NLP Evaluation in trouble!



On the Need to Measure LLM Data Contamination for each Benchmark

Oscar Sainz¹, Jon Ander Campos², Iker García-Ferrero¹, Julen Etxaniz¹, Oier Lopez de Lacalle¹ and Eneko Agirre¹

The data contamination problem

- Evaluation using annotated benchmarks is in trouble, e.g. LLMs trained on test split of a benchmark.
- Contamination can cause an overestimation of the performance with respect to their non-contaminated counterparts. Translated to wrong scientific conclusions being published while other correct ones are discarded.
- The extent of the problem is unknown.

In this position paper we argue for a community action:

1st Workshop on Data Contamination CONDA@ACL conda-workshop.github.io



- 1) Agree on automatic and semi-automatic methods to detect when data from a benchmark was exposed to a model.
- 2) Collate evidence in a collaborative index of contamination.
- 3) Flag papers that draw conclusions that are compromised by contamination.
- 4) Revise reviewing protocols to check for contamination issues.



		(90.0%) Containinated		
GPT-3	ANLI R1		(20.0%) Contaminated	Link
GLaM	ANLI R2	(96.8%) Contaminated		<u>Link</u>
FLAN	ANLI R2	(97.9%) Contaminated		Link

LLM Contamination Index

hitz-zentroa.github.io/Im-contamination

	Contam		Model	Subst pe	Avg. Contam. %	n	\overline{X}	μ_n	Zn
Uniform D. F.		Dataset		Clean	0	7391	80.0	82.5 -	5.73
(MBE MEE) (DT	0.7			Not Clean	67.5	2651	89.5	82.4	9.50
(MBE+MEE+MPT)	0%		70B	Not Clean	11.5	9194	81.6	82.5	-2.27
SAT E	39 %			Distric	86.1	848	92.2	82.5	7.42
SAI Evidence-Based Reading &	12 %	-7.11 C (I - 40)		Dirty	0	7391	70.5	73.3	-5.46
SATMAN		HellaSwag $(L = 40)$		Clean	67.5	2651	81.3	73.4	9.17
SAI Math	7%		7B	Not Clean	11.5	9194	72.4	73.4	-2.06
GRE Quantitative	35 %			Not Dirty	86.1	848	83.7	73.3	6.84
GRE Verbal	25 %			Dirty	0011	2006	62.2	65.3	-4.08
GRE Writing	100 %			Clean	0.05	700	82.7	65.3	9.71
USABO Semifinal Exam 2020	100 //			Not Clean	85.12	1195	62.7	65.3	-3.50
LISNCO I	3%		70B	Not Dirty	2.73	4165	95.8	65.3	9.80
Medical Kernel Section Exam 2022	5%			Dirty	94.5	520	40.8	42.9	-2.75
Self Assessment P	10.07	MMLU-Humanities $(L = 50)$		Clean	0.05	3996	5 40.0	42.9	6.50
Codeforces Bati	19 %	TALLAR -		Not Clean	85.2	709	- 41.1	42.0	-2.25
AP Art II'r	0 %		7B	Not Dirty	2.73	4185	5 41.1	42.9	6.49
AP Art History	17 %			Dirty	94.5	520	56.9	42.0	0.47
AP Biology	1%			Diry	0.02	11862	2 68.0	68.9	-2.00
AP Calculus BC	3%			Clean	84.7	218	0 73.5	68.9	4.64
AP Chemistry	16 %	11 (7 50)	70B	Not Clean	3.18	1250	6 67.7	68.9	-2.75
AP Eng. Lang. and Comp.	79 %	MMLU-Overall $(L = 50)$	100	Not Dirty	94.4	153	6 78.2	68.9	7.87
AP Eng. Lit. and Comp.	92 %			Dirty	74.4				
CDT-2	4%			FΙΛ	N				
GPT-3	4%			FLA	N				
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	4 % <i>Dotal</i> Dirty <i>Acc/F1/BLEU</i> 7353 44.3 11873 69.9 9536 69.9 9536 65.8 10000 90.3 66.8 10000 90.3 5270 90.2 13270 90.2 1953 75.8 15 953 75.8 15 953 73.4 15 953 73.4	Dirty Clean Clean Clean Count Acc/F1/BLEU Count Percentage 7315 54.1 38 1% 8898 29.5 638 7% 107 67.1 1435 7% 110 88.2 3867 14% 10 88.2 3890 39% 55 86.2 109 39% 58 76.3 1312 40%		FLA Dataset DROP SQUADV2 ANLI R1 ANLI R2 ReCoRD MultiRC PIOA	Metric Total count acc/F1// F1 9,536 22 F1 11,873 41 acc 1,000 48 acc 1,000 48 acc 1,000 42 acc 10,000 4 acc 4,848 75 acc 1,838 23	A Cle BLEU cou 4 6 .3 10 .1 1 .9 2 .9 2 .5 4 1, 3.7 8 3.7 8	ean (int acc/l i1 06 4 203 972 996 9972 996	Clean F1/BLEU 33.0 38.7 57.1 38.1 4.5 75.7 23.3 45.3	% clean 0.6 0.9 1.4 2.1 32.0 40.7 48.7 59.8

ChatGPT generating CoNLL03

ChatGPT generating ACE05



Self-assessment is not enough: External audit of contamination is needed.

Different contamination types

- Guideline contamination: happens when the annotation guidelines for a specific dataset are seen by the model. Potentially affecting zero and few-shot evaluations.
- Raw text contamination: happens when the original text (previous to annotation) is seen by the model. Using Wikipedia for example. Tasks where the labels are directly found in the original text are particularly affected.
- Annotation contamination: happens when the annotations (labels) of the target benchmark are exposed to the model during training. <u>The worst case scenario.</u>

Contamination on different steps

 Contamination after deployment: Recent LLM advancements prompted the research community to assess closed models on standard benchmarks. The financial success of closed systems hinges on model performance. Thus, companies actively audit user inputs and retrain systems for subpar task performance. Iterative improvements based on API access and user input expose these models to evaluation data.

Measuring contamination

We differentiate open and closed LLMs by whether the pretraining data is publicly available or not.

- Open LLMs: Most of the research on data contamination pre-training data with string-matching analyzes
- Pretraining: Gathering vast internet text, making complete filtering challenging. Avoiding absolute data contamination is unrealistic due to diverse testing datasets. Allowing researchers access to and queries on pretraining data can prevent corrupted evaluations.
- Instruction tuning: Machine-generated datasets like Self-Instruct, Unnatural Instructions, Alpaca data, or SharedGPT aim to distill closed models' capabilities into open alternatives. However, distilling from a contaminated model might pose a potential data contamination issue.

operations. Many tools have been developed to audit large-scale pretraining data: ROOTS (Piktus et al., 2023), Data Portraits (Marone and Durme, 2023) and WIMBD (Elazar et al., 2023) among others. However, there is no <u>currently agreed-upon methodology to measure the level</u> of contamination.

• Closed LLMs: The recent most popular models GPT-3, LLaMA, Bard, or GPT-4 do not share their pre-training data. Detecting data-contamination under this scenario is much more challenging. Fortunately, in the last couple of months several advances have been made (Golchin and Surdeanu, 2023; Yucheng, 2023; Shi et al., 2023; Oren et al., 2023). Still, <u>a common evaluation framework for</u> <u>contamination is needed</u> for comparative assessment.