Data Contamination Report from the 2024 CONDA Shared Task

Oscar Sainz[1] **Iker García-Ferrero**[1] **Alon Jacovi**[2]
Jon Ander Campos[3] **Yanai Elazar**[4,5] **Eneko Agirre**[1] **Yoav Goldberg**[2,4]

Wei-Lin Chen[6,7] **Jenny Chim**[8] **Leshem Choshen**[9,10] **Luca D'Amico-Wong**[11]
Melissa Dell[11] **Run-Ze Fan**[12] **Shahriar Golchin**[13] **Yucheng Li**[14] **Pengfei Liu**[12]
Bhavish Pahwa[15] **Ameya Prabhu**[16,17] **Suryansh Sharma**[18] **Emily Silcock**[11]
Kateryna Solonko **David Stap**[19] **Mihai Surdeanu**[20] **Yu-Min Tseng**[21]
Vishaal Udandarao[22,23] **Zengzhi Wang**[12] **Ruijie Xu**[12] **Jinglin Yang**[11]

[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU [2]Bar Ilan University [3]Cohere
[4]Allen Institute for Artificial Intelligence [5]University of Washington [6]National Taiwan University
[7]University of Virginia [8]Queen Mary University of London [9]MIT-IBM Watson AI Lab [10]MIT
[11]Harvard University [12]Shanghai Jiao Tong University [13]University of Arizona [14]University of Surrey
[15]Microsoft Research [16]Tübingen AI Center [17]University of Tübingen
[18]Indian Institute of Technology Kharagpur [19]University of Amsterdam [20]University of Arizona
[21]National Taiwan University [22]University of Tuebingen [23]University of Cambridge

Contact: conda-workshop@googlegroups.com

Abstract

The 1st Workshop on Data Contamination (CONDA 2024) focuses on all relevant aspects of data contamination in natural language processing, where data contamination is understood as situations where evaluation data is included in pre-training corpora used to train large scale models, compromising evaluation results. The workshop fostered a shared task to collect evidence on data contamination in current available datasets and models. The goal of the shared task and associated database is to assist the community in understanding the extent of the problem and to assist researchers in avoiding reporting evaluation results on known contaminated resources. The shared task provides a structured, centralized public database for the collection of contamination evidence, open to contributions from the community via GitHub pool requests. This first compilation paper is based on 566 reported entries over 91 contaminated sources from a total of 23 contributors. The details of the individual contamination events are available in the platform.[1] The platform continues to be online, open to contributions from the community.

1 Introduction

Data contamination, where evaluation data is inadvertently included in pre-training corpora of large-scale models, and language models (LMs) in particular, has become a concern in recent times (Sainz et al., 2023a; Jacovi et al., 2023). The growing scale of both models and data, coupled with massive web crawling, has led to the inclusion of segments from evaluation benchmarks in the pre-training data of LMs (Dodge et al., 2021; OpenAI et al., 2024; Anil et al., 2023; Elazar et al., 2024). The scale of internet data makes it difficult to prevent this contamination from happening, or even detect when it has happened (Bommasani et al., 2022; Mitchell et al., 2023).

Crucially, when evaluation data becomes part of pre-training data, it introduces biases and can artificially inflate the performance of LMs on specific tasks or benchmarks (Magar and Schwartz, 2022; Magnusson et al., 2023; Merrill et al., 2024). This poses a challenge for fair and unbiased evaluation of models, as their performance may not accurately reflect their generalization capabilities (Hupkes et al., 2023).

Although a growing number of papers and state-of-the-art models mention issues of data contamination (Brown et al., 2020; Wei et al., 2022; Chowdhery et al., 2022; OpenAI et al., 2024; Anil et al., 2023; Touvron et al., 2023), there is little in the way of organized and compiled knowledge about real, documented cases of contamination in practice (Sainz et al., 2023a). Addressing data contamination is a shared responsibility among researchers, developers, and the broader community.

This report compiles the evidence reported in the

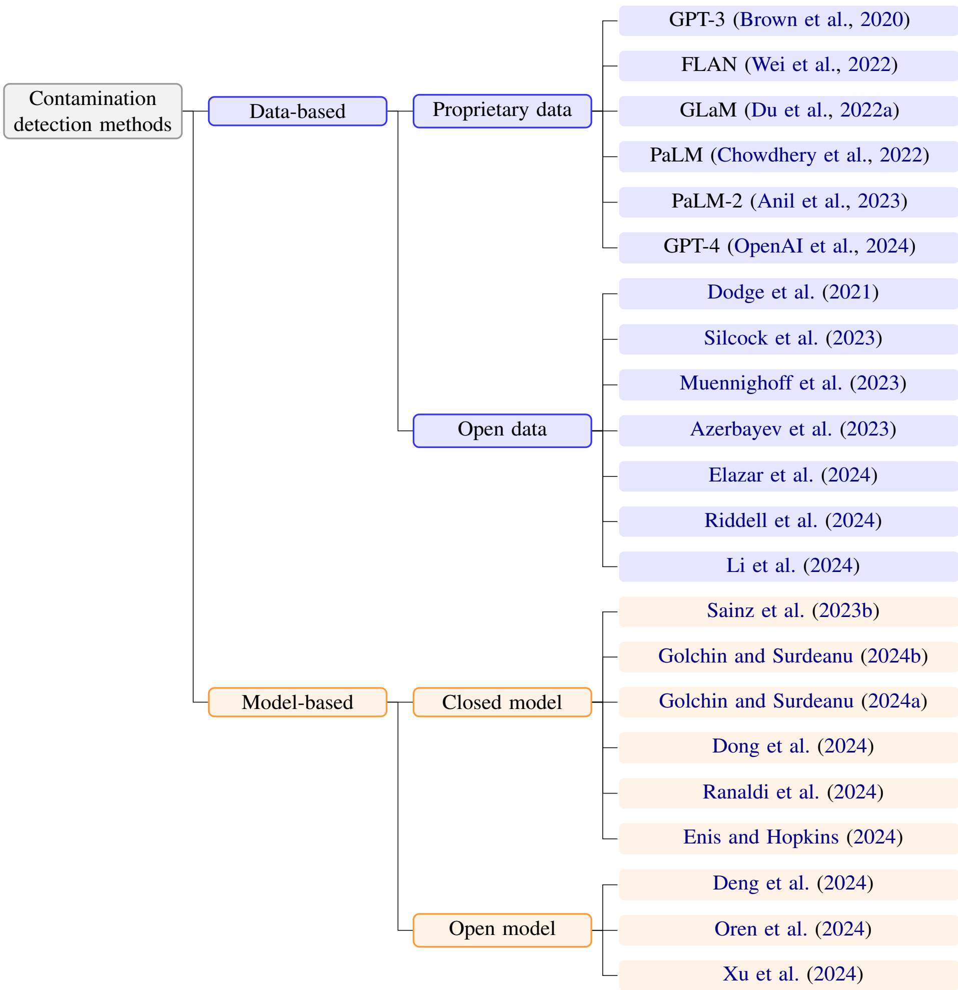

Figure 1: Taxonomy of papers that report contamination evidence. Including LLM's papers and technical reports, papers about methods for detecting contamination, and papers about corpus analysis.

Data Contamination Database[1] as part of the Data Contamination Workshop.[2] As the Shared Task of the workshop, researchers were invited to discover cases of contamination in available corpora and models, and submit evidence of their discovery. The submissions to the database were collected and compiled on June 23rd, 2024, to be included in this report, but the database continues to run and grow. Overall we collected 566 submissions from 23 contributors, where each submission included a detailed contamination report, indicating the estimated percentage of contaminated data. We continue to operate the database, and expect to update it with newer datasets and models as they come out, as well as new report about existing contaminated (or uncontaminated) evaluations.

This report first presents the methodology for collecting evidence, as well as existing papers that report data contamination (Section 2). We also report the evidence collected in the Data Contamination Database (Section 3), followed by an overview of the trends and statistics in the database, that inform a high-level perspective on the state of data contamination in NLP today (Section 4).

2 Methodology and Previous Work

Collecting all the contamination evidence —or lack of it— was done openly, through pull requests, and subject to discussions before the admission. Contributors were asked to fill in the information about several aspects, such as the *contaminated resource* (a training corpus or model), the *evaluation dataset* which was found in the contaminated source, a breakdown of the percentage of contamination found in each split of the dataset (train, development, and test), an optional reference to a paper that describes the methodology behind the

[1] https://huggingface.co/spaces/CONDA-Workshop/Data-Contamination-Database

[2] https://conda-workshop.github.io/

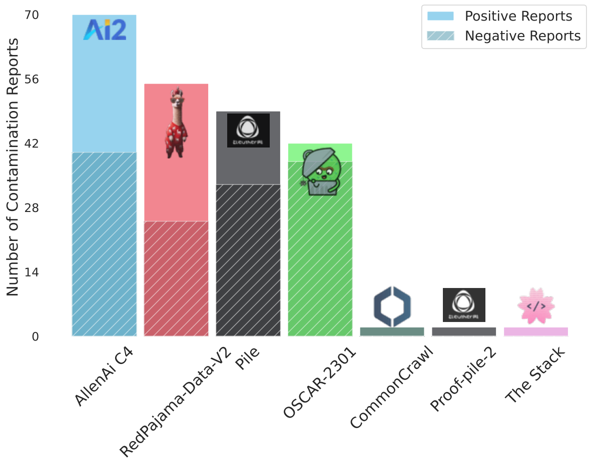

Figure 2: Number of test sets reported for each corpus often used in pre-training.

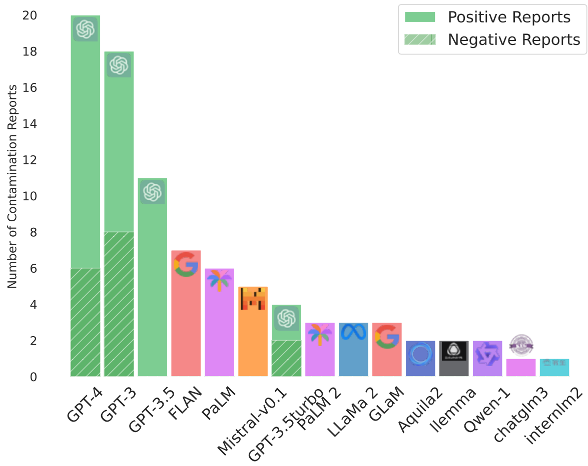

Figure 3: Number of test sets reported for each pre-trained model.

submission, as well as whether the contamination detection method was *data-based* or *model-based*. The contributions provided the HuggingFace Hub id of models, corpus, and datasets when possible. In addition, contributors must provide the evidence or a reference to the scientific paper that reported the evidence originally. Figure 1 shows the taxonomy of the papers that reported contamination evidence in the shared task.[3] We split the these methods into two: *data-based* and *model-based* approaches.

Data-based approaches are methods that inspect the pre-training corpora to find contamination evidence. Data-based approaches typically involve string or sub-string matching techniques such as 13-gram overlap (Brown et al., 2020; Wei et al., 2022), 50-character overlap (OpenAI et al., 2024) or even full-string overlap (Elazar et al., 2024). In Figure 1 we differentiate between *Proprietary* and *Open* data. Papers that fall in the category of *Proprietary data* are usually LLMs technical reports that run post-hoc data contamination evaluations to identify and remove evaluation instances that appear in the pre-training corpora (Brown et al., 2020; Wei et al., 2022; OpenAI et al., 2024). Papers that fall in the *open data* category usually involve corpus analysis tools (Dodge et al., 2021; Elazar et al., 2024) or LLMs with publicly available pre-training data (Azerbayev et al., 2023).

Model-based approaches are those methods that try to estimate the contamination of a model by prompting or analyzing the output, without accessing the pre-training data. These methods are formulated as Membership Inference Attacks (MIA) and range from asking LLMs to generate verbatim of the actual evaluation data (Sainz et al., 2023b; Golchin and Surdeanu, 2024b) to analyzing the actual output probabilities given by the model (Oren et al., 2024). We differentiate between methods applicable to *closed* and *open* models. Methods applicable to *closed* models are usually applicable to *open* models, but not the other way around due to the limitations established by the API or interface providers.

The collected evidences come from different approaches and sources, making them hardly comparable. For transparency, we included in the database information about the source of the evidence and the link to the discussion. We encourage the users to assess how the evidence was collected for their datasets of interest.

3 Compilation of Evidence

The report includes 42 contaminated sources (training corpora or models), 91 datasets, and 566 contamination entries, including 432 contamination events (20 train-set, 95 dev-set, 317 test-set) and 144 non-contamination events, where a contamination event is taken as any report above 0% of contamination. The database contains, for each split (train, dev, and test) of each evaluation dataset, what percentage was found to be contaminated by a subset of the contamination sources (corpora or models). We analyze separately the contaminated corpora and models.

[3] Note that there are many other works on data contamination detection. In this report we focus on works that were used to detect contamination for this report. We leave a more detailed coverage survey for future work.

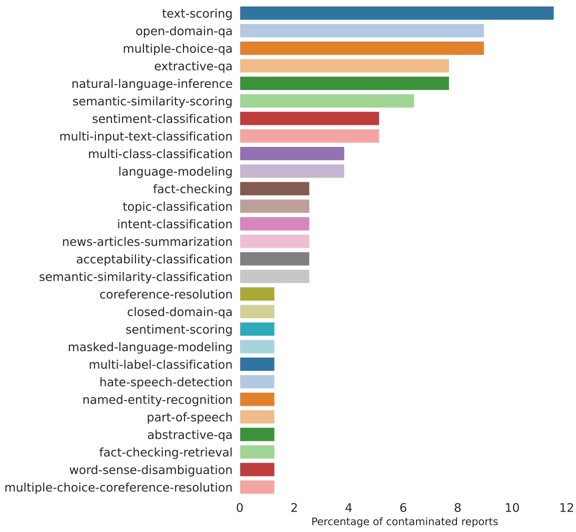

Figure 4: Percentage of contaminated report per task

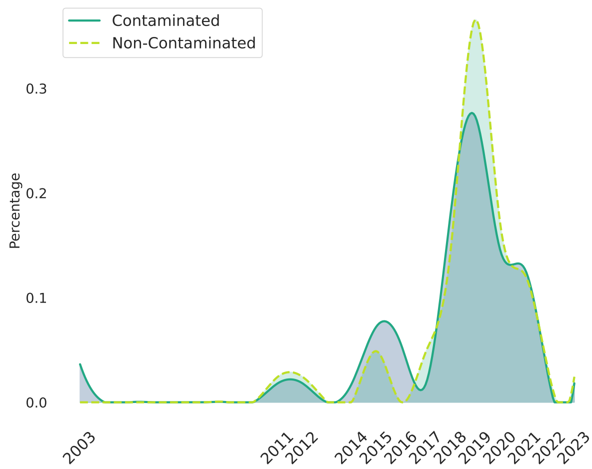

Figure 5: Publication year of the test sets included in the data contamination report.

Contaminated corpora. Figure 2 shows the number of reported test sets for each corpus often used to pre-train language models. The reported corpora are mainly based on CommonCrawl snapshots, GitHub, or a mix of sources. For CommonCrawl-based corpora, there are 35 events reported for C4 (Raffel et al., 2023), 32 for RedPajama v2 (Computer, 2023), 29 for OSCAR (Jansen et al., 2022; Abadji et al., 2022, 2021; Kreutzer et al., 2022; Ortiz Su'arez et al., 2020; Ortiz Su'arez et al., 2019) and 6 for CommonCrawl (Rana, 2010) itself. Regarding the GitHub data, there are 2 events reported for the TheStack (Kocetkov et al., 2022) project. The corpora with various sources, the Pile (Gao et al., 2020) and ProofPile (Azerbayev et al., 2023), have 30 and 2 reported contamination events respectively. There is also 1 report for xP3 (Muennighoff et al., 2022), which is a collection of prompts for different NLP datasets.[4]

Table 1 shows for each corpus often used to pre-train language models, the contamination events involving development or test splits. Please refer to the online database for full details of each report.

Contaminated models. Figure 3 details the number of contamination events involving test sets that were reported, organised according to each pre-trained model. Most reported evidence is for closed models, for instance: 24 for GPT-3 (Brown et al., 2020), 17 for GLaM (Du et al., 2022a), 16 for GPT-4 (OpenAI et al., 2024), 13 for GPT-3.5 (Brown et al., 2020), 8 for PaLM (Chowdhery et al., 2022), 3 for PaLM-2 (Anil et al., 2023), 2 for GPT-3.5 Turbo (Brown et al., 2020) and 1 for Calude 3 Opus. In the case of open models: there are 14 reported events for models fine-tuned with FLAN data (Wei et al., 2022), 5 for Mistral (Jiang et al., 2023), 3 for Llama 2 (Touvron et al., 2023), 2 for Qwen (Bai et al., 2023), Llema (Azerbayev et al., 2023) and Aquila 2; and a single one for mT0 and Bloom-Z (Muennighoff et al., 2022).

Table 2 shows for each pre-trained language model, the contamination events involving development or test splits. Please refer to the online database for full details of each report.

4 Noteworthy Trends and Statistics

In this section, we analyze the report to identify trends in the data that could lead to better identification of compromised evaluation datasets or to prevent data contamination in the first place.

Which are the most contaminated tasks? Figure 4 shows the percentage of data contamination per task. We use the task_id assigned to each dataset in the Hugging Face hub. Text-scoring, QA, and multiple-choice-qa are among the most contaminated task types. These types of tasks include very popular datasets such as MMLU (multiple-choice-qa), GLUE (text-scoring), and ai2_arc (multiple-choice-qa), which are standard benchmarks for measuring the performance of LLMs. These benchmarks are implemented in community leaderboards

[4]The report indicates the use of validation data from a specific dataset as training.

Contaminated Source	Evaluation Set
allenai/c4 (Raffel et al., 2023),	sem_eval_2014_task_1 (Marelli et al., 2014), race, nyu-mll/glue (Wang et al., 2019b), amazon_reviews_multi (Keung et al., 2020), liar (Wang, 2017), reddit_tifu (Kim et al., 2018), stsb_multi_mt (May, 2021), wiki_qa (Yang et al., 2015), gigaword (Graff et al., 2003), piqa (Bisk et al., 2020), esnli (Camburu et al., 2018), scitail (Khot et al., 2018), snli (Bowman et al., 2015), ibm/duorc (Saha et al., 2018), math_qa (Amini et al., 2019), swag (Zellers et al., 2018), wiki_bio (Lebret et al., 2016), xnli (Conneau et al., 2018), allenai/scicite (Cohan et al., 2019), aeslc (Zhang and Tetreault, 2019), billsum (Kornilova and Eidelman, 2019), AMR-to-Text, winograd_wsc (Levesque et al., 2012), squadshifts (Miller et al., 2020), head_qa (Vilares and Gómez-Rodríguez, 2019), xsum (Narayan et al., 2018), health_fact (Kotonya and Toni, 2020), EdinburghNLP/xsum (Narayan et al., 2018), UCLNLP/adversarial_qa (Bartolo et al., 2020), paws (Zhang et al., 2019), sick, super_glue (Wang et al., 2019a), paws-x (Yang et al., 2019), scan, lama (Petroni et al., 2019, 2020)
CommonCrawl (Rana, 2010)	allenai/ai2_arc (Clark et al., 2018), tau/commonsense_qa (Talmor et al., 2019), ceval/ceval-exam (Huang et al., 2023), cais/mmlu (Hendrycks et al., 2021a), Rowan/hellaswag (Zellers et al., 2019), winogrande (Levesque et al., 2012)
EleutherAI/pile (Gao et al., 2020)	sem_eval_2014_task_1 (Marelli et al., 2014), nyu-mll/glue (Wang et al., 2019b), amazon_reviews_multi (Keung et al., 2020), mbpp, openai_humaneval (Chen et al., 2021), liar (Wang, 2017), stsb_multi_mt (May, 2021), wiki_qa (Yang et al., 2015), gigaword (Graff et al., 2003), piqa (Bisk et al., 2020), esnli (Camburu et al., 2018), scitail (Khot et al., 2018), snli (Bowman et al., 2015), ibm/duorc (Saha et al., 2018), swag (Zellers et al., 2018), xnli (Conneau et al., 2018), allenai/scicite (Cohan et al., 2019), aeslc (Zhang and Tetreault, 2019), billsum (Kornilova and Eidelman, 2019), winograd_wsc (Levesque et al., 2012), squadshifts (Miller et al., 2020), head_qa (Vilares and Gómez-Rodríguez, 2019), xsum (Narayan et al., 2018), health_fact (Kotonya and Toni, 2020), UCLNLP/adversarial_qa (Bartolo et al., 2020), paws (Zhang et al., 2019), sick, super_glue (Wang et al., 2019a), paws-x (Yang et al., 2019), scan
oscar-corpus/OSCAR-2301 (Jansen et al., 2022; Abadji et al., 2022, 2021; Kreutzer et al., 2022; Ortiz Su'arez et al., 2020; Ortiz Su'arez et al., 2019)	sem_eval_2014_task_1 (Marelli et al., 2014), crows_pairs (Nangia et al., 2020), nyu-mll/glue (Wang et al., 2019b), race, amazon_reviews_multi (Keung et al., 2020), openai_humaneval (Chen et al., 2021), liar (Wang, 2017), stsb_multi_mt (May, 2021), wiki_qa (Yang et al., 2015), gigaword (Graff et al., 2003), piqa (Bisk et al., 2020), esnli (Camburu et al., 2018), scitail (Khot et al., 2018), snli (Bowman et al., 2015), math_qa (Amini et al., 2019), swag (Zellers et al., 2018), xnli (Conneau et al., 2018), allenai/scicite (Cohan et al., 2019), aeslc (Zhang and Tetreault, 2019), billsum (Kornilova and Eidelman, 2019), winograd_wsc (Levesque et al., 2012), squadshifts (Miller et al., 2020), head_qa (Vilares and Gómez-Rodríguez, 2019), xsum (Narayan et al., 2018), health_fact (Kotonya and Toni, 2020), UCLNLP/adversarial_qa (Bartolo et al., 2020), paws (Zhang et al., 2019), sick, super_glue (Wang et al., 2019a)
togethercomputer/RedPajama-Data-V2 (Computer, 2023)	sem_eval_2014_task_1 (Marelli et al., 2014), race, nyu-mll/glue (Wang et al., 2019b), amazon_reviews_multi (Keung et al., 2020), liar (Wang, 2017), stsb_multi_mt (May, 2021), wiki_qa (Yang et al., 2015), gigaword (Graff et al., 2003), piqa (Bisk et al., 2020), esnli (Camburu et al., 2018), scitail (Khot et al., 2018), snli (Bowman et al., 2015), ibm/duorc (Saha et al., 2018), math_qa (Amini et al., 2019), swag (Zellers et al., 2018), xnli (Conneau et al., 2018), allenai/scicite (Cohan et al., 2019), aeslc (Zhang and Tetreault, 2019), billsum (Kornilova and Eidelman, 2019), winograd_wsc (Levesque et al., 2012), squadshifts (Miller et al., 2020), head_qa (Vilares and Gómez-Rodríguez, 2019), xsum (Narayan et al., 2018), health_fact (Kotonya and Toni, 2020), UCLNLP/adversarial_qa (Bartolo et al., 2020), mc_taco, paws (Zhang et al., 2019), samsum (Gliwa et al., 2019), sick, super_glue (Wang et al., 2019a), paws-x (Yang et al., 2019), scan
bigscience/xP3 (Muennighoff et al., 2022)	facebook/flores (NLLB-Team et al., 2022)
EleutherAI/proof-pile-2 (Azerbayev et al., 2023)	gsm8k (Cobbe et al., 2021), hendrycks/competition_math (Hendrycks et al., 2021b)
bigcode/the-stack (Kocetkov et al., 2022)	openai_humaneval (Chen et al., 2021), mbpp

Table 1: A summary of the *dev* or *test* sets found at above 0% contamination in each corpus often used to pre-train models.

such as the Open LLM Leaderboard.[5]

Are older datasets more compromised than newer datasets? Figure 5 shows the percentage of total test sets included in contamination events per year. We present data for test sets in both contamination events (>0% contamination) and non-contamination events (0% contamination). As shown in the figure, older datasets are more likely to be compromised, while newer datasets are more likely to be reported as non-contaminated. However, from 2021 to 2023, the percentages of datasets reported as compromised and non-compromised are very similar. Thus, using newer datasets is not always an effective method to prevent data contamination.

We further explore the relationship between the year of publication of the datasets and instances of contamination by examining the reported data contamination for the three models with the most instances of data contamination: GPT-4, GPT-3, and

[5] https://hf.co/spaces/open-llm-leaderboard/

Contaminated Source	Evaluation Set
GPT-3 (Brown et al., 2020)	Reversed Words , race , quac (Choi et al., 2018), Anagrams 1 , Cycled Letters , mandarjoshi/trivia_qa (Joshi et al., 2017), ibragim-bad/arc_easy (Clark et al., 2018), SAT Analogies, piqa (Bisk et al., 2020), Rowan/hellaswag (Zellers et al., 2019), wmt/wmt16 (Bojar et al., 2016), stanfordnlp/coqa (Reddy et al., 2019), cimec/lambada (Paperno et al., 2016), natural_questions (Kwiatkowski et al., 2019), winograd_wsc (Levesque et al., 2012), ucinlp/drop (Dua et al., 2019), rmanluo/RoG-webqsp, rajpurkar/squad_v2 (Rajpurkar et al., 2018, 2016), allenai/openbookqa (Mihaylov et al., 2018), Symbol Insertion, Anagrams 2 , super_glue (Wang et al., 2019a), ibragim-bad/arc_challenge (Clark et al., 2018), facebook/anli (Nie et al., 2020)
GPT-3.5 (Brown et al., 2020)	samsum (Gliwa et al., 2019), yelp_review_full (Zhang et al., 2015a), imdb (Maas et al., 2011), ag_news (Zhang et al., 2015b), nyu-mll/glue (Wang et al., 2019b), conll2003 (Tjong Kim Sang and De Meulder, 2003), winogrande (Levesque et al., 2012), rajpurkar/squad_v2 (Rajpurkar et al., 2018, 2016), cais/mmlu (Hendrycks et al., 2021a), EdinburghNLP/xsum (Narayan et al., 2018), allenai/openbookqa (Mihaylov et al., 2018), xlangai/spider (Yu et al., 2018), truthful_qa (Lin et al., 2022)
GPT-4 (OpenAI et al., 2024)	samsum (Gliwa et al., 2019), yelp_review_full (Zhang et al., 2015a), gsm8k (Cobbe et al., 2021), imdb (Maas et al., 2011), ibragim-bad/arc_challenge (Clark et al., 2018), nyu-mll/glue (Wang et al., 2019b), ucinlp/drop (Dua et al., 2019), winogrande (Levesque et al., 2012), openai_humaneval (Chen et al., 2021), ag_news (Zhang et al., 2015b), EdinburghNLP/xsum (Narayan et al., 2018), cais/mmlu (Hendrycks et al., 2021a), Rowan/hellaswag (Zellers et al., 2019), allenai/openbookqa (Mihaylov et al., 2018), truthful_qa (Lin et al., 2022), bigbench (Srivastava et al., 2023)
PaLM 2 (Anil et al., 2023)	EdinburghNLP/xsum (Narayan et al., 2018), csebuetnlp/xlsum (Hasan et al., 2021), wiki_lingua (Ladhak et al., 2020)
GPT-3.5-turbo (Brown et al., 2020)	openai_humaneval (Chen et al., 2021), HumanEval_R (Chen et al., 2021)
FLAN (Wei et al., 2022)	natural_questions (Kwiatkowski et al., 2019), mandarjoshi/trivia_qa (Joshi et al., 2017), story_cloze (Sharma et al., 2018), piqa (Bisk et al., 2020), super_glue (Wang et al., 2019a), ibragim-bad/arc_challenge (Clark et al., 2018), ucinlp/drop (Dua et al., 2019), rajpurkar/squad_v2 (Rajpurkar et al., 2018, 2016), ibragim-bad/arc_easy (Clark et al., 2018), Rowan/hellaswag (Zellers et al., 2019), allenai/openbookqa (Mihaylov et al., 2018), facebook/anli (Nie et al., 2020), winogrande (Levesque et al., 2012), wmt/wmt16 (Bojar et al., 2016)
GLaM (Du et al., 2022a)	stanfordnlp/coqa (Reddy et al., 2019), natural_questions (Kwiatkowski et al., 2019), mandarjoshi/trivia_qa (Joshi et al., 2017), story_cloze (Sharma et al., 2018), cimec/lambada (Paperno et al., 2016), piqa (Bisk et al., 2020), super_glue (Wang et al., 2019a), ibragim-bad/arc_challenge (Clark et al., 2018), race , quac (Choi et al., 2018), winograd_wsc (Levesque et al., 2012), rajpurkar/squad_v2 (Rajpurkar et al., 2018, 2016), ibragim-bad/arc_easy (Clark et al., 2018), Rowan/hellaswag (Zellers et al., 2019), allenai/openbookqa (Mihaylov et al., 2018), facebook/anli (Nie et al., 2020), winogrande (Levesque et al., 2012)
LLaMa 2-13B (Touvron et al., 2023)	allenai/openbookqa (Mihaylov et al., 2018), winogrande (Levesque et al., 2012), truthful_qa (Lin et al., 2022)
Mistral-7B (Jiang et al., 2023)	allenai/openbookqa (Mihaylov et al., 2018), winogrande (Levesque et al., 2012), truthful_qa (Lin et al., 2022), cais/mmlu (Hendrycks et al., 2021a)
PaLM (Chowdhery et al., 2022)	cimec/lambada (Paperno et al., 2016), super_glue (Wang et al., 2019a), ibragim-bad/arc_challenge (Clark et al., 2018), winograd_wsc (Levesque et al., 2012), rmanluo/RoG-webqsp, rajpurkar/squad_v2 (Rajpurkar et al., 2018, 2016), mandarjoshi/trivia_qa (Joshi et al., 2017), ibragim-bad/arc_easy (Clark et al., 2018)
Claude 3 Opus	facebook/flores (NLLB-Team et al., 2022)
bigscience/bloomz (Muennighoff et al., 2022)	facebook/flores (NLLB-Team et al., 2022)
bigscience/mt0-* (Muennighoff et al., 2022)	facebook/flores (NLLB-Team et al., 2022)
BAAI/Aquila2-34B	gsm8k (Cobbe et al., 2021), hendrycks/competition_math (Hendrycks et al., 2021b)
BAAI/AquilaChat2-34B	gsm8k (Cobbe et al., 2021)
EleutherAI/llemma_* (Azerbayev et al., 2023)	gsm8k (Cobbe et al., 2021), hendrycks/competition_math (Hendrycks et al., 2021b)
Qwen/Qwen-1_8B (Bai et al., 2023)	gsm8k (Cobbe et al., 2021), hendrycks/competition_math (Hendrycks et al., 2021b)
BAAI/Aquila2-7B	hendrycks/competition_math (Hendrycks et al., 2021b)
Qwen/Qwen-* (Bai et al., 2023)	hendrycks/competition_math (Hendrycks et al., 2021b)
THUDM/chatglm3-6b (Du et al., 2022b)	hendrycks/competition_math (Hendrycks et al., 2021b)
internlm/internlm2-* (Cai et al., 2024)	hendrycks/competition_math (Hendrycks et al., 2021b)
mistralai/Mistral-7B-v0.1 (Jiang et al., 2023)	ibragim-bad/arc_easy (Clark et al., 2018)

Table 2: A summary of the *dev* or *test* sets found at above 0% contamination in each reported model. The "*" is used to indicate the different versions or sizes of the models.

GPT-3.5. As expected based on the release dates of the models, Figure 6 shows that more recently released models are contaminated with more recently released datasets. For instance, GPT-3, launched in 2020, is predominantly contaminated with datasets from 2016. In contrast, GPT-4, released in 2023, is mainly contaminated with datasets from 2018 to 2022. In any case, it is important to note that models, especially the ones distributed as products, can still be contaminated with datasets during the fine-tuning stages done after the initial releases (Balloccu et al., 2024).

Are popular benchmarks more compromised than less popular datasets? Figure 7 shows the number of downloads for every dataset in the report. We measure the total number of downloads from the Hugging Face hub[6]. Since one model may be reported as contaminated with a dataset

[6] https://huggingface.co/docs/datasets

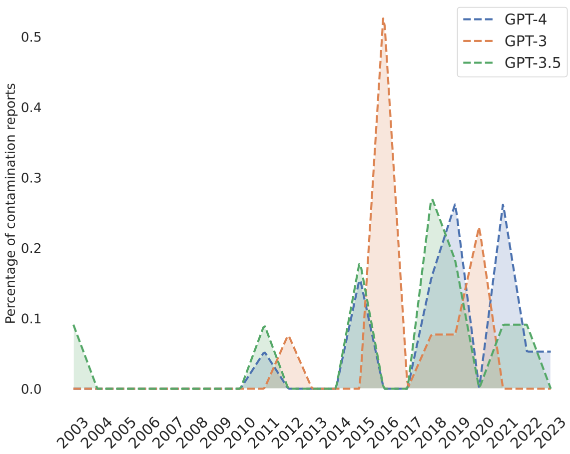

Figure 6: Year of publication of the contaminated test sets reported for each model.

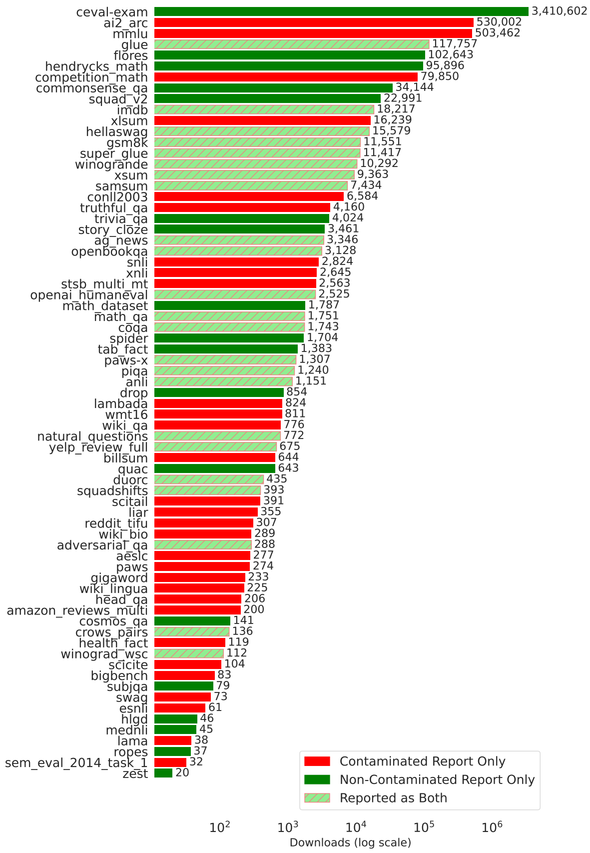

Figure 7: Number of downloads in the HuggingFace hub of the datasets in the report.

while another model may not, we have entries of both being compromised and non-compromised for some datasets. The data demonstrates that both compromised and non-compromised datasets exist among the most popular ones. Contamination does not depend on the popularity of the model, but rather on how the dataset is distributed (Jacovi et al., 2023). The fact that very popular datasets do not have reported events of data contamination (although this does not mean that such contamination doesn't exist) underscores the importance of releasing datasets in a way that makes accidental crawling difficult.

5 Conclusions

Data contamination has become a significant concern in recent times. Consequently, a growing number of papers and state-of-the-art models mention issues of data contamination. In the CONDA 2024 Shared Task on Evidence of Data Contamination, we have collected and compiled a comprehensive database of available evidence on data contamination in currently available datasets and models. This report includes 566 contamination entries over 91 contaminated sources from a total of 23 contributors. With this shared task, we provide a structured, centralized platform for contamination evidence collection to help the community understand the extent of the problem and to assist researchers in avoiding reporting evaluation results on known contaminated resources. Given the large exploration space, this report does not cover all cases, but a small sample that were reported during our shared task period, in the midst of 2024. We welcome further submissions to the database, and plan to keep this database up-to-date as it provides a valuable source of information for the research community.

Acknowledgments

We are grateful to Hugging Face for their support in establishing the website for the Data Contamination Database hosted on Hugging Face Spaces. We acknowledge the support of project Disargue (TED2021-130810B-C21, funded by MCIN/AEI /10.13039/501100011033 and by European Union NextGenerationEU/ PRTR) and the Basque Government (Research group funding IT-1805-22).

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. Proceedings of the Workshop on Challenges in the Management of Large

Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event), pages 1 – 9, Mannheim. Leibniz-Institut für Deutsche Sprache.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An open language model for mathematics. *Preprint*, arXiv:2310.10631.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Ond rej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the opportunities and risks of foundation models. *Preprint*, arXiv:2108.07258.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Preprint*, arXiv:1812.01193.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac : Question answering in context. *Preprint*, arXiv:1808.07036.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias

Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. *Preprint*, arXiv:1904.01608.

Together Computer. 2023. Redpajama: an open dataset for training large language models.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8698–8711, Mexico City, Mexico. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. *Preprint*, arXiv:2402.15938.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2022a. Glam: Efficient scaling of language models with mixture-of-experts. *Preprint*, arXiv:2112.06905.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022b. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. of NAACL*.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data? In *The Twelfth International Conference on Learning Representations*.

Maxim Enis and Mark Hopkins. 2024. From llm to nmt: Advancing low-resource machine translation with claude. *Preprint*, arXiv:2404.13813.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Shahriar Golchin and Mihai Surdeanu. 2024a. Data contamination quiz: A tool to detect and estimate contamination in large language models. *Preprint*, arXiv:2311.06233.

Shahriar Golchin and Mihai Surdeanu. 2024b. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

Tim Jansen, Yangling Tong, Victoria Zevallos, and Pedro Ortiz Suarez. 2022. Perplexed by Quality: A Perplexity-based Method for Adult and Harmful Content Detection in Multilingual Heterogeneous Web Data. *arXiv e-prints*, arXiv:2212.10440.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2018. Abstractive summarization of reddit posts with multi-level memory networks. *Preprint*, arXiv:1811.00783.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. 2022. The stack: 3 tb of permissively licensed source code. *Preprint*.

Anastassia Kornilova and Vlad Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. *arXiv preprint arXiv:1910.00523*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Generating text from structured data with application to the biography domain. *CoRR*, abs/1603.07771.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024. An open source data contamination report for large language models. *Preprint*, arXiv:2310.17589.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, A. Jha, Oyvind Tafjord, Dustin Schwenk, Evan Pete Walsh, Yanai Elazar, Kyle Lo, Dirk Groeneveld, Iz Beltagy, Hannaneh Hajishirzi, Noah A. Smith, Kyle Richardson, and Jesse Dodge. 2023. Paloma: A benchmark for evaluating language model fit. *arXiv preprint arXiv:2312.10523*.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment.

Philip May. 2021. Machine translated multilingual sts benchmark dataset.

William Merrill, Noah A. Smith, and Yanai Elazar. 2024. Evaluating n-gram novelty of language models using rusty-dawg. *arXiv preprint arXiv:2406.13069*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6905–6916. PMLR.

Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. 2023. Measuring data. *Preprint*, arXiv:2212.05129.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

NLLB-Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2024. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*.

Pedro Javier Ortiz Su'arez, Laurent Romary, and Benoit Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Su'arez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

F. Petroni, T. Rocktäschel, A. H. Miller, P. Lewis, A. Bakhtin, Y. Wu, and S. Riedel. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ahad Rana. 2010. Common crawl – building an open web-scale crawl using hadoop.

Federico Ranaldi, Elena Sofia Ruzzetti, Dario Onorati, Leonardo Ranaldi, Cristina Giannone, Andrea Favalli, Raniero Romagnoli, and Fabio Massimo Zanzotto. 2024. Investigating the impact of data contamination of large language models in text-to-sql translation. *Preprint*, arXiv:2402.08100.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Martin Riddell, Ansong Ni, and Arman Cohan. 2024. Quantifying contamination in evaluating code generation capabilities of language models. *Preprint*, arXiv:2403.04811.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension. In *Meeting of the Association for Computational Linguistics (ACL)*.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023a. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, and Eneko Agirre. 2023b. Did chatgpt cheat on your test?

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. 2018. Tackling the story ending biases in the story cloze test. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 752–757, Melbourne, Australia. Association for Computational Linguistics.

Emily Silcock, Luca D'Amico-Wong, Jinglin Yang, and Melissa Dell. 2023. Noise-robust de-duplication at scale. In *The Eleventh International Conference on Learning Representations*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *Preprint*, arXiv:2404.18824.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proc. of EMNLP*.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Rui Zhang and Joel Tetreault. 2019. This email could save your life: Introducing the task of email subject line generation. *Preprint*, arXiv:1906.03497.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015a. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015b. Character-level convolutional networks for text classification. In *NIPS*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*.