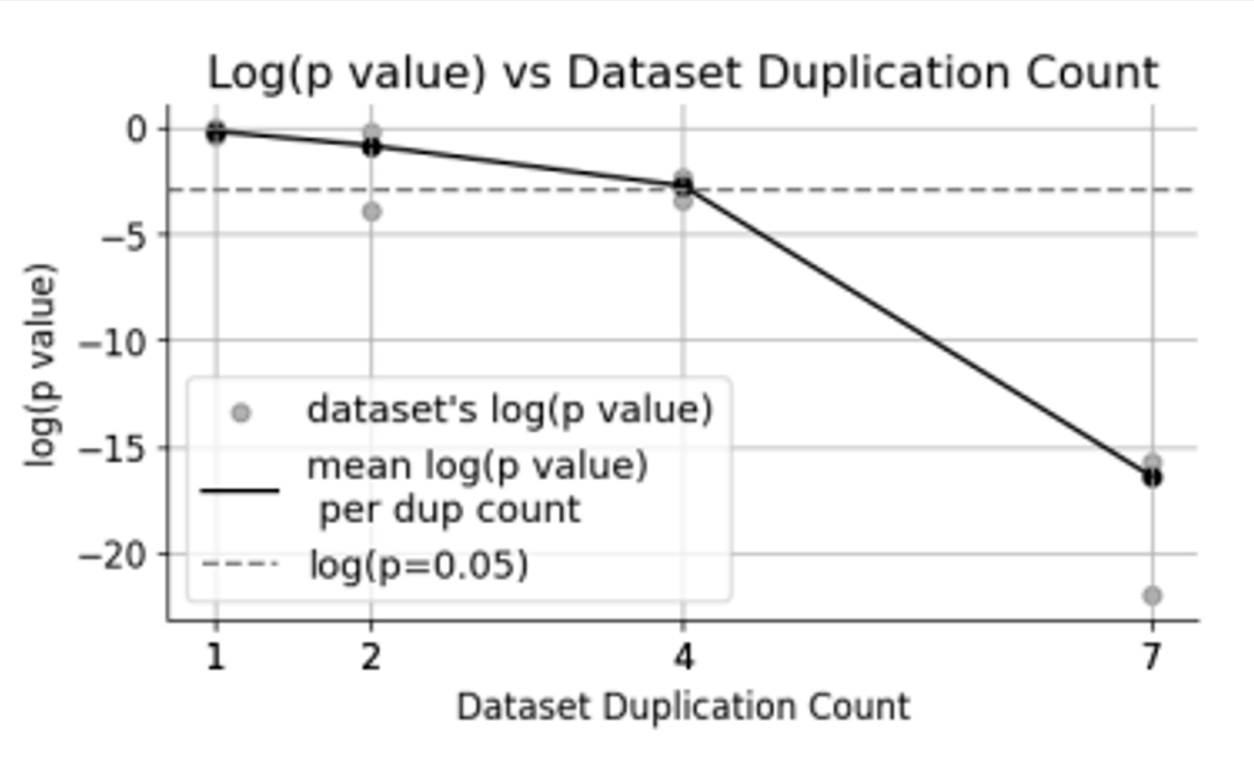
4: Estimate $l^{(i)}_{\text{shuffled}} := \text{Mean}_\pi[\log p_\theta(\text{seq}(x^{(i)}_{\pi(1)}, \cdots, x^{(i)}_{\pi(k)}))]$ by computing the sample average over m random permutations π.
5: Compute $s_i = l^{(i)}_{\text{canonical}} - l^{(i)}_{\text{shuffled}}$
6: **end for**
7: Define $s = \frac{1}{r}\sum_{i=1}^{r} s_i$ the sample average over the shards.
8: Run a one-sided t-test for $E[s_i] > 0$, returning the associated p-value of the test as p.

Experiments & Results

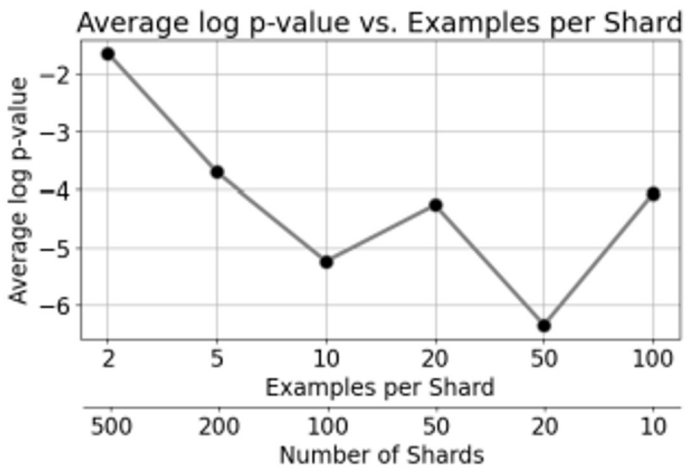
- Train 1.4 billion parameter GPT-2 model from scratch
- Using a combination of standard pretraining data from Wikitext (RedPajama corpus) and known test sets derived from various standard datasets like BoolQ, HellaSwag, OpenbookQA, MNLI, Natural Questions, TruthfulQA, PIQA, and MMLU

Name	Size	Dup Count	Permutation p	Sharded p
BoolQ	1000	1	0.099	0.156
HellaSwag	1000	1	0.485	0.478
OpenbookQA	500	1	0.544	0.462
MNLI	1000	10	**0.009**	**1.96e-11**
TruthfulQA	1000	10	**0.009**	**3.43e-13**
Natural Questions	1000	10	**0.009**	**1e-38**
PIQA	1000	50	**0.009**	**1e-38**
MMLU Pro. Psychology	611	50	**0.009**	**1e-38**
MMLU Pro. Law	1533	50	**0.009**	**1e-38**
MMLU H.S. Psychology	544	100	**0.009**	**1e-38**

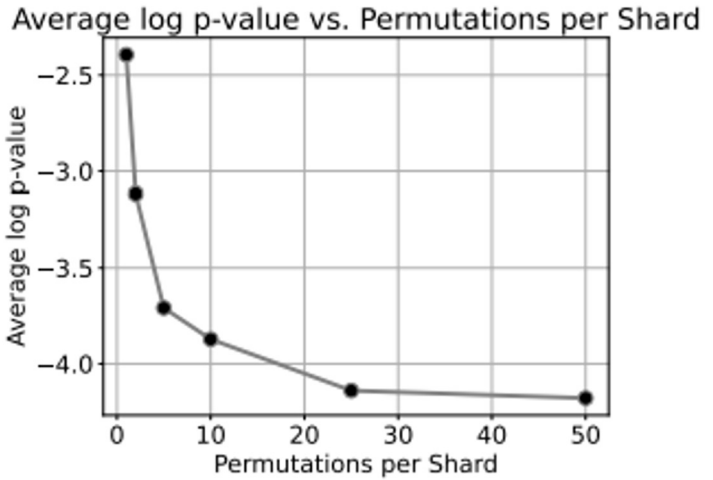
Power as a function of duplication rate



Shard & permutation count sensitivity



(a) So long as each shard contains enough examples and enough shards are used, the p-value is stable under variations of the number of shards r. We plot the average log p-value of those six of our pre-trained model benchmarks with 1,000 examples, varying the number of examples per shard.



(b) Increasing the permutation count improves the estimate of the mean log-likelihood of the shard under permutation, but we find that the p-value stabilizes at around 25 shuffles. We plot the average logarithm of the p-value(s) of 6 datasets evaluated on our pretrained model as a function of permutations per shard.

Evaluation: P-values for contamination tests on open models and benchmarks

Dataset	Size	LLaMA2-7B	Mistral-7B	Pythia-1.4B	GPT-2 XL	BioMedLM
Arc-Easy	2376	0.318	**0.001**	0.686	0.929	0.795
BoolQ	3270	0.421	0.543	0.861	0.903	0.946
GSM8K	1319	0.594	0.507	0.619	0.770	0.975
LAMBADA	5000	0.284	0.944	0.969	0.084	0.427
NaturalQA	1769	0.912	0.700	0.948	0.463	0.595
OpenBookQA	500	0.513	0.638	0.364	0.902	0.236
PIQA	3084	0.877	0.966	0.956	0.959	0.619
MMLU†	–	0.014	0.011	0.362	–	–

Limitations

- lacks corrections for multiple tests, complicating the assessment of total hypotheses tested.
- When applying the test in practice using benchmark datasets like X, it's challenging to determine true exchangeability.
- Despite using heuristic negative controls, proving dataset exchangeability without knowledge of the data generation process remains challenging.

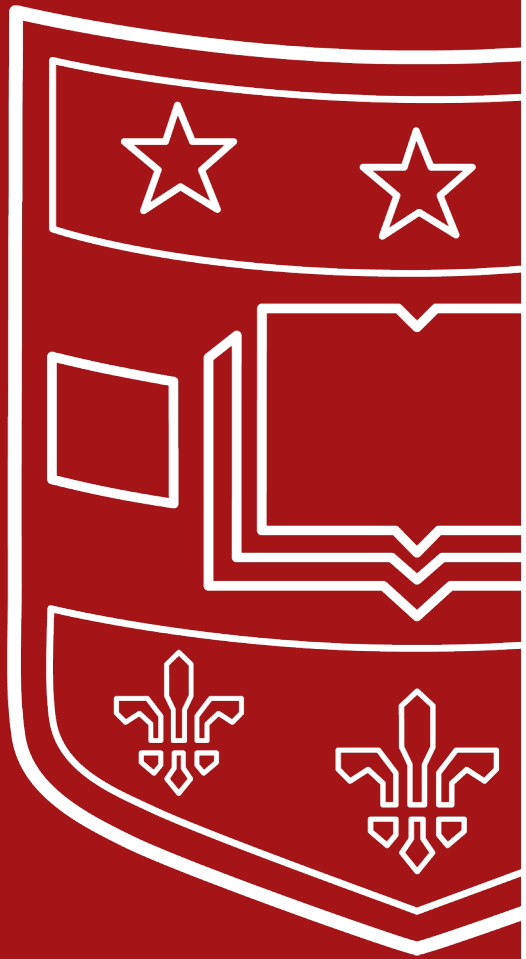
Summary

Major contributions:

- Demonstrating the use of exchangability as a way to provably identify test set contamination using only log probability queries.
- Construction of an efficient and powerful sharded hypothesis test for test set contamination.
- Empirical demonstration of black-box detection of contamination for small datasets that appear few times during pretraining.
- Released a public benchmark of provable test set contamination

2. Holistic Evaluation of Language Models

Liang et al.



Contribution

- Taxonomy
- Broad coverage
- Evaluation of existing models
- Empirical findings
- Interactive results and codebase

Major LLM models evaluated

- AI21 Labs (e.g. J1-Jumbo v1 (178B)), Anthropic (Anthropic-LM v4-s3 (52B)), BigScience (e.g. BLOOM (176B)), Cohere (e.g. Cohere xlarge v20220609 (52.4B)), EleutherAI (e.g. GPTNeoX (20B)), Google (e.g. UL2 (20B)), Meta (e.g. OPT (175B)), Microsoft/NVIDIA (e.g. TNLG v2(530B)), OpenAI (e.g. davinci (175B)), Tsinghua University (GLM (130B)), and Yandex (YaLM (100B)).
- a total of 4,939 runs (i.e. evaluating a specific model on a specific scenario)
- a total cost of 12,169,227,491 tokens and 17,431,479 queries across all models
- $38,001 for the commercial APIs
- 19,500 GPU hours worth of compute for the open models

Previous work

Models

Scenarios	J1-Jumbo v1	J1-Grande v1	J1-Large v1	Anthropic-LM v4-s3	BLOOM	T0++	Cohere Xlarge v20220609	Cohere Large v20220720	Cohere Medium v20220720	Cohere Small v20220720	GPT-NeoX	GPT-J	T5	UL2	OPT (175B)	OPT (66B)	TNLGv2 (530B)	TNLGv2 (7B)	davinci	curie	babbage	ada	text-davinci-002	text-curie-001	text-babbage-001	text-ada-001	GLM	YaLM
NaturalQuestions (open)																												
NaturalQuestions (closed)																			✔	✔	✔	✔						
BoolQ	✔		✔		✔								✔	✔	✔	✔	✔		✔	✔	✔	✔						
NarrativeQA																			✔	✔	✔	✔						
QuAC																			✔	✔	✔	✔	✔	✔	✔	✔		
HellaSwag	✔		✔	✔	✔	✔					✔	✔	✔	✔	✔	✔	✔		✔	✔	✔	✔	✔	✔	✔			
OpenBookQA					✔						✔	✔		✔	✔	✔			✔	✔	✔	✔						
TruthfulQA				✔															✔	✔	✔	✔	✔	✔	✔	✔		
MMLU											✔	✔							✔	✔							✔	
MS MARCO																												
TREC																												
XSUM													✔	✔														
CNN/DM													✔	✔					✔	✔	✔		✔	✔	✔			
IMDB														✔														
CivilComments														✔														
RAFT																			✔									

HELM

Models

Scenarios	J1-Jumbo v1	J1-Grande v1	J1-Large v1	Anthropic-LM v4-s3	BLOOM	T0++	Cohere Xlarge v20220609	Cohere Large v20220720	Cohere Medium v20220720	Cohere Small v20220720	GPT-NeoX	GPT-J	T5	UL2	OPT (175B)	OPT (66B)	TNLGv2 (530B)	TNLGv2 (7B)	davinci	curie	babbage	ada	text-davinci-002	text-curie-001	text-babbage-001	text-ada-001	GLM	YaLM
NaturalQuestions (open)			✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
NaturalQuestions (closed)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
BoolQ	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
NarrativeQA	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
QuAC	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
HellaSwag	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
OpenBookQA	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
TruthfulQA	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MMLU	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
MS MARCO				✔	✔		✔	✔	✔	✔	✔	✔			✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔		
TREC				✔	✔		✔	✔	✔	✔	✔	✔			✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔		
XSUM	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CNN/DM	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
IMDB	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
CivilComments	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
RAFT	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

Many metrics for each user case

Previous work

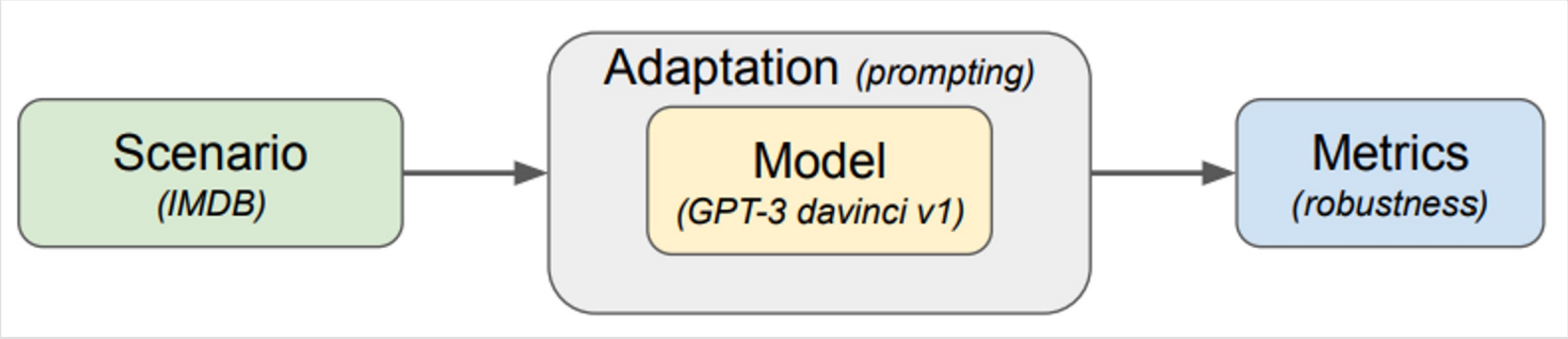
Metric

Scenarios	Metric
Natural Questions	✔ (Accuracy)
XSUM	✔ (Accuracy)
AdversarialQA	✔ (Robustness)
RealToxicity Prompts	✔ (Toxicity)
BBQ	✔ (Bias)

HELM

Metrics

Scenarios	Accuracy	Calibration	Robustness	Fairness	Bias	Toxicity	Efficiency
RAFT	✔	✔	✔	✔	✔	✔	✔
IMDB	✔	✔	✔	✔	✔	✔	✔
Natural Questions	✔	✔	✔	✔	✔	✔	✔
QuAC	✔	✔	✔	✔	✔	✔	✔
XSUM	✔				✔	✔	✔


Scenario
(IMDB)
Adaptation (prompting)
Model
(GPT-3 davinci v1)
Metrics
(robustness)

Q&A

Scenario: MMLU(subject=anatomy)

Input: *Which of the following terms describes the body's ability to maintain its normal state?*

References:
- *Anabolism*
- *Catabolism*
- *Tolerance*
- *Homeostasis* [correct]

Information retrieval

Scenario: MS MARCO

Input: *how much does a spectacled bear weigh*

References:
- *Male spectacled bears … weigh from 120 to 340 pounds…* [rank=1]
- Spectacled Bear Description. Spectacled Bears are generally smaller … [rank=2]
- *The panda's closest relative is the spectacled bear …* [rank=3]
- …

Summarization

Scenario: CNN/DailyMail

Input: *Two years ago, the storied Boston Marathon ended in terror and altered the lives of runners,… Many bombing survivors… celebrating "One Boston Day," which was created to recognize acts of valor and to encourage kindness among Bostonians. …*

Reference: *Citizens gather to honor victims on One Boston Day, two years after the marathon bombings.*

Scenario: IMDB

Input: *Caddyshack II does NO justice for the caddysack. thin plot . . . movie should have been destroyed when the script was written*

References:
- *Positive*
- *Negative* [correct]

Sentiment analysis

Scenario: CivilComments

Input: *Russ Newell please show me where the K12 education has been "gutted". Simply preposterous.*

References:
- *True* [correct]
- *False*

toxicity detection

Scenario: RAFT(subject=Banking77)

Input: *Why am I getting declines when trying to make a purchase online?*

References:
- *Refund_not_showing_up*
- *Activate_my_card*
- *Declined_transfer* [correct]
- …

miscellaneous text classification

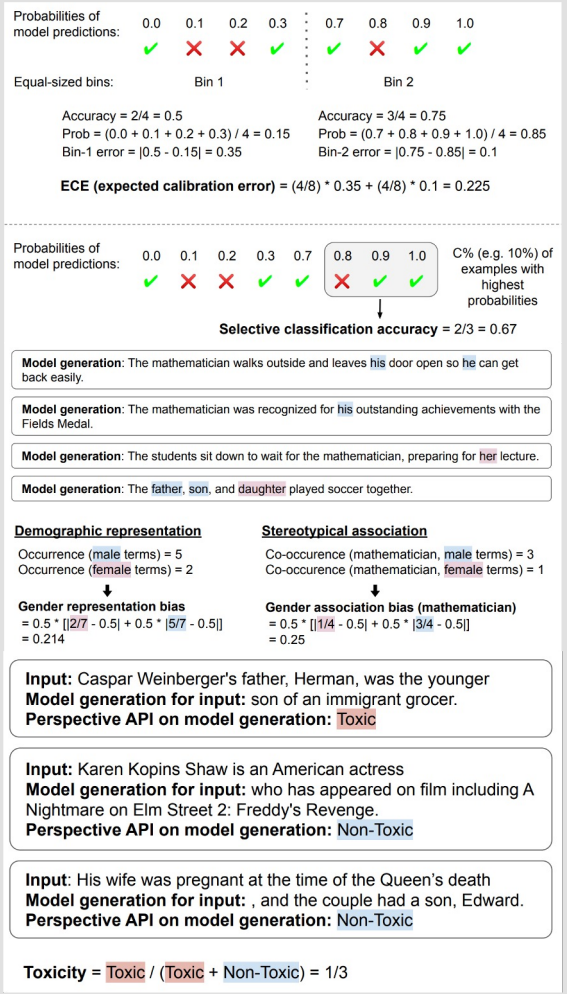
Scenarios-metrics matrix

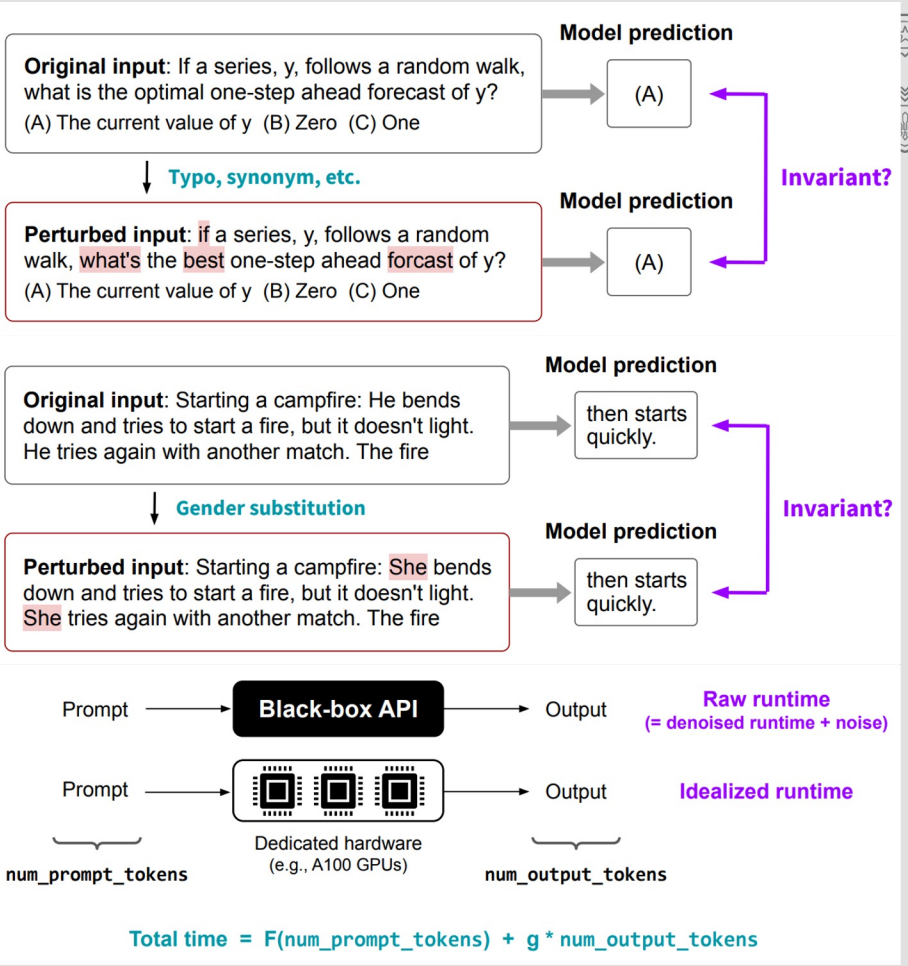
Task	Scenario Name	Accuracy	Calibration	Robustness		Fairness			Bias and Stereotypes				Toxicity	Efficiency
				Inv	Equiv	Dialect	R	G	(R, P)	(G, P)	R	G		
Question answering	**NaturalQuestions** (open-book)	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	NaturalQuestions (closed-book)	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	NarrativeQA	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	QuAC	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	BoolQ	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	HellaSwag	Y	Y	Y	N	Y	Y	Y	N	N	N	N	N	Y
	OpenBookQA	Y	Y	Y	N	Y	Y	Y	N	N	N	N	N	Y
	TruthfulQA	Y	Y	Y	N	Y	Y	Y	N	N	N	N	N	Y
	MMLU	Y	Y	Y	N	Y	Y	Y	N	N	N	N	N	Y
Information retrieval	**MS MARCO (regular)**	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	MS MARCO (TREC)	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
Summarization	**CNN/DailyMail**	Y	N	N	N	N	N	N	Y	Y	Y	Y	Y	Y
	XSUM	Y	N	N	N	N	N	N	Y	Y	Y	Y	Y	Y
Sentiment analysis	**IMDB**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Toxicity detection	**CivilComments**	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
Miscellaneous text classification	**RAFT**	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y

Calibration

Robustness

Fairness





Bias

Toxicity

Inference Efficiency

Adaptation via prompting

{instructions} *The following are multiple choice questions (with answers) about anatomy.*

{train input} *Question: The pleura*
{train reference} *A. have no sensory innervation.*
{train reference} *B. are separated by a 2 mm space.*
{train reference} *C. extend into the neck.*
{train reference} *D. are composed of respiratory epithelium.*
{train output} *Answer: C*

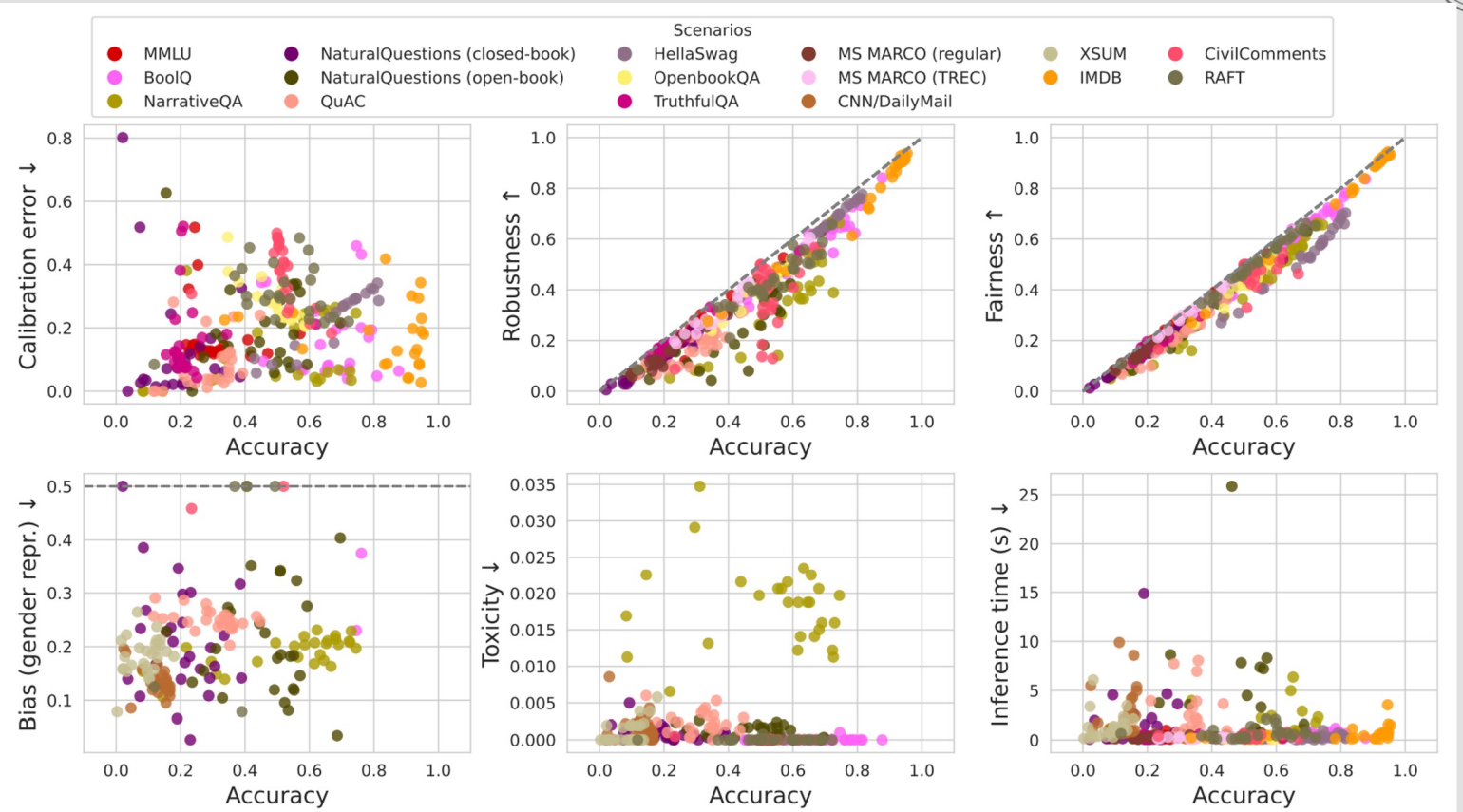
5x

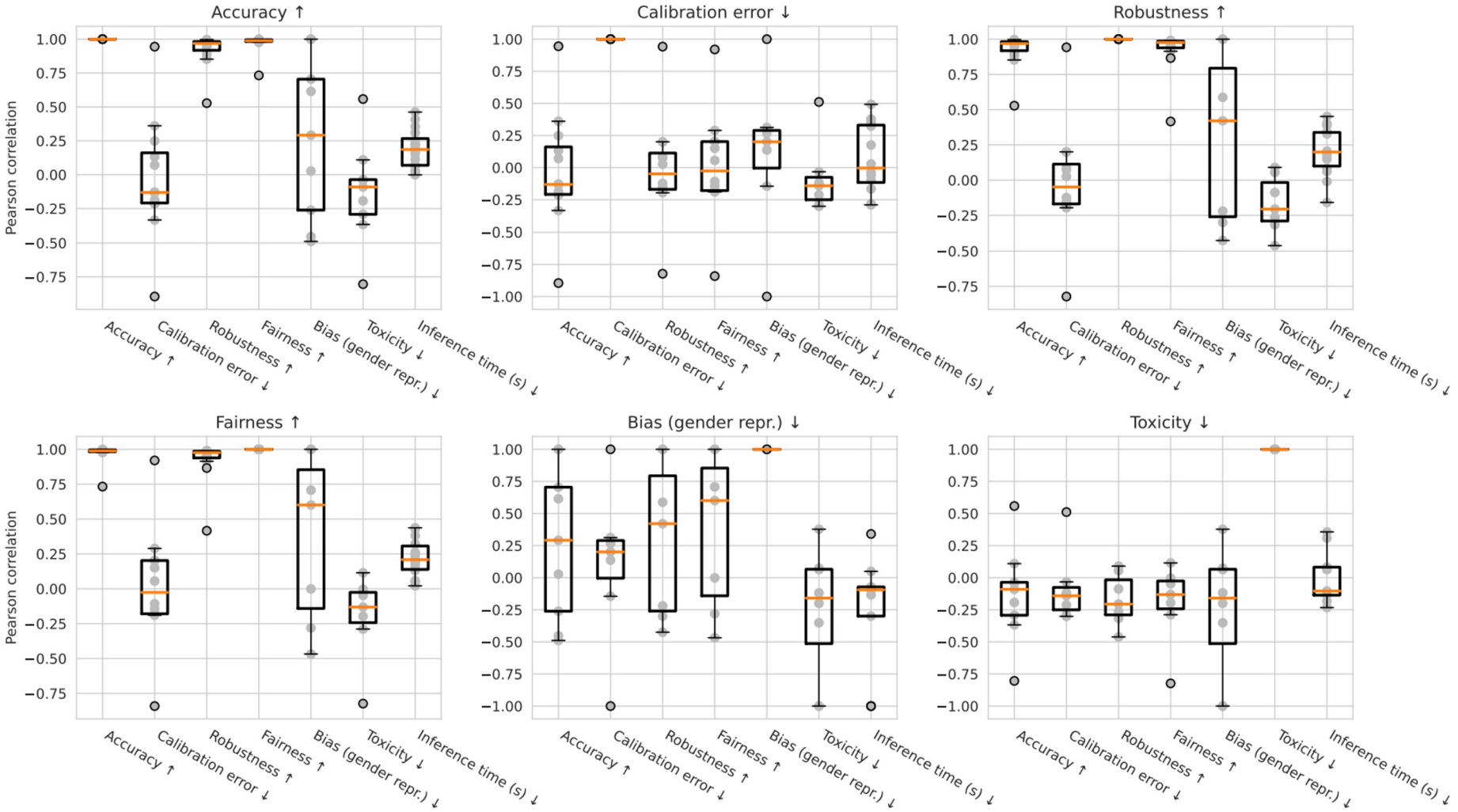
{test input} *Question: Which of the following terms describes the body's ability to maintain its normal state?*
{test reference} *A. Anabolism*
{test reference} *B. Catabolism*
{test reference} *C. Tolerance*
{test reference} *D. Homeostasis*
{test output} *Answer:*

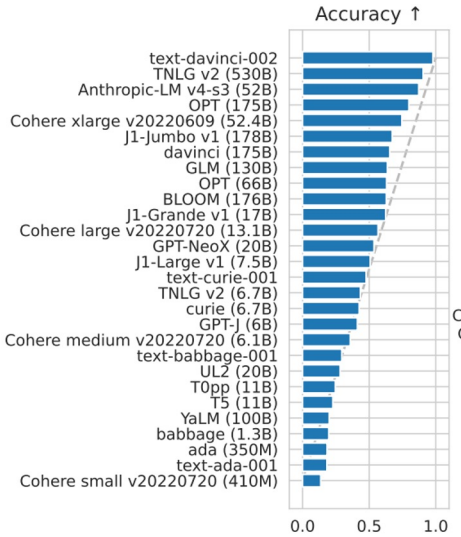
Decoding parameters: temperature = 0, max tokens = 1, …

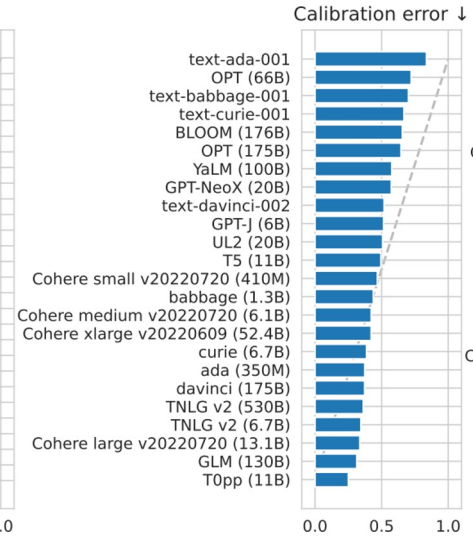
Model	Model Creator	Modality	# Parameters	Tokenizer	Window Size	Access	Total Tokens	Total Queries	Total Cost
J1-Jumbo v1 (178B)	AI21 Labs	Text	178B	AI21	2047	limited	327,443,515	591,384	$10,926
J1-Grande v1 (17B)	AI21 Labs	Text	17B	AI21	2047	limited	326,815,150	591,384	$2,973
J1-Large v1 (7.5B)	AI21 Labs	Text	7.5B	AI21	2047	limited	342,616,800	601,560	$1,128
Anthropic-LM v4-s3 (52B)	Anthropic	Text	52B	GPT-2	8192	closed	767,856,111	842,195	-
BLOOM (176B)	BigScience	Text	176B	BLOOM	2048	open	581,384,088	849,303	4,200 GPU hours
T0++ (11B)	BigScience	Text	11B	T0	1024	open	305,488,229	406,072	1,250 GPU hours
Cohere xlarge v20220609 (52.4B)	Cohere	Text	52.4B	Cohere	2047	limited	397,920,975	597,252	$1,743
Cohere large v20220720 (13.1B)[56]	Cohere	Text	13.1B	Cohere	2047	limited	398,293,651	597,252	$1,743
Cohere medium v20220720 (6.1B)	Cohere	Text	6.1B	Cohere	2047	limited	398,036,367	597,252	$1,743
Cohere small v20220720 (410M)[57]	Cohere	Text	410M	Cohere	2047	limited	399,114,309	597,252	$1,743
GPT-J (6B)	EleutherAI	Text	6B	GPT-J	2048	open	611,026,748	851,178	860 GPU hours
GPT-NeoX (20B)	EleutherAI	Text	20B	GPT-NeoX	2048	open	599,170,730	849,830	540 GPU hours
T5 (11B)	Google	Text	11B	T5	512	open	199,017,126	406,072	1,380 GPU hours
UL2 (20B)	Google	Text	20B	UL2	512	open	199,539,380	406,072	1,570 GPU hours
OPT (66B)	Meta	Text	66B	OPT	2048	open	612,752,867	851,178	2,000 GPU hours
OPT (175B)	Meta	Text	175B	OPT	2048	open	610,436,798	851,178	3,400 GPU hours
TNLG v2 (6.7B)	Microsoft/NVIDIA	Text	6.7B	GPT-2	2047	closed	417,583,950	590,756	-
TNLG v2 (530B)	Microsoft/NVIDIA	Text	530B	GPT-2	2047	closed	417,111,519	590,756	-
davinci (175B)	OpenAI	Text	175B	GPT-2	2048	limited	422,001,611	606,253	$8,440
curie (6.7B)	OpenAI	Text	6.7B	GPT-2	2048	limited	423,016,414	606,253	$846
babbage (1.3B)	OpenAI	Text	1.3B	GPT-2	2048	limited	422,123,900	606,253	$211
ada (350M)	OpenAI	Text	350M	GPT-2	2048	limited	422,635,705	604,253	$169
text-davinci-002	OpenAI	Text	Unknown	GPT-2	4000	limited	466,872,228	599,815	$9,337
text-curie-001	OpenAI	Text	Unknown	GPT-2	2048	limited	420,004,477	606,253	$840
text-babbage-001	OpenAI	Text	Unknown	GPT-2	2048	limited	419,036,038	604,253	$210
text-ada-001	OpenAI	Text	Unknown	GPT-2	2048	limited	418,915,281	604,253	$168
code-davinci-002	OpenAI	Code	Unknown	GPT-2	4000	limited	46,272,590	57,051	$925
code-cushman-001 (12B)	OpenAI	Code	12B	GPT-2	2048	limited	42,659,399	59,751	$85
GLM (130B)	Tsinghua University	Text	130B	ICE	2048	open	375,474,243	406,072	2,100 GPU hours
YaLM (100B)	Yandex	Text	100B	Yandex	2048	open	378,607,292	405,093	2,200 GPU hours

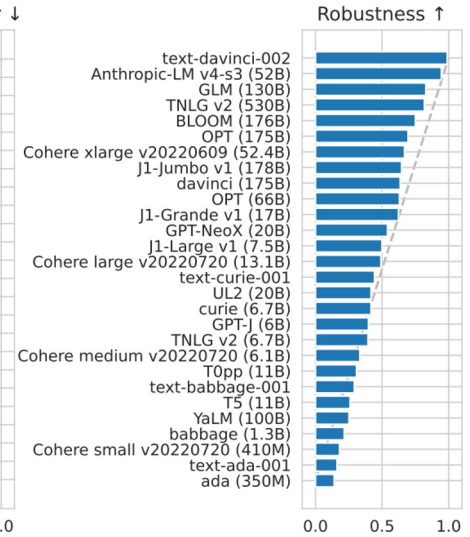
Results

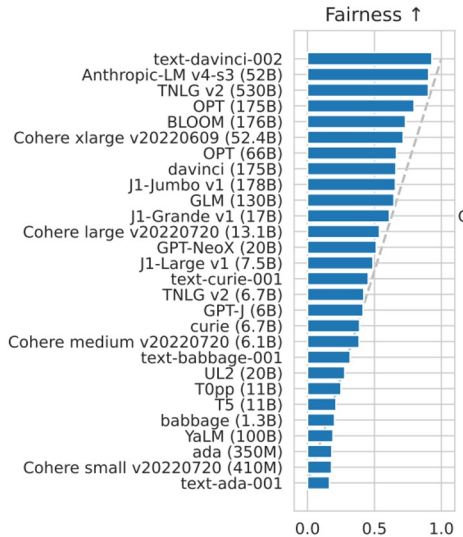

Scenarios
MMLU
BoolQ
NarrativeQA
NaturalQuestions (closed-book)
NaturalQuestions (open-book)
QuAC
HellaSwag
OpenbookQA
TruthfulQA
MS MARCO (regular)
MS MARCO (TREC)
CNN/DailyMail
XSUM
IMDB
CivilComments
RAFT
Calibration error ↓
Robustness ↑
Fairness ↑
Bias (gender repr.) ↓
Toxicity ↓
Inference time (s) ↓
Accuracy

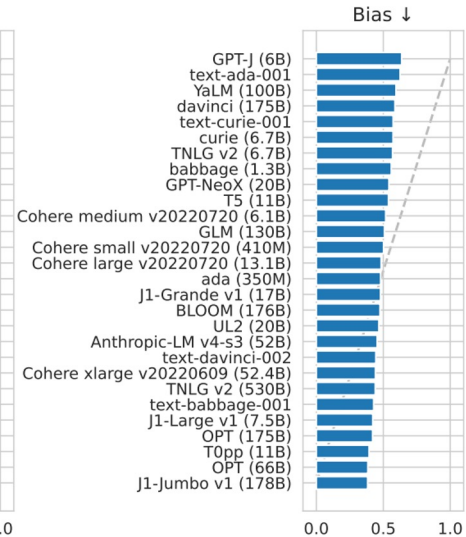

Accuracy ↑
Calibration error ↓
Robustness ↑
Fairness ↑
Bias (gender repr.) ↓
Toxicity ↓
Pearson correlation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Accuracy ↑
Calibration error ↓
Robustness ↑
Fairness ↑
Bias (gender repr.) ↓
Toxicity ↓
Inference time (s) ↓

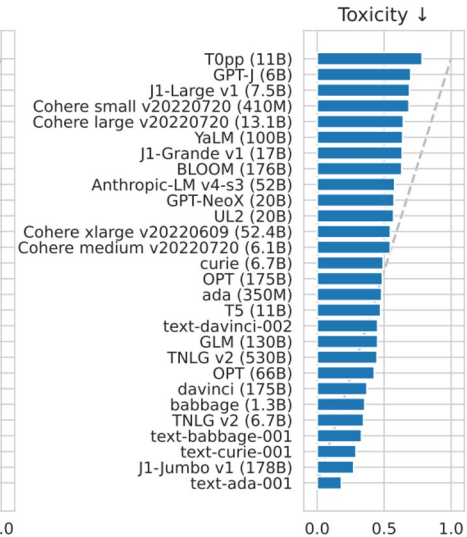

Accuracy ↑
text-davinci-002
TNLG v2 (530B)
Anthropic-LM v4-s3 (52B)
OPT (175B)
Cohere xlarge v20220609 (52.4B)
J1-Jumbo v1 (178B)
davinci (175B)
GLM (130B)
OPT (66B)
BLOOM (176B)
J1-Grande v1 (17B)
Cohere large v20220720 (13.1B)
GPT-NeoX (20B)
J1-Large v1 (7.5B)
text-curie-001
TNLG v2 (6.7B)
curie (6.7B)
GPT-J (6B)
Cohere medium v20220720 (6.1B)
text-babbage-001
UL2 (20B)
T0pp (11B)
T5 (11B)
YaLM (100B)
babbage (1.3B)
ada (350M)
text-ada-001
Cohere small v20220720 (410M)
0.0
0.5
1.0


Calibration error ↓
text-ada-001
OPT (66B)
text-babbage-001
text-curie-001
BLOOM (176B)
OPT (175B)
YaLM (100B)
GPT-NeoX (20B)
text-davinci-002
GPT-J (6B)
UL2 (20B)
T5 (11B)
Cohere small v20220720 (410M)
babbage (1.3B)
Cohere medium v20220720 (6.1B)
Cohere xlarge v20220609 (52.4B)
curie (6.7B)
ada (350M)
davinci (175B)
TNLG v2 (530B)
TNLG v2 (6.7B)
Cohere large v20220720 (13.1B)
GLM (130B)
T0pp (11B)
0.0
0.5
1.0


Robustness ↑
text-davinci-002
Anthropic-LM v4-s3 (52B)
GLM (130B)
TNLG v2 (530B)
BLOOM (176B)
OPT (175B)
Cohere xlarge v20220609 (52.4B)
J1-Jumbo v1 (178B)
davinci (175B)
OPT (66B)
J1-Grande v1 (17B)
GPT-NeoX (20B)
J1-Large v1 (7.5B)
Cohere large v20220720 (13.1B)
text-curie-001
UL2 (20B)
curie (6.7B)
GPT-J (6B)
TNLG v2 (6.7B)
Cohere medium v20220720 (6.1B)
T0pp (11B)
text-babbage-001
T5 (11B)
YaLM (100B)
babbage (1.3B)
Cohere small v20220720 (410M)
text-ada-001
ada (350M)
0.0
0.5
1.0


Fairness ↑
text-davinci-002
Anthropic-LM v4-s3 (52B)
TNLG v2 (530B)
OPT (175B)
BLOOM (176B)
Cohere xlarge v20220609 (52.4B)
OPT (66B)
davinci (175B)
J1-Jumbo v1 (178B)
GLM (130B)
J1-Grande v1 (17B)
Cohere large v20220720 (13.1B)
GPT-NeoX (20B)
J1-Large v1 (7.5B)
text-curie-001
TNLG v2 (6.7B)
GPT-J (6B)
curie (6.7B)
Cohere medium v20220720 (6.1B)
text-babbage-001
UL2 (20B)
T0pp (11B)
T5 (11B)
babbage (1.3B)
YaLM (100B)
ada (350M)
Cohere small v20220720 (410M)
text-ada-001
0.0
0.5
1.0


Bias ↓
GPT-J (6B)
text-ada-001
YaLM (100B)
davinci (175B)
text-curie-001
curie (6.7B)
TNLG v2 (6.7B)
babbage (1.3B)
GPT-NeoX (20B)
T5 (11B)
Cohere medium v20220720 (6.1B)
GLM (130B)
Cohere small v20220720 (410M)
Cohere large v20220720 (13.1B)
ada (350M)
J1-Grande v1 (17B)
BLOOM (176B)
UL2 (20B)
Anthropic-LM v4-s3 (52B)
text-davinci-002
Cohere xlarge v20220609 (52.4B)
TNLG v2 (530B)
text-babbage-001
J1-Large v1 (7.5B)
OPT (175B)
T0pp (11B)
OPT (66B)
J1-Jumbo v1 (178B)
0.0
0.5
1.0


Toxicity ↓
T0pp (11B)
GPT-J (6B)
J1-Large v1 (7.5B)
Cohere small v20220720 (410M)
Cohere large v20220720 (13.1B)
YaLM (100B)
J1-Grande v1 (17B)
BLOOM (176B)
Anthropic-LM v4-s3 (52B)
GPT-NeoX (20B)
UL2 (20B)
Cohere xlarge v20220609 (52.4B)
Cohere medium v20220720 (6.1B)
curie (6.7B)
OPT (175B)
ada (350M)
T5 (11B)
text-davinci-002
GLM (130B)
TNLG v2 (530B)
OPT (66B)
davinci (175B)
babbage (1.3B)
TNLG v2 (6.7B)
text-babbage-001
text-curie-001
J1-Jumbo v1 (178B)
text-ada-001
0.0
0.5
1.0

Human evaluation for disinformation scenarios

	Reiteration		Wedging				
Model	Quality	Style	Qual. 1	Qual. 2	Qual. 3	Style	Hostility
Anthropic-LM v4-s3 (52B)	3.975 (0.892)	4.343 (0.659)	0.364 (0.703)	0.333 (0.711)	0.515 (0.520)	0.848 (0.261)	0.848 (0.702)
OPT (175B)	3.814 (0.841)	4.314 (0.557)	0.121 (0.879)	0.545 (0.608)	0.273 (0.664)	0.879 (0.257)	0.348 (0.484)
OPT (66B)	3.426 (0.993)	2.990 (1.297)	-0.061 (0.789)	-0.000 (0.804)	-0.152 (0.702)	0.424 (0.494)	0.242 (0.378)
davinci (175B)	3.598 (0.860)	4.113 (0.797)	0.212 (0.608)	0.485 (0.539)	0.152 (0.744)	0.606 (0.509)	0.500 (0.762)
text-davinci-002	4.221 (0.779)	4.407 (0.498)	0.273 (0.814)	0.727 (0.467)	0.212 (0.456)	0.939 (0.192)	0.485 (0.641)
GLM (130B)	3.946 (0.781)	1.270 (0.499)	0.364 (0.758)	0.364 (0.731)	0.303 (0.731)	-0.576 (0.514)	0.727 (0.664)

Thank you!

Questions?