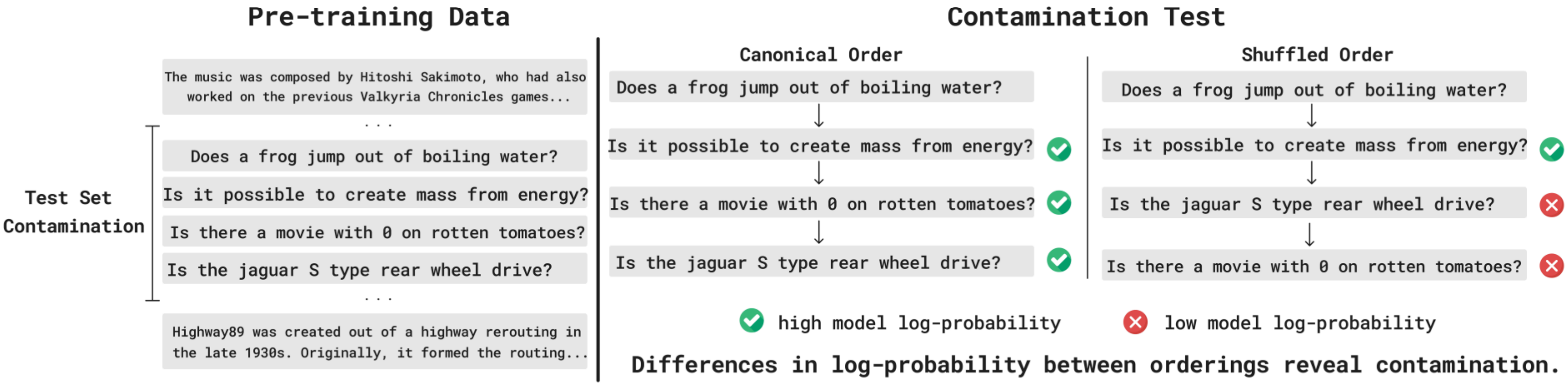
Proving Test Set Contamination in Black Box Language Model

Proving Test Set Contamination in Black Box Language Models

Yonatan Oren[1]*, **Nicole Meister**[1]*, **Niladri Chatterji**[1]*, **Faisal Ladhak**[2], **Tatsunori B. Hashimoto**[1]
[1]Stanford University, [2]Columbia University
yonatano@cs.stanford.edu
{nmeist, niladric, thashim}@stanford.edu
faisal@cs.columbia.edu

Jiayu Yao

Background



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? ✗

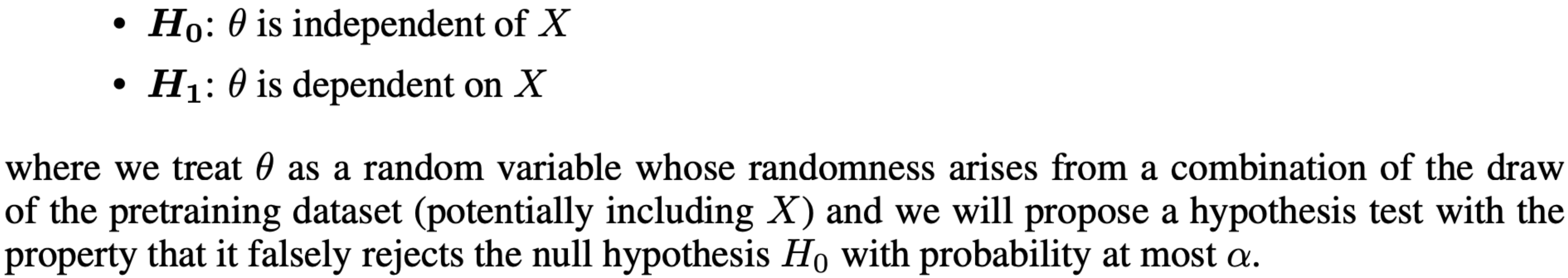
Is there a movie with 0 on rotten tomatoes? ✗

✓ high model log-probability ✗ low model log-probability

Differences in log-probability between orderings reveal contamination.

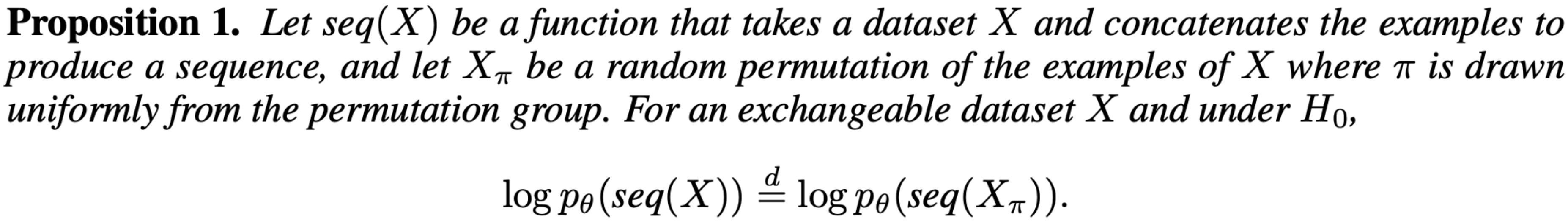
Figure 1: Given a pre-training dataset contaminated with the BoolQ(Clark et al., 2019) test set (left), we detect such contamination by testing for exchangability of the dataset (right). If a model has seen a benchmark dataset, it will have a preference for the canonical order (i.e. the order that examples are given in public repositories) over randomly shuffled examples orderings. We test for these differences in log probabilities, and aggregate them across the dataset to provide false positive rate guarantees.

Background



- $\boldsymbol{H_0}$: θ is independent of X
- $\boldsymbol{H_1}$: θ is dependent on X

where we treat θ as a random variable whose randomness arises from a combination of the draw of the pretraining dataset (potentially including X) and we will propose a hypothesis test with the property that it falsely rejects the null hypothesis H_0 with probability at most α.



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,*

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

Statstics

θ depend on X $\quad logp_\theta\left(seq(X)\right) \gg logp_\theta\left(seq(X_\pi)\right)$

Under Null hypothesis $\quad \left\{logp_\theta\left(seq(X_{\pi_1})\right), \cdots, logp_\theta\left(seq(X_{\pi_n})\right)\right\} \sim Uniform$

$$p := \mathbb{E}[\mathbb{1}\{\log p_\theta(\mathrm{seq}(X)) < \log p_\theta(\mathrm{seq}(X_\pi))\}].$$

Monte Carlo estimate

$$\hat{p} := \frac{\sum_{i=1}^{m} \mathbb{1}\{\log p_\theta(\mathrm{seq}(X)) < \log p_\theta(\mathrm{seq}(X_{\pi_m}))\} + 1}{m+1}$$

P-value $\quad 1/(m+1) \qquad \mathcal{O}\left(m|X|\right)$

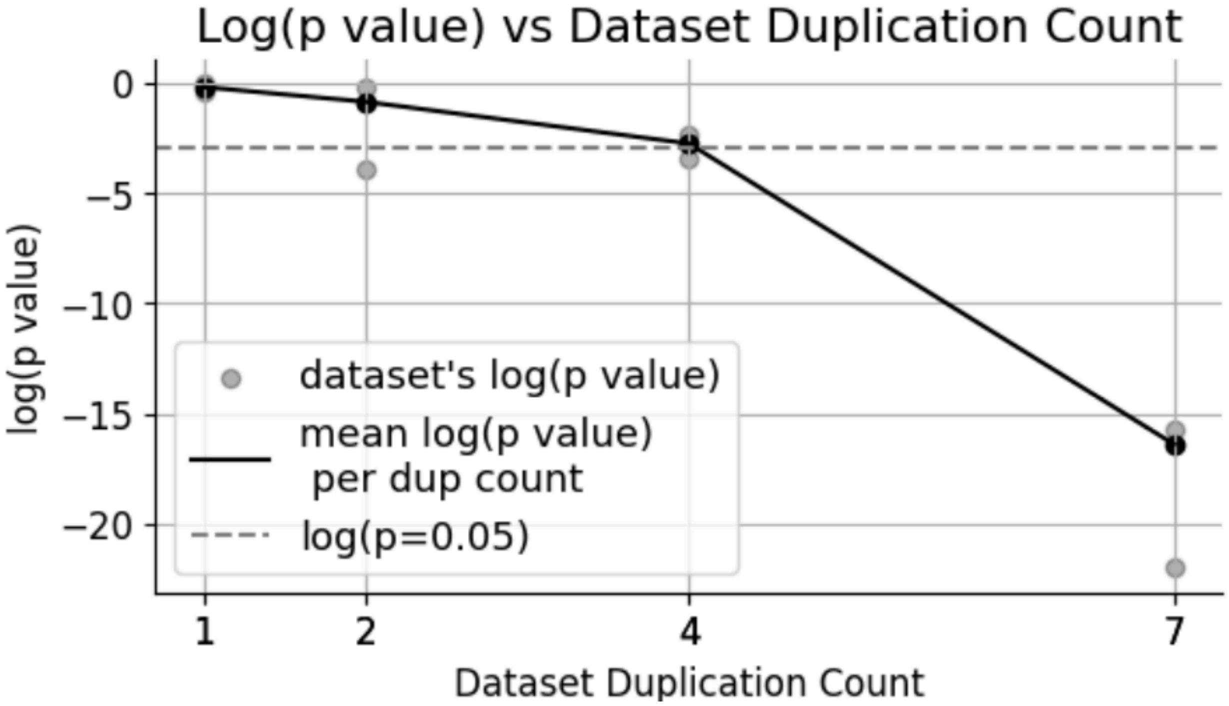
Statstics

$$S_1 = (X_1, X_2, \cdots, X_k)$$

Each Shards $s_i := \log p_\theta(\mathrm{seq}(X)) - \mathrm{Mean}_\pi(\log p_\theta(\mathrm{seq}(X_\pi))).$

$$s = \frac{1}{r}\sum_{i=1}^{r} s_i$$

Experiment


Log(p value) vs Dataset Duplication Count
log(p value)
0
−5
−10
−15
−20
1
2
4
7
Dataset Duplication Count
● dataset's log(p value)
mean log(p value) per dup count
log(p=0.05)

Experiment

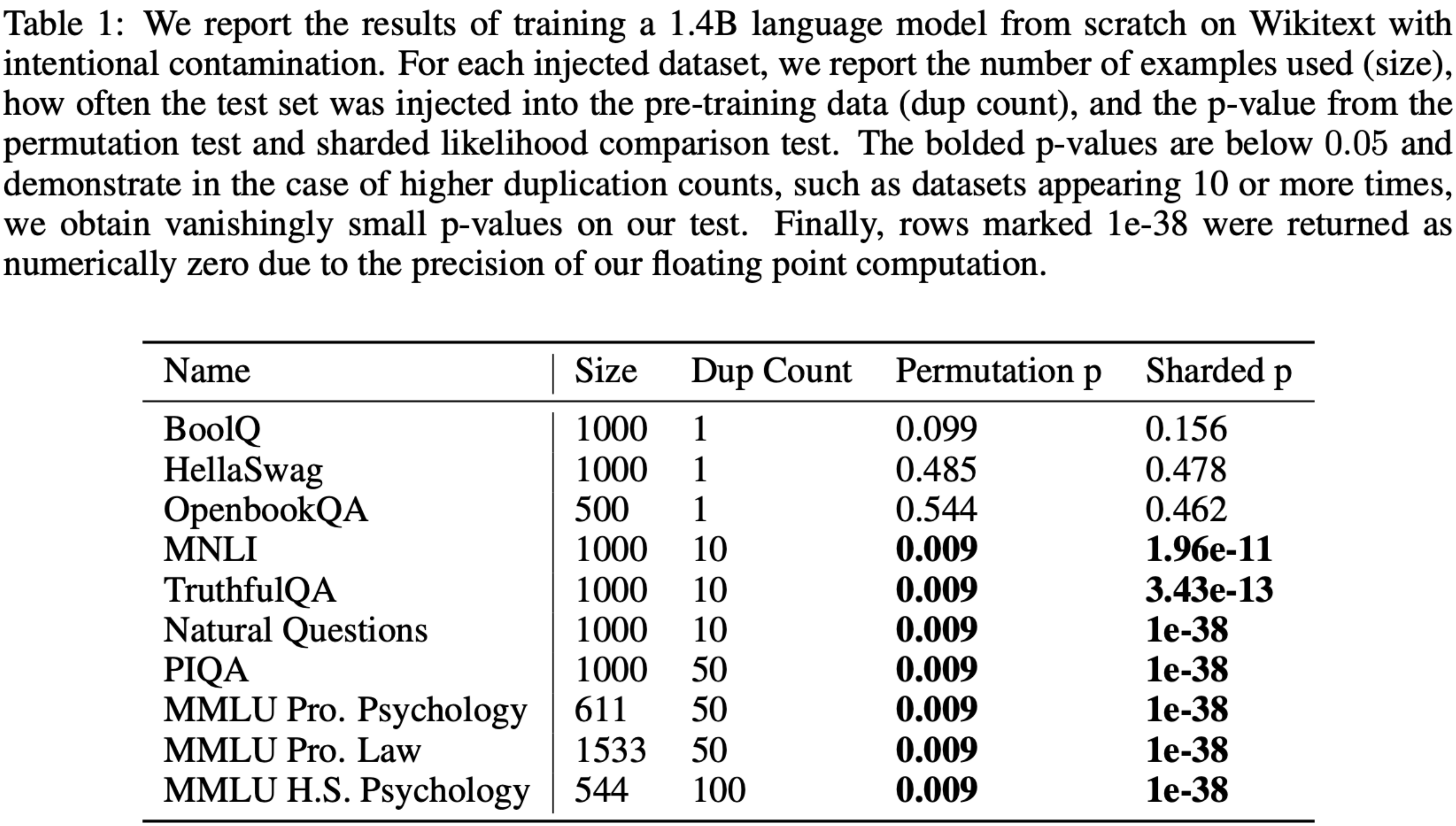


Table 1: We report the results of training a 1.4B language model from scratch on Wikitext with intentional contamination. For each injected dataset, we report the number of examples used (size), how often the test set was injected into the pre-training data (dup count), and the p-value from the permutation test and sharded likelihood comparison test. The bolded p-values are below 0.05 and demonstrate in the case of higher duplication counts, such as datasets appearing 10 or more times, we obtain vanishingly small p-values on our test. Finally, rows marked 1e-38 were returned as numerically zero due to the precision of our floating point computation.

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**

Experiment

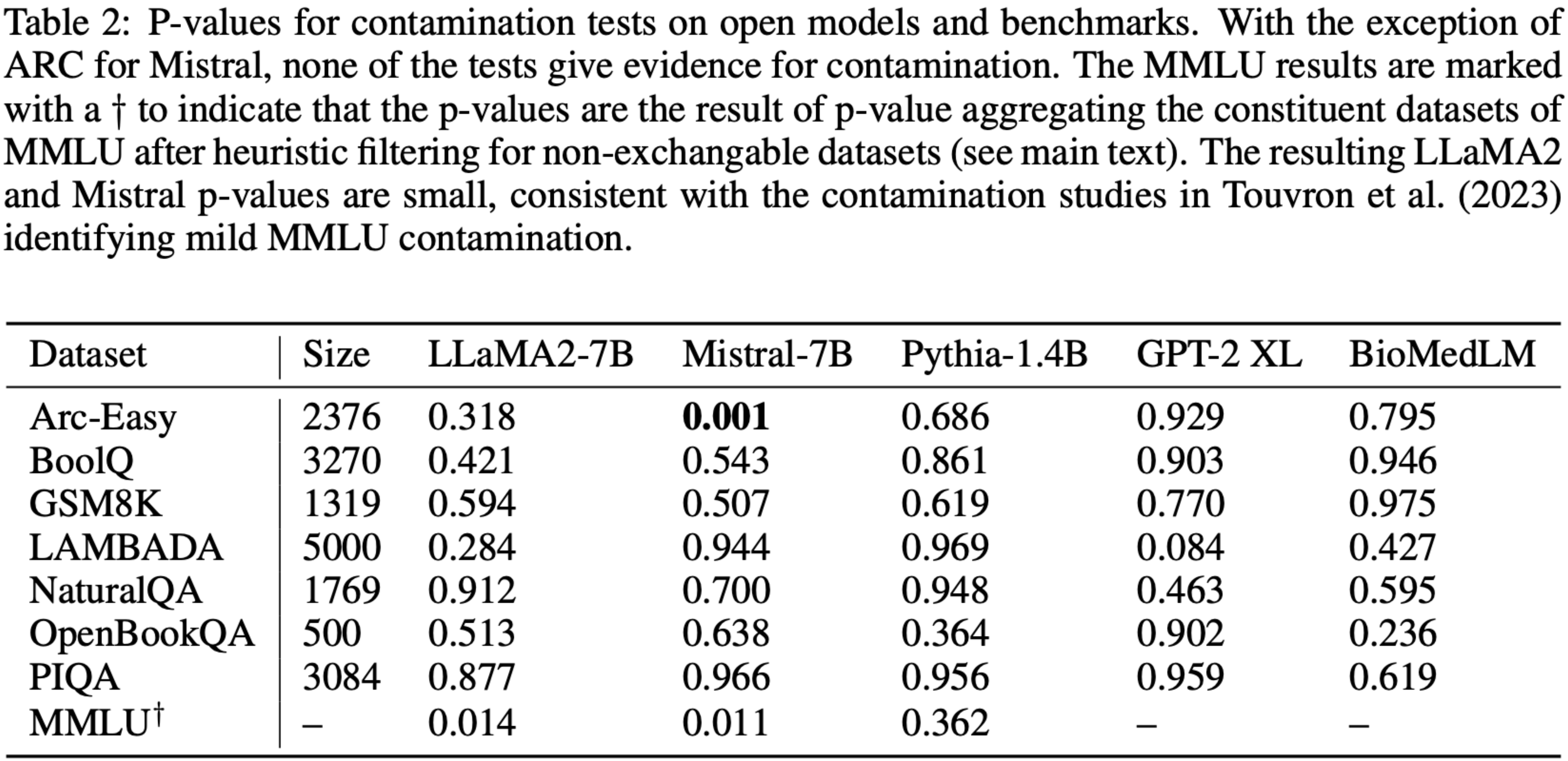


Table 2: P-values for contamination tests on open models and benchmarks. With the exception of ARC for Mistral, none of the tests give evidence for contamination. The MMLU results are marked with a † to indicate that the p-values are the result of p-value aggregating the constituent datasets of MMLU after heuristic filtering for non-exchangable datasets (see main text). The resulting LLaMA2 and Mistral p-values are small, consistent with the contamination studies in Touvron et al. (2023) identifying mild MMLU contamination.

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	–	–

Thanks