NLP Evaluation in trouble!

On the Need to Measure LLM Data Contamination for each Benchmark

(1) Universidad del País Vasco Euskal Herriko Unibertsitatea | HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology (2) cohere

Oscar Sainz[1], Jon Ander Campos[2], Iker García-Ferrero[1], Julen Etxaniz[1], Oier Lopez de Lacalle[1] and Eneko Agirre[1]

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**:

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.

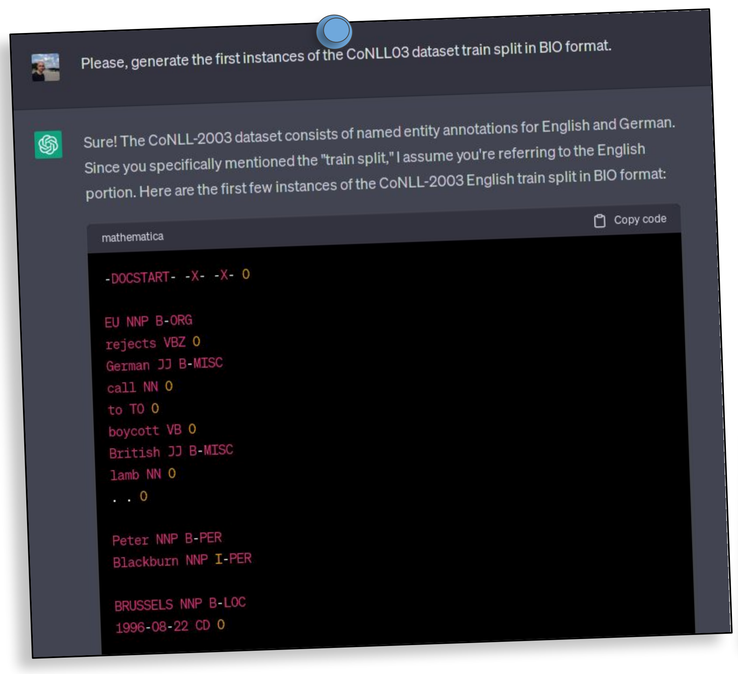
1st Workshop on Data Contamination CONDA@ACL

conda-workshop.github.io

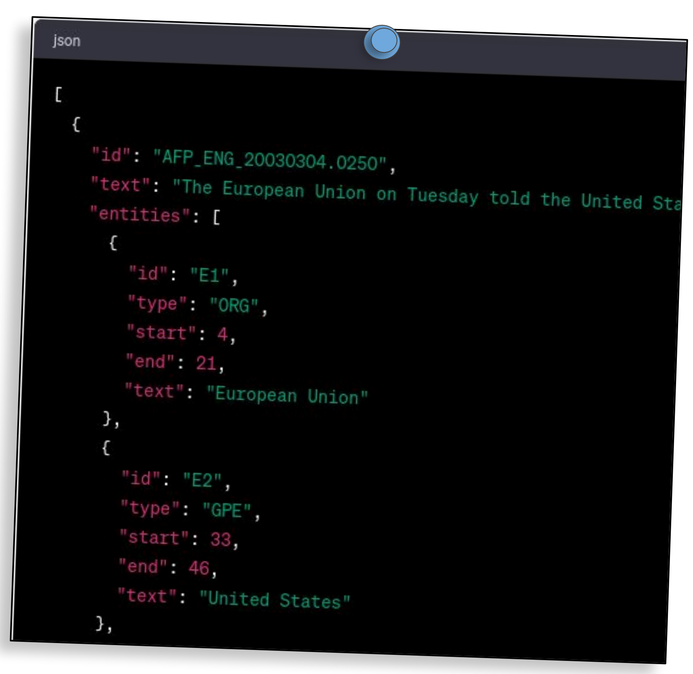
Model	Dataset ↑	Train split	Dev split	Test split	Source
ChatGPT	ACE05	Suspicious	Suspicious	Suspicious	
GPT-3.5	AG News	Clean		Clean	Link
GPT-4	AG News	Contaminated		Contaminated	Link
GLaM	ANLI R1		(96.2%) Contaminated		Link
FLAN	ANLI R1		(98.6%) Contaminated		Link
GPT-3	ANLI R1			(20.0%) Contaminated	Link
GLaM	ANLI R2		(96.8%) Contaminated		Link
FLAN	ANLI R2		(97.9%) Contaminated		Link

LLM Contamination Index

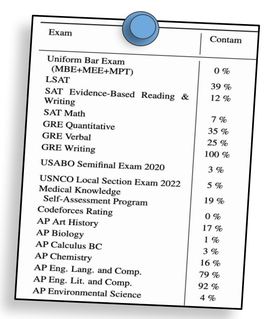
hitz-zentroa.github.io/lm-contamination



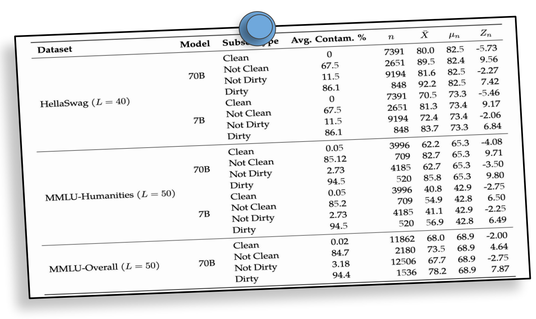
ChatGPT generating CoNLL03



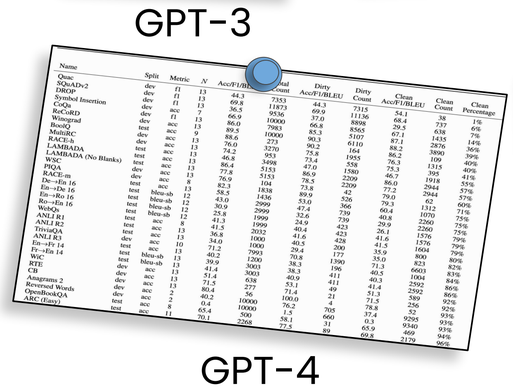
ChatGPT generating ACE05



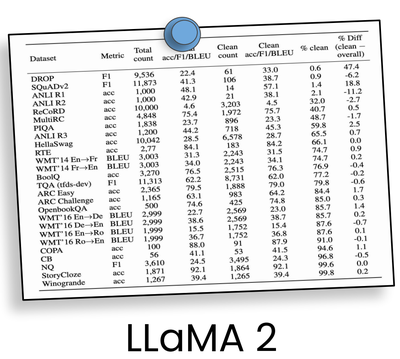
GPT-3



FLAN



GPT-4



LLaMA 2

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. The worst case scenario.

Contamination on different steps

- **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.
- **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 analyzes pre-training data with string-matching 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 currently agreed-upon methodology to measure the level 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, a common evaluation framework for contamination is needed for comparative assessment.