

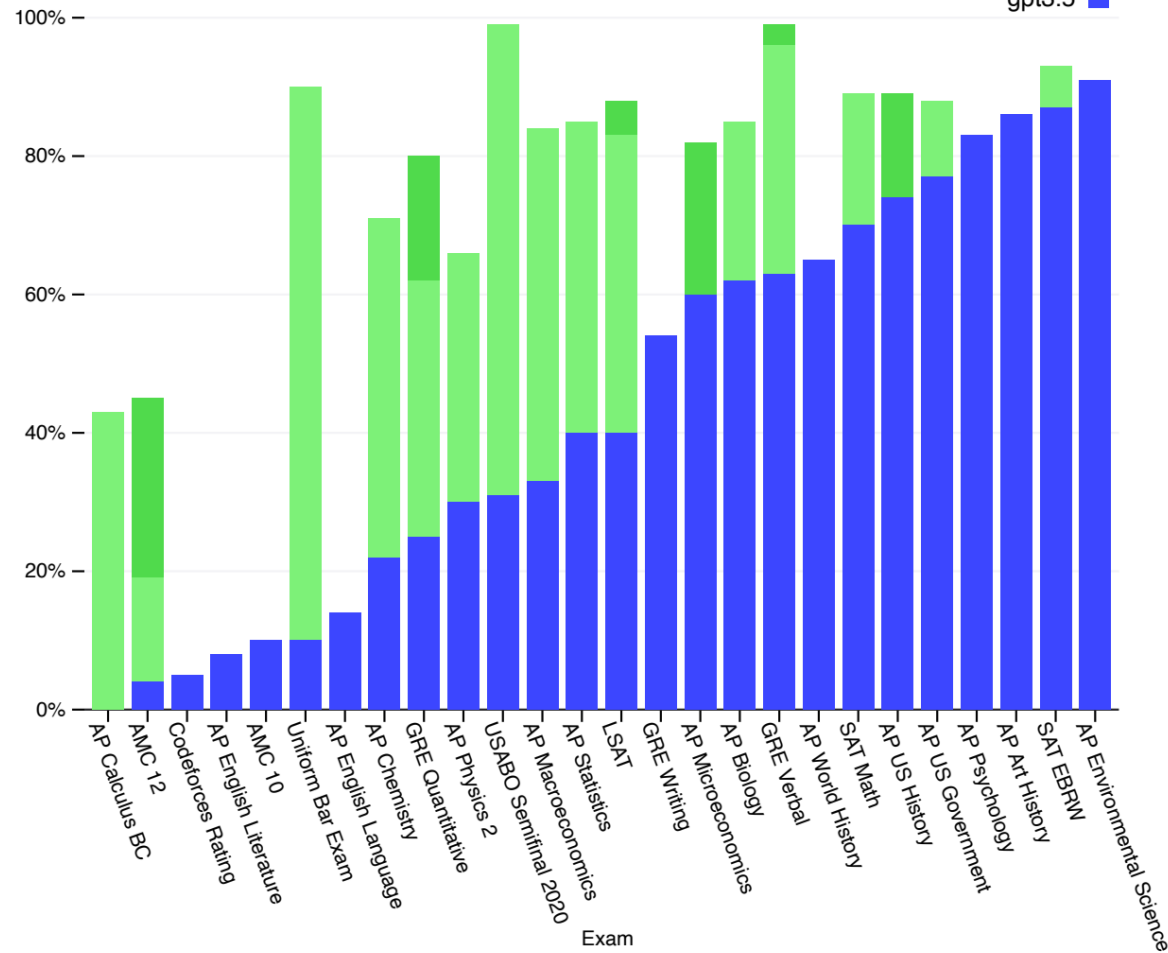
Proving Test Set Contamination in Black Box Language Models

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

Language models perform remarkably well on benchmarks

Exam results (ordered by GPT-3.5 performance)

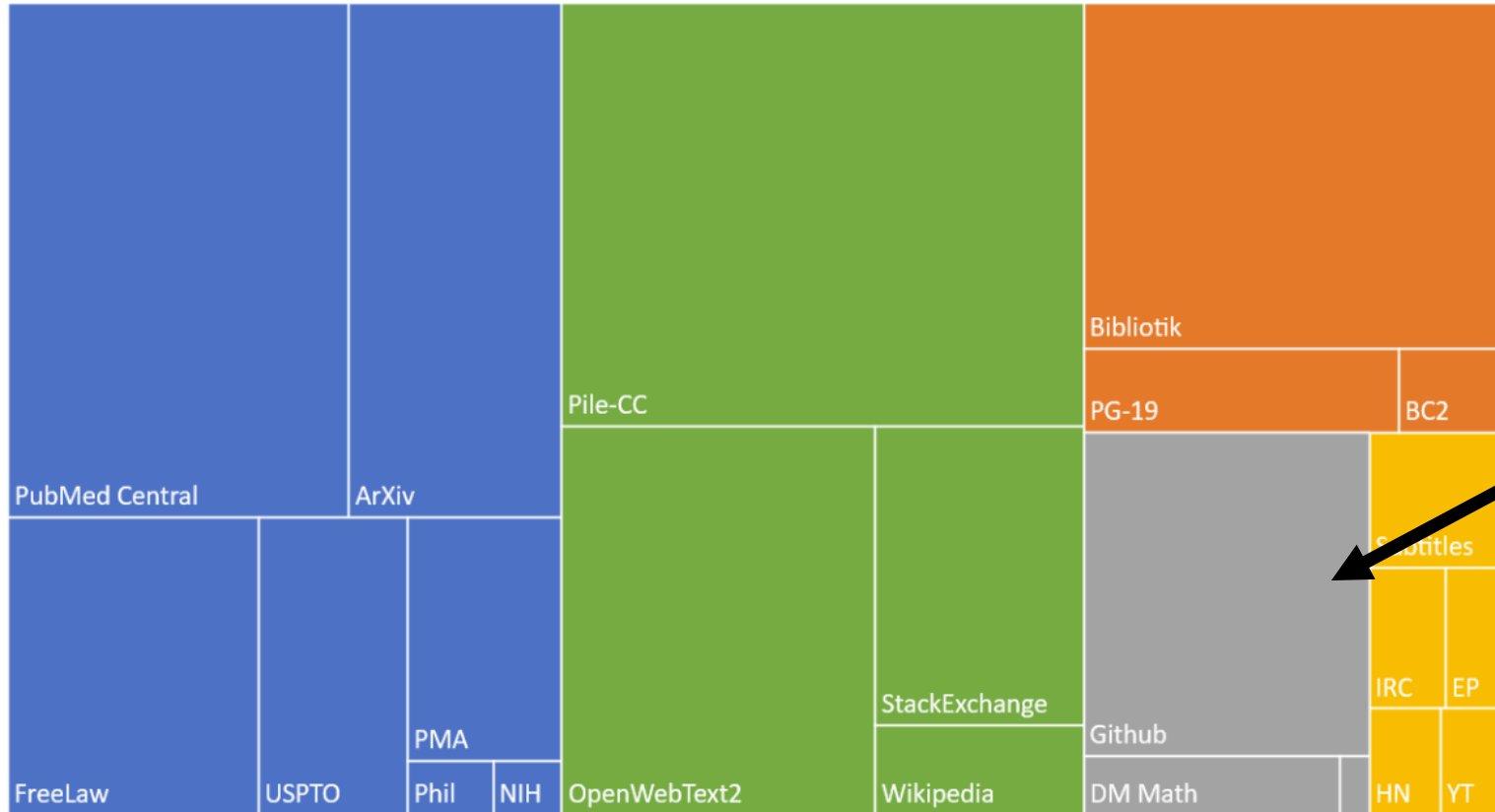
Estimated percentile lower bound (among test takers)



What's in a language model's training data?

Composition of the Pile by Category

■ Academic ■ Internet ■ Prose ■ Dialogue ■ Misc



... maybe your test set is in here?




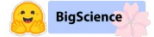
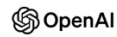







Modern pre-training datasets are massive, with minimal curation involved

How do we know closed LLMs are not seeing
(and memorizing) test sets in their training data?

Pre-training datasets are the “secret sauce”

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index

	 Meta	 BigScience	 OpenAI	 stability.ai	 Google	 ANTHROPIC	 cohere	 AI21labs	 Inflection	 amazon	Average
	Llama 2	BLOOMZ	GPT-4	Stable Diffusion 2	PaLM 2	Claude 2	Command	Jurassic-2	Inflection-1	Titan Text	
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	57%	52%	47%	47%	41%	39%	31%	20%	20%	13%	

Major Dimensions of Transparency

We need a way to audit closed language models for contamination



Horace He
@cHHillee

I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces.

Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems.

This strongly points to contamination.

1/4

g's Race	implementation, math	🚩 ⭐	greedy, implementation	🚩 ⭐
nd Chocolate	implementation, math	🚩 ⭐	Cat?	🚩 ⭐
triangle!	brute force, geometry, math	🚩 ⭐	Actions	🚩 ⭐
	greedy, implementation, math	🚩 ⭐	Interview Problem	🚩 ⭐

...



Susan Zhang
@suchenzang

I think Phi-1.5 trained on the benchmarks. Particularly, GSM8K.



Susan Zhang
@suchenzang · Sep 12

Let's take github.com/openai/grade-s...

If you truncate and feed this question into Phi-1.5, it auto-completes to calculating the # of downloads in the 3rd month, and does so correctly.

Change the number a bit, and it answers correctly as well.

1/🤖



Public claims and heuristic tests exist, but **without proof**

Existing work: promising, but lacks provable guarantees

Time Travel in LLMs: Tracing Data Contamination in Large Language Models (Golchin et. al.)

Min-K-Prob: Detecting Pre-training Data from Large Language Models (Shi et. al.)

To the Cutoff... and Beyond? A Longitudinal Perspective on LLM Data Contamination (Roberts et. al.)

Provably Detecting Test Set Contamination

Goal: Provide a provable (false positive rate) guarantee for detecting test set contamination.

Our Setup:

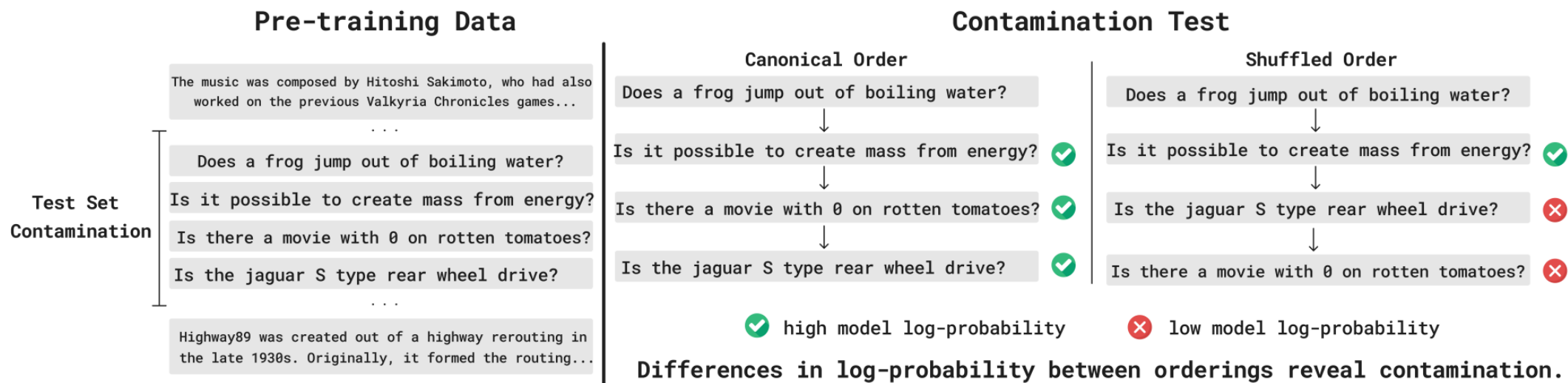
Given a test set X and access to log-probability queries from a language model θ , we want to test the null hypothesis,

H_0 : the test set X and the model θ are independent R.V.s.

We will present a test that falsely rejects H_0 with probability at most α .

Exploiting the exchangeability of datasets

Most datasets are *exchangeable*



Key Insight: a preference by the model for a “canonical” ordering of an exchangeable dataset must result from contamination.

A Statistical Test for Contamination

Permutation Test: shuffle and compute log probs

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

A contamination detector based on $\hat{p} < \alpha$ has a FP rate of at most α .

Sharded Rank Comparison Test: aggregates many smaller shuffled subsequences

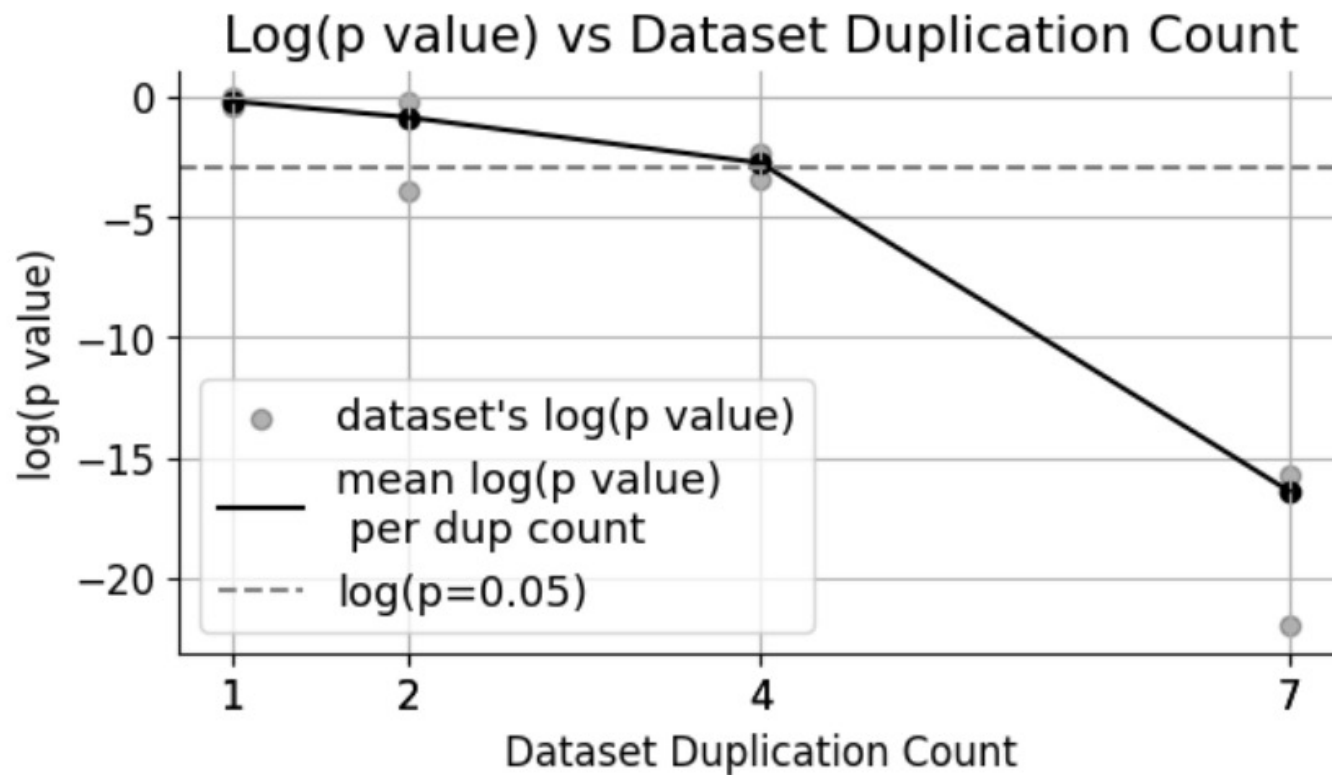
Pretraining with intentional contamination

We pretrain a 1.4B language model from scratch on 20B tokens from Wikipedia **with test sets injected randomly** at various duplication counts.

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

100% detection rate for duplication count ≥ 10

Detection at low duplication counts



Around 50% detection rate for 2-4 duplicates

Testing real models for 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	–	–

- **Did not find evidence of contamination (except for Mistral on Arc-Easy)**
- **MMLU results are consistent with preexisting contamination studies (Touvron et. al.)**

Takeaways

Proving test set contamination is (sometimes) possible.

Low dup count is a major open problem in contamination detection.

At high dup count, we do not see evidence of pervasive contamination.

Acknowledgements!

