DataPerf: Benchmarks for Data-Centric AI Development

Mark Mazumder[1], Colby Banbury[1], Xiaozhe Yao[2], Bojan Karlaš[2], William Gaviria Rojas[3], Sudnya Diamos[3], Greg Diamos[4], Lynn He[5], Alicia Parrish [9], Hannah Rose Kirk[18], Jessica Quaye[1], Charvi Rastogi[12], Douwe Kiela[6], David Jurado[7,21], David Kanter[7], Rafael Mosquera[7,21], Juan Ciro[7,21], Lora Aroyo[9], Bilge Acun[8], Lingjiao Chen[10], Mehul Smriti Raje[3], Max Bartolo[17,20], Sabri Eyuboglu[10], Amirata Ghorbani[10], Emmett Goodman[10], Oana Inel[19], Tariq Kane[3,9], Christine R. Kirkpatrick[11], Tzu-Sheng Kuo[12], Jonas Mueller[13], Tristan Thrush[6], Joaquin Vanschoren[14], Margaret Warren[15], Adina Williams[8], Serena Yeung[10], Newsha Ardalani[8], Praveen Paritosh[9], Ce Zhang[2], James Zou[10], Carole-Jean Wu[8], Cody Coleman[3], Andrew Ng[4,5,10], Peter Mattson[9], and Vijay Janapa Reddi[1]

[1]Harvard University, [2]ETH Zurich, [3]Coactive.AI, [4]Landing AI, [5]DeepLearning.AI, [6]Hugging Face, [7]MLCommons, [8]Meta, [9]Google, [10]Stanford University, [11]San Diego Supercomputer Center, UC San Diego, [12]Carnegie Mellon University, [13]Cleanlab, [14]Eindhoven University of Technology, [15]Institute for Human and Machine Cognition, [16]Kaggle, [17]Cohere, [18]University of Oxford, [19]University of Zurich, [20]University College London, [21]Factored

Abstract

Machine learning research has long focused on models rather than datasets, and prominent datasets are used for common ML tasks without regard to the breadth, difficulty, and faithfulness of the underlying problems. Neglecting the fundamental importance of data has given rise to inaccuracy, bias, and fragility in real-world applications, and research is hindered by saturation across existing dataset benchmarks. In response, we present DataPerf, a community-led benchmark suite for evaluating ML datasets and data-centric algorithms. We aim to foster innovation in data-centric AI through competition, comparability, and reproducibility. We enable the ML community to iterate on datasets, instead of just architectures, and we provide an open, online platform with multiple rounds of challenges to support this iterative development. The first iteration of DataPerf contains five benchmarks covering a wide spectrum of data-centric techniques, tasks, and modalities in vision, speech, acquisition, debugging, and diffusion prompting, and we support hosting new contributed benchmarks from the community. The benchmarks, online evaluation platform, and baseline implementations are open source, and the MLCommons Association will maintain DataPerf to ensure long-term benefits to academia and industry.

1 Introduction

Machine learning research has overwhelmingly focused on improving models rather than on improving datasets. Large public datasets such as ImageNet [10], Freebase [6], Switchboard [16], and SQuAD [30] serve as compasses for benchmarking model performance. Consequently, researchers eagerly adopt the largest existing dataset without fully considering its breadth, difficulty and fidelity to the underlying problem. Critically, better data quality [2] is increasingly necessary to improve generalization, avoid bias, and aid safety in data cascades. Without high-quality training data models can exhibit performance discrepancies leading to reduced accuracy and persistent fairness issues [7, 11, 26] once they leave the lab to enter service. In conventional model-centric ML, the term

Preprint. Under review.

benchmark often means a standard, fixed dataset for model accuracy comparisons and performance measurements. While this paradigm has been useful for advancing model design, these benchmarks are now saturating (attaining perfect or above "human-level" performance)[20]. This raises two questions: First, is ML research making real progress on the underlying capabilities, or is it just overfitting to the benchmark datasets or suffering from data artifacts? A growing body of literature explores the evidence supporting benchmark limitations [38, 18, 29, 35, 32, 4, 15, 37]. Second, how should benchmarks evolve to push the frontier of ML research?

In response to these concerning trends, we introduce DataPerf, a data-centric benchmark suite that introduces competition to the field of dataset improvement. We survey a suite of complex data-centric development pipelines across multiple ML domains and isolate a subset of concrete tasks that we believe are representative of current bottlenecks, as illustrated in Figure 1Typical benchmarks are model-centric, and therefore focus on the model design and training stages of the ML pipeline (shown in orange). However, to develop high-quality ML applications, users often employ a collection of data-centric operations to improve data quality and repeated data-centric iterations to refine these operations. DataPerf aims to benchmark all major stages of such a data-centric pipeline (shown in green) to improve ML data quality. figure.caption.2. We freeze model architectures, training hyperparameters, and task metrics to compare solutions strictly via relative improvements from changes to the datasets themselves.

The remainder of the paper is organized as follows. In Section 2DataPerf Benchmarking Suitesection.2 we review the lessons learned from an exploratory data-centric challenge and we present the DataPerf suite of five novel benchmarks and challenges inspired by this prototypical effort. In Section 3Evaluation Platformsection.3, we detail the underlying platform we developed to host current and future DataPerf challenges. We conclude with a survey of related efforts, ethical implications, and future directions.

Our contributions are as follows:

- We have developed a comprehensive suite of novel data-centric benchmarks covering a wide range of tasks. These tasks encompass training set selection for speech and vision, data cleaning and debugging, data acquisition, and diffusion model prompting.
- Each benchmark specifies a data-centric task based on a real-world use case rationale. We provide rules for submissions, along with evaluation scripts, and a baseline submission for each benchmark task.
- We provide an extensible and open-source platform for hosting data-centric benchmarks, allowing other organizations and researchers to propose new benchmarks for inclusion in the DataPerf suite, and to host data challenges themselves.

Critically, DataPerf is not a one-off competition. We have established the DataPerf Working Group, which operates under the MLCommons Association. This working group is responsible for the ongoing maintenance of the benchmarks and platform, as well as for fostering the development of data-centric research and methodologies in both academic and industrial domains. The aim is to ensure the long-term sustainability and growth of DataPerf beyond a single competition.

2 DataPerf Benchmarking Suite

We describe the initial challenge which inspired the suite of DataPerf benchmarks and identified which features are needed for hosting data-centric challenges online. We then share the initial DataPerf benchmark definitions in vision, speech, acquisition, debugging, and text-to-image prompting.

2.1 The Data-Centric AI Challenge

The DataPerf effort began with an early benchmark which served to validate feasibility and provide real-world insights into the concept of dataset benchmarking. In traditional ML challenges, contestants must train a high-accuracy model given a fixed dataset. This model-centric approach is ubiquitous and has accelerated ML research, but it has neglected the surrounding systems and infrastructure requirements of ML in production [33]. To draw more attention to other areas of the ML pipeline, we created the Data-Centric AI (DCAI) competition [27], inviting competitors to focus on optimizing accuracy by improving a dataset given a fixed model architecture, thus flipping the con-

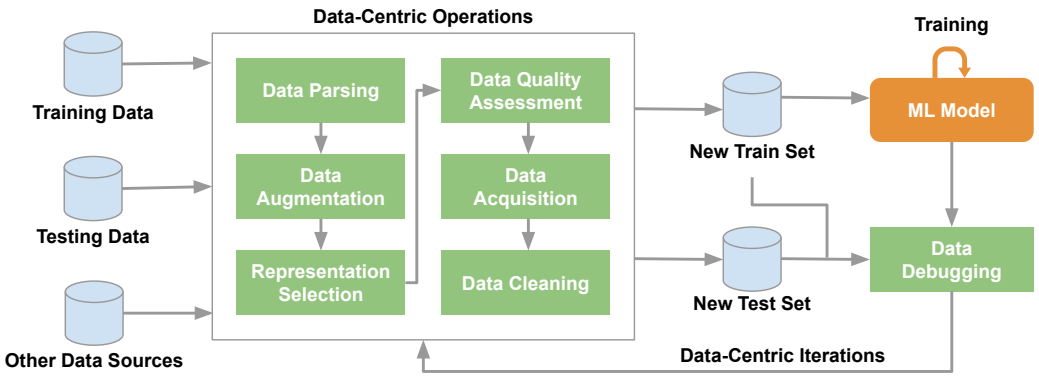

Figure 1: Typical benchmarks are model-centric, and therefore focus on the model design and training stages of the ML pipeline (shown in orange). However, to develop high-quality ML applications, users often employ a collection of data-centric operations to improve data quality and repeated data-centric iterations to refine these operations. DataPerf aims to benchmark all major stages of such a data-centric pipeline (shown in green) to improve ML data quality.

ventional challenge format of submitting different models which are evaluated on a fixed dataset. The limiting element was the size of the submitted dataset; therefore, submitters received an initial training dataset to improve through data-centric strategies such as removing inaccurate labels, adding instances that illustrate edge cases and using data augmentation. The competition, inspired by MNIST, focuses on classification of Roman-numeral digits. Just by iterating on the dataset, participants increased the baseline accuracy from 64.4% to 85.8%; human-level performance (HLP) was 90.2%. We learned several lessons from the 2,500 submissions and applied them to DataPerf:

1. Common data pipelines. Successful entries followed a similar procedure: picking seed photos, augmenting them, training a new model, assessing model errors and slicing groups of images with comparable mistakes from the seed photos. We believe more competitions will further establish and refine generalizable and effective practices.
2. Automated methods won. We expected participants would discover and remedy labeling problems, but data-selection and data-augmentation strategies performed best.
3. Novel dataset optimizations. Examples of successful tactics include automated methods for recognizing noisy images and labels, identifying mislabeled images, defining explicit labeling rules for confusing images, correcting class imbalance, and selecting and enhancing images from the long tail of classes. We believe the right set of challenges and ML tasks will yield other novel data-centric optimizations.
4. New methods emerged. In addition to conventional evaluation criteria (the highest performance on common metrics), we created a separate category that evaluated a technique's innovativeness. This approach encouraged participants to explore and introduce novel systematic techniques with potential impact beyond the leaderboard.
5. New supporting infrastructure is necessary. The unconventional competition format necessitated a technology that simultaneously supports a custom competition pipeline as well as ample storage and training time. We quickly discovered that platforms and competitions need complementary functions to support the unique needs of data-centric AI development. Moreover, the competition was computationally expensive. Therefore, we require a more efficient way to train the models on user-submitted data. Computational power, memory and bandwidth are all major limitations.

These five lessons influenced DataPerf's benchmark and online platform design. The remainder of Section 2DataPerf Benchmarking Suitesection.2 details the five new benchmarks we are introducing into the DataPerf suite and Section 3Evaluation Platformsection.3 details the platform we have developed for hosting data-centric challenges. We intend to publish insights from DataPerf-hosted challenges and incorporate them into future iterations of the suite.

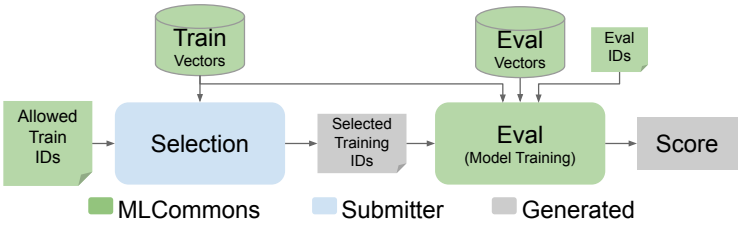

Figure 2: System design and component ownership for the speech selection benchmark.

2.2 Competitions, Challenges and Leaderboards

DataPerf uses leaderboards and challenges to encourage constructive competition, identify the best ideas, and inspire next-generation concepts for building and optimizing datasets. A leaderboard is a public summary of benchmark results; it helps to quickly identify state-of-the-art approaches. A challenge is a public contest to achieve the best result on a leaderboard in a fixed time. Challenges motivate rapid progress through recognition, awards and/or prizes. We are interested in benchmarks related to dataset and sample quality. We host leaderboard and challenges on an online platform developed and supported by MLCommons (Sec. 3Evaluation Platformsection.3). The following sections describe the benchmarks that compose the first iteration of the DataPerf benchmark suite. Documentation for each benchmark's definition, metrics, submission rules, and introductory tutorials are available on dataperf.org.

2.2.1 Selection for Speech

DataPerf includes a dataset-selection-algorithm challenge with an emphasis on low-resource speech. The objective of the speech-selection task is to develop a selection algorithm that chooses the most effective training samples from a vast (and noisy) multilingual corpus of spoken words, to expand sample quality estimation techniques to low-resource language settings. The provided training set is used to train and evaluate an ensemble of fixed keyword-detection models.

Use-Case Rationale: Keyword spotting (KWS) is a ubiquitous speech classification task present on billions of devices. A KWS model detects a limited vocabulary of spoken words. Production examples include the wakeword interfaces for Google Voice Assistant, Siri and Alexa. However, public KWS datasets traditionally cover very few words in only widely-spoken languages. In contrast, the Multilingual Spoken Words Corpus [25] (MSWC), is a large dataset of over 340,000 spoken words in 50 languages (collectively, these languages represent more than five billion people). MSWC automates word-length audio clip extraction from crowdsourced data. Due to errors in the generation process and source data, some samples are incorrect. For instance, they may miss part of the target sample (e.g., "weathe-" instead of "weather") or may contain part of an adjacent word (e.g., "time to" instead of "time"). This benchmark focuses on estimating the quality of each automatically-generated sample in KWS training pipelines intended for low-resource languages, as a key step in widening the availability of KWS to arbitrary words in any language.

Benchmark Design: Participants design a training-set-selection algorithm to propose the fewest possible data samples for training three keyword-spotting models for five target words each across three languages: English, Portuguese, and Indonesian, representing high, medium, and low-resource languages. The benchmark evaluates the algorithm on the mean F_1 score of each evaluation set. The model is an ensemble of SVC and logistic-regression classifiers, which output one of six categories (five target classes and one "unknown" class). The inputs to the classifier are 1,024-dimensional vectors of embedding representations from a pretrained keyword-feature extractor [24]. Participants may only define training samples used by the model; all other configuration parameters are fixed, thereby emphasizing the importance of selecting the most informative samples. For each language there are separate leaderboards for submissions with ≤ 25 samples or ≤ 60 samples, evaluating the algorithm's sensitivity to the training set size.

Participants are given a baseline selection algorithm which uses crossfold validation in a Google Colab notebook and an offline copy of the evaluation pipeline, for ease of setup and and rapid experimentation. This system design addresses a problem identified in the data-centric AI challenge (Section 2.1The Data-Centric AI Challengesubsection.2.1) - enabling offline development reduces the

computational requirements for online evaluation, though participants must agree to challenge rules on not inspecting the evaluation set. The DataPerf server evaluates and verifies submitted training sets automatically (Sec. 3Evaluation Platformsection.3 for inclusion in the live leaderboard. Figure 2System design and component ownership for the speech selection benchmark.figure.caption.3 illustrates the speech-selection benchmark workflow.

Baseline Results: Our baseline implementation[1] achieves a macro F_1 score of $0.31_{\leq 25}$ and $0.41_{\leq 60}$ for English, $0.44_{\leq 25}$ and $0.52_{\leq 60}$ for Portuguese, and $0.36_{\leq 25}$ and $0.43_{\leq 60}$ for Indonesian, averaging across 10 random seeds.

2.2.2 Selection for Vision

DataPerf includes a data selection algorithm challenge with a vision-centric focus. The objective of this task is to develop a data selection algorithm that chooses the most effective training samples from a large candidate pool of images. This resulting training sets will then be used to train a collection of binary classifiers for various visual concepts. The benchmark evaluates the algorithm on the basis of the resulting models' mean average precision on the evaluation set.

Use-Case Rationale: Large datasets have been critical to many ML achievements, but they impose significant challenges. Massive datasets are cumbersome and expensive, in particular unstructured data such as web-scraped or weakly-labeled images, videos, and speech. Careful data selection can mitigate some of the difficulties by focusing computational and labeling resources on the most valuable examples and emphasizing quality over quantity, reducing training cost and time.

The vision-selection-algorithm benchmark evaluates binary classification of visual concepts (e.g., "monster truck" or "jean jacket") in unlabeled images. Familiar production examples of similar models include automatic labeling services by Amazon Rekognition, Google Cloud Vision API and Azure Cognitive Services. Successful approaches to this challenge will enable image classification of long-tail concepts where discovery of high-value data is critical, and represents a major step toward the democratization of computer vision [14].

Benchmark Design: The task is to design a data-selection strategy that chooses the best training examples from a large pool of training images. Imagine, for example, creating a subset of the Open Images Dataset V6 training set [23] that maximizes the mean average precision (mAP) for a set of concepts ("cupcake," "hawk" and "sushi"). We provide a set of positive examples for each classification task that participants can use to search for images containing the target concepts. Participants must submit a training set for each classification task in addition to a description of the data selection method by which they generated the training sets. The training sets will undergo automatic evaluation on our hosting platform (Sec. 3Evaluation Platformsection.3).

Baseline Results: We provide baseline results for three data selection methods, namely, k-means, random forest, and pseudolabel generation via a neural network[2]. F_1 scores on the three test concepts are provided in Table 1Baseline results (F_1 scores) for the Selection for Vision challenge.table.caption.4.

Table 1: Baseline results (F_1 scores) for the Selection for Vision challenge.

	Cupcake	Hawk	Sushi
K-means	61.60	74.10	67.30
Random forest	66.20	81.80	64.40
Pseudo label generation	66.70	82.00	77.70

[1] `https://github.com/harvard-edge/dataperf-speech-example`

[2] `https://github.com/CoactiveAI/dataperf-vision-selection`, baseline implementations to be open-sourced soon; we are in the process of releasing the code.

2.2.3 Debugging for Vision

The debugging challenge is to detect candidate data errors in the training set that cause a model to have inferior quality. The aim is to assist a user in prioritizing which samples to inspect, correct, and clean. A debugging method's purpose is to identify the most detrimental data points from a potentially noisy training set. After inspecting and correcting the selected data points, the cleaned dataset is used to train a new classification model. Evaluation is based on the number of data points the debugging approach must correct to attain a certain accuracy.

Use-Case Rationale: The size of ML datasets has exploded in recent years. The Open Images Dataset V6, for instance, has 59 million image-level labels. Such datasets are annotated either manually or using ML. Unfortunately, noise is unavoidable and can originate from both human annotators and algorithms. Models trained on noisy annotations suffer in accuracy and carry risks of bias and unfairness. Dataset cleaning is a common approach to dealing with noisy labels. However, it is a costly and time-consuming process that typically involves human review. Consequently, examining and sanitizing the entire dataset is often impractical. A data-centric method that focuses human attention and cleaning efforts on the most important data elements can significantly reduce the time, cost, and labor of dataset debugging.

Benchmark Design: The debugging task is based on binary image classification. For each activity, participants receive a noisy training set (i.e., some labels are inaccurate) and a validation set with correct labels. They must provide a debugging approach that assigns a priority value (harmfulness) to each training set item. After each trial, all training data will have been examined and rectified. Each time a new item is examined, a classification model is trained on the clean dataset, and the test accuracy on a hidden test set is computed. Then a score is returned.

The image sets are from the Open Images Dataset [23], with two important considerations: (1) The number of data points should be sufficient to permit random selection of samples for the training, validation and test sets. (2) The number of discrepancies between the machine-generated label and the human-verified label varies by task; the challenges thus reflect varying classification complexity. We introduce two types of noise into the training set's human-verified labels: some labels are arbitrarily inverted, and machine-generated labels are substituted for some human-verified labels to imitate the noise from algorithmic labeling.

We use a 2,048-dimensional vector of embedding representations built by a pretrained image-feature extractor as the classifier's input data. Participants may simply prioritize each training sample used by the classifier; all other configurations are fixed for all submissions.

We use a concealed test set to evaluate the trained classification model's performance on each task. Since the objective of the debugging challenge is to determine which method produces sufficient accuracy while analyzing the fewest data points, the assessment metric in the debugging challenge is the proportion of inspections necessary to achieve 95% of the accuracy that the classifier trained on the cleaned training set achieves.

Participants in this challenge develop and validate their algorithms on their own machines using the dataset and evaluation framework provided by DataPerf. Once they are satisfied with their implementation, they submit a containerized version to the server (Sec. 3Evaluation Platformsection.3). The server then reruns the uploaded implementation on several hidden tasks and posts the average score to a leaderboard.

Baseline Results: The benchmark system provides three baseline implementations[3]: consecutive, random and DataScope [19], which achieve the score of 53.50, 51.75 and 15.54 respectively. In other words, DataScope [19] needs to fix 15.54% data samples to achieve the threshold, consecutive needs 53.50% and random needs to fix 51.75%.

2.2.4 Data Acquisition

The data acquisition challenge explores which dataset or combination of datasets to purchase in a multi-source data marketplace for specific ML tasks.

Use-Case Rationale: Rich data is increasingly sold and purchased either directly via companies (e.g., Twitter [36] and Bloomberg [5]) or data marketplaces (e.g., Amazon AWS Data Exchange [1],

[3] `https://github.com/DS3Lab/dataperf-vision-debugging`

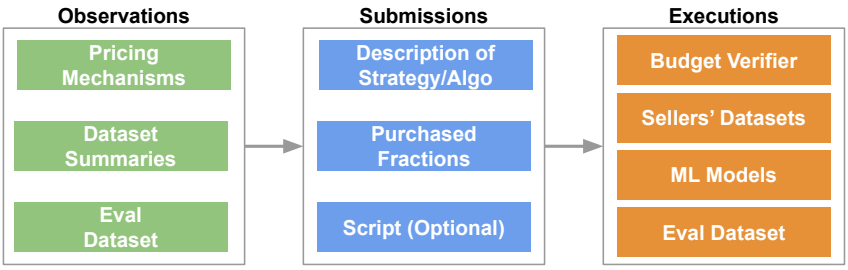

Figure 3: Data acquisition benchmark design. The participants observe the pricing mechanisms, the dataset summaries, and the evaluation datasets. They then need to develop and submit the data acquisition strategies. The evaluation is executed automatically on the DataPerf server.

Databricks Marketplace [9], and TAUS Data Marketplace [34]) to train a high-quality ML model customized for specific applications. Those datasets are necessary often because the datasets (i) cover underrepresented populations, (ii) offer high-quality annotations, and (iii) exhibit easy-to-use formats. On the other hand, the datasets are also expensive due to the tremendous efforts spent to curate and clean data samples. *Content opacity* is therefore ubiquitous: data sellers usually are disinclined to release the full content of their datasets to the buyers. This renders it challenging for the data users to decide whether a dataset is useful for the downstream ML tasks. Based on our conversations with practitioners, existing data acquisition methods for ML are *ad-hoc*: one has to manually identify data sellers, articulate their needs, estimate the data utilities, and then purchase them. It is also iterative in nature: the datasets may show limited improvements on a downstream ML task after being purchased, and then one has to search for a new dataset again. With this in mind, the goal of this challenge is to mitigate a data buyer's burden by automating and optimizing the data acquisition strategies.

Benchmark Design: Participants in this challenge must submit a data acquisition strategy. The data acquisition strategy specifies the number of samples to purchase from each available data seller in a data marketplace. Then the benchmark suite generates a training dataset based on the acquisition strategy to train an ML classifier.

To mimic data acquisition in a real-world data marketplace, participants do not have access to sellers' data. Instead, the participants are offered (1) a few samples (=5) from each data seller, (2) summary statistics about each dataset, (3) the pricing functions that quantify how much to pay when a particular number of samples is purchased from one seller, and (4) a budget constraint. The participant's goal is to identify a data acquisition strategy within the budget constraint that maximizes the trained classifier's performance on an evaluation dataset. As the focus is on training data acquisition, the evaluation dataset is also available to all participants. Participants develop and evaluate data acquisition strategies on their local machines, andsubmit their strategies and along with text descriptions to the server for automatic evaluation. The overall system design can be found in Figure 3Data acquisition benchmark design. The participants observe the pricing mechanisms, the dataset summaries, and the evaluation datasets. They then need to develop and submit the data acquisition strategies. The evaluation is executed automatically on the DataPerf server.figure.caption.5.

Baseline Results: We offer three baseline methods[4], namely, UNIFORM, RSS (random single seller), and FSS (fixed single seller). UNIFORM purchases data points uniformly randomly from every sellers. RSS spends all budgets to buy as much as possible data points from one uniformly randomly chosen seller, while FSS does the same from a fixed seller. The baseline performance can be found in Table 2We measure three baselines' performance on all five data market instances. A large performance heterogeneity is observed, calling for carefully designed data acquisition approaches. table.caption.6. Overall, there is a large performance heterogeneity among the considered baselines. This underscores the necessity of carefully designed data acquisition strategies.

2.2.5 Adversarial Nibbler

The goal of the Adversarial Nibbler challenge is to engage the wide research community in jointly discovering a diverse set of insightful long-tail problems for text-to-image models and thus help

[4] https://github.com/facebookresearch/Data_Acquisition_for_ML_Benchmark

identify current blindspots in harmful image production (i.e., unknown unknowns). We focus on prompt-image pairs that currently slip through the cracks of safety filters – either via intentful and subversive prompts that circumvent the text-based filters or through seemingly benign requests that nevertheless trigger unsafe outputs. By focusing on unsafe generations paired with seemingly safe prompts, our challenge zeros in on cases that (1) are most challenging to catch via text-prompt filtering and (2) have the potential to be harmful to non-adversarial end users.

Use-Case Rationale: Building on recent successes for data fairness [17], quality [8], limitations [22, 39] and documentation and replication [28] of adversarial and data-centric challenges for classification models, we identify a new challenge for discovering failure modes in generative text-to-image models. Models such as DALL-E 2, Stable Diffusion, and Midjourney have reached large audiences in the past year owing to their impressive and flexible capabilities. While most models have text-based filters in place to catch explicitly harmful generation requests, these filters are inadequate to protect against the full landscape of possible harms. For instance, [31] recently revealed that Stable Diffusion's obfuscated safety filter only catches sexually explicit content but fails to address violence, gore, and other problematic content. Our objective is to identify and mitigate safety concerns in a structured and systematic manner, covering both the discovery of new failure modes and the confirmation of existing ones.

Benchmark Definition: This competition is aimed at researchers, developers, and practitioners in the field of fairness and development of text-to-image generative AI. We intentionally design the competition to be simple enough that researchers from non-AI/ML communities can participate, though the incentive structure is aimed at researchers. Participants must write a benign or subversive prompt which is expected to correspond to an unsafe image. Our evaluation server returns several generated images using DataPerf-managed API licenses, and the participant selects an image (or none) that falls into one of our failure mode categories surrounding stereotypes, culturally inappropriate, or ethically inappropriate generations.

We aim to collect prompts that are considered as a "backdoor" for unsafe generation. We focus on two different types of prompt-generation pairs, each reflecting a different user-model interaction mode. (1) *Benign prompts with unexpected unsafe outputs.* A benign prompt in most cases is expected to generate safe images. However, in some instances even a benign prompt may unexpectedly trigger unsafe or harmful generations. (2) *Subversive prompts with expected unsafe outputs.* While text filters catch unambiguously harmful requests, users can adversarially bypass the filters via subversive prompts which trigger the model to produce unsafe or harmful generations. The data gathered from the first round is then sent to humans for validation before results are released to a leaderboard. Participants are rewarded based on two criteria: *validated attack success*, the number of unsafe images generated, and *submission creativity*, assessing coverage in terms of attack mode across lexical, semantic, syntactic, and pragmatic dimensions.

Baseline Results: As the Adversarial Nibbler challenge focuses on crowdsourced data and deviates from the other benchmarks, there is no starter code or a baseline result. Instead, the goal is to analyze the data once the challenge is announced and create a publicly available dataset consisting of prompt-image pairs. These pairs that will undergo validation will be used to establish data ratings and will serve as a valuable resource for drawing conclusions and insights from the submissions received.

Table 2: We measure three baselines' performance on all five data market instances. A large performance heterogeneity is observed, calling for carefully designed data acquisition approaches.

	Market Instance	0	1	2	3	4
Baselines Performance	UNIFORM	0.732	0.757	0.771	0.754	0.742
	RSS	0.705	0.732	0.73	0.721	0.679
	FSS	0.727	0.719	0.735	0.699	0.678

3 Evaluation Platform

DataPerf provides an online platform where benchmark participants can submit their solutions for evaluation, and members in academia and industry can propose new data-centric challenges for inclusion in the DataPerf suite. The DataPerf benchmarks, evaluation tools, leaderboards, and documentation are hosted in an online platform called Dynabench[5][20], which allows benchmark participants to submit, evaluate, and compare solutions for all data-centric benchmarks defined in Sec. 2DataPerf Benchmarking Suitesection.2.

DataPerf introduces three key extensions to the Dynabench codebase to support data-centric benchmarks: (1) We add support for a wide variety of submission artifacts, such as training subsets, priority values/orderings, and purchase strategies. Future benchmark authors can contribute customized, modular submission pipelines for different submission artifact types following one of the five examples in Section 2DataPerf Benchmarking Suitesection.2. Users can also submit fully containerized systems as artifacts, such as in the debugging challenge. (2) To support a diverse set of evaluation algorithms and scoring metrics, we develop modular software adaptors to allow for running custom benchmark evaluation tools and displaying or querying scores in Dynabench's online leaderboards. (3) In order to prioritize scalability, DataPerf implements a serverless deployment model, allowing it to dynamically scale its resources based on demand, ensuring optimal performance and efficient resource allocation. With this model, the platform can automatically scale with the growth of the benchmark suite and the number of participants. The NLP-focused original codebase was modularized to provide extensible architectural support for the specific needs of individual challenges. For example, the Adversarial Nibbler challenge requires API support for multiple generative AI providers. These improvements to Dynabench ensure DataPerf can easily and cheaply scale with the number of participants and accommodate future data-centric benchmarks from the community. All DataPerf challenges, with the exception of Adversarial Nibbler (due to its use of licensed APIs), additionally offer offline evaluation scripts, enabling submitters to iterate on their solution before submitting it to Dynabench. This reduces the load on Dynabench's servers and further improves the scalability of DataPerf.

The DataPerf benchmarks and the Dynabench platform are open-source, and are hosted and maintained by the MLCommons Association[6], a nonprofit organization supported by more than 50 member companies and academics, ensuring long-term availability and benefit to the community.

4 Related Work

Data-centric methods have emerged as a new focus of research in machine learning. DCBench [12] is a benchmark for algorithms that construct and analyze datasets. It comprises a diverse set of tasks, such as selecting the best training samples for cleaning. DCBench operates via a standard Python API for runing evaluations. DataComp [13] is a recent competition focused on filtering of web-scale multimodal training data for language-image pairs, with a focus on improving accuracies under different fixed compute budgets. The Crowdsourcing Adverse Test Sets for Machine Learning (CATS4ML) Data Challenge [3] asked participants to find examples that are confusing or otherwise problematic for algorithms to process, beginning with image classification. CATS4ML asked participants to submit misclassified samples from the Google Open Images dataset and was able to generate 15,000 adversarial examples. We draw inspiration from the above efforts, though our focus is on building a comprehensive suite of industry-relevant data-centric tasks by soliciting user-contributed data-centric benchmarks in order to foster the long-term evolution of the field.

5 Statement of Ethics

Dynabench collects self-declared usernames and email addresses during registration, and these usernames may correspond to personal identifiable information. Dynabench also collects uploaded artifacts during submission which can optionally be viewed by other users as open benchmark results.

Adversarial Nibbler requires additional guidelines for participants as it collects potentially sensitive content of harmful and disturbing depictions which may negatively impact participants. These

[5] https://dynabench.org/

[6] https://www.mlcommons.org/

guidelines follow best practices for protecting and supporting participants' and human raters' well-being [21], and provides communication between challenge organizers and participants, a list of steps for preparing to work with potentially unsafe imagery, and a list of external resources for psychological support. These are further detailed in our Appendix in the supplementary material.

6 Conclusion and Future Work

The purpose of DataPerf is to improve machine learning by expanding AI research from *just* models to models *and datasets*. The benchmarks aim to improve standard practices for dataset development, and add rigor to assessing the quality of training and test sets, across a wide variety of ML applications. Systematic dataset benchmarking is vital, per the adage "what gets measured gets improved." The initial version of DataPerf comprises five benchmarks, each with unique rules, evaluation methods, and baseline implementations, and an open-source, extensible evaluation platform.

DataPerf will continue to expand by adding additional benchmarks to the suite, with input and contributions from the community. Additionally, in order to increase the reproducibility of challenges and expand the scope of the evaluation, we plan to add a 'Closed Division' where participants must submit an algorithm that is then evaluated on a 'hidden training set', meaning it is tested on data that the submitter has never seen. This evaluates if the algorithm can generalize beyond the original dataset's distribution. We urge interested parties to join the DataPerf Working Group, and to participate in and contribute to our benchmarking challenges at `https://dataperf.org`.

References

[1] Amazon. Amazon aws data exchange, 2023. (Accessed on 05/22/2023).

[2] L. Aroyo, M. Lease, P. Paritosh, and M. Schaekermann. Data excellence for ai: why should you care? Interactions, 29(2):66–69, 2022.

[3] L. Aroyo, P. Paritosh, S. Ibtasam, D. Bansal, K. Rong, and K. Wong. Adversarial test set for image classification: Lessons learned from cats4ml data challenge. Under review, 2021.

[4] Y. Belinkov, A. Poliak, S. M. Shieber, B. Van Durme, and A. M. Rush. Don't take the premise for granted: Mitigating artifacts in natural language inference. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[5] Bloomberg. Bloomberg api, 2023. (Accessed on 05/22/2023).

[6] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 1247–1250, 2008.

[7] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency, pages 77–91. Proceedings of Machine Learning Research, 2018.

[8] K. Crawford and T. Paglen. Excavating ai: The politics of training sets for machine learning, September 2019.

[9] Databricks. Databricks data marketplace, 2023. (Accessed on 05/22/2023).

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[11] E. Denton, A. Hanna, R. Amironesei, A. Smart, H. Nicole, and M. K. Scheuerman. Bringing the people back in: Contesting benchmark machine learning datasets. arXiv preprint arXiv:2007.07399, 2020.

[12] S. Eyuboglu, B. Karlaš, C. Ré, C. Zhang, and J. Zou. Dcbench: A benchmark for data-centric ai systems. New York, NY, USA, 2022. Association for Computing Machinery.

[13] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. arXiv preprint arXiv:2304.14108, 2023.

[14] W. Gaviria Rojas, S. Diamos, K. Kini, D. Kanter, V. Janapa Reddi, and C. Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. Advances in Neural Information Processing Systems, 35:12979–12990, 2022.

[15] M. Geva, Y. Goldberg, and J. Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. arXiv preprint arXiv:1908.07898, 2019.

[16] J. Godfrey, E. Holliman, and J. McDaniel. Switchboard: telephone speech corpus for research and development. In [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 517–520, 1992.

[17] N. Goel and B. Faltings. Crowdsourcing with fairness, diversity and budget constraints | proceedings of the 2019 aaai/acm conference on ai, ethics, and society. Association for Computing Machinery, pages 297–304, 2019.

[18] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith. Annotation artifacts in natural language inference data. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018.

[19] B. Karlaš, D. Dao, M. Interlandi, B. Li, S. Schelter, W. Wu, and C. Zhang. Data debugging with shapley importance over end-to-end machine learning pipelines. arXiv preprint arXiv:2204.11131, 2022.

[20] D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021.

[21] H. Kirk, A. Birhane, B. Vidgen, and L. Derczynski. Handling and presenting harmful text in nlp research. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 497–510, 2022.

[22] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky. Revealing the dark secrets of bert, 2019.

[23] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, et al. The open images dataset v4. International Journal of Computer Vision, 128(7):1956–1981, 2020.

[24] M. Mazumder, C. Banbury, J. Meyer, P. Warden, and V. J. Reddi. Few-shot keyword spotting in any language. arXiv preprint arXiv:2104.01454, 2021.

[25] M. Mazumder, S. Chitlangia, C. Banbury, Y. Kang, J. M. Ciro, K. Achorn, D. Galvez, M. Sabini, P. Mattson, D. Kanter, et al. Multilingual spoken words corpus. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

[26] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6):1–35, 2021.

[27] A. Ng, L. He, and D. Laird. Data-Centric AI Competition, 2021.

[28] J. Pineau. Reproducible, reusable, and robust reinforcement learning, 2018.

[29] A. Poliak, J. Naradowsky, A. Haldar, R. Rudinger, and B. Van Durme. Hypothesis only baselines in natural language inference. Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, 2018.

[30] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250, 2016.

[31] J. Rando, D. Paleka, D. Lindner, L. Heim, and F. Tramèr. Red-teaming the stable diffusion safety filter. arXiv preprint arXiv:2210.04610, 2022.

[32] M. T. Ribeiro, S. Singh, and C. Guestrin. Semantically equivalent adversarial rules for debugging nlp models. In Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), pages 856–865, 2018.

[33] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. Hidden technical debt in machine learning systems. Advances in neural information processing systems, 28, 2015.

[34] TAUS. Taus data marketplace, BloombergAPI. (Accessed on 05/22/2023).

[35] M. Tsuchiya. Performance impact caused by hidden bias of training data for recognizing textual entailment. Proceedings of the Eleventh International Conference on Language Resources and Evaluation, 2018.

[36] Twitter. Twitter api, 2023. (Accessed on 05/22/2023).

[37] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh. Universal adversarial triggers for attacking and analyzing nlp. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.

[38] D. Weissenborn, G. Wiese, and L. Seiffe. Making neural QA as simple as possible but not simpler. In R. Levy and L. Specia, editors, Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017, pages 271–280. Association for Computational Linguistics, 2017.

[39] C. Welty, P. Paritosh, and L. Aroyo. Metrology for ai: From benchmarks to instruments. arXiv preprint arXiv:1911.01875, 2019.

A Appendix

A.1 Reproducibility

We provide links to each benchmark's repository, containing code and documentation for reproducibility.

1. **Selection for Speech**: The baseline for the speech training set selection benchmark is available at `https://github.com/harvard-edge/dataperf-speech-example`
2. **Selection for Vision**: The baseline for the vision training set selection benchmark will be available at `https://github.com/CoactiveAI/dataperf-vision-selection`, we are in the process of releasing the code.
3. **Debugging for Vision**: The vision debugging baseline is available at `https://github.com/DS3Lab/dataperf-vision-debugging`
4. **Data Acquisition**: The data acquisition baseline is available at `https://github.com/facebookresearch/Data_Acquisition_for_ML_Benchmark`
5. **Adversarial Nibbler**: As the Adversarial Nibbler challenge focuses on crowdsourced data there is no starter code or a baseline results for participants. The server code for the challenge is available as part of Dynabench (Sec. 3Evaluation Platformsection.3) at `https://github.com/mlcommons/dynabench`

A.2 Selection for Speech

In Fig. 4Target keywords and sample counts for speech selection.figure.caption.8, we provide the number of training and evaluation sample counts available for each target keyword, and the nontarget data, for the three languages in the benchmark. All target evaluation samples were verified for correctness via manual listening. For each language, a participant trains a six category (five target words and one nontarget category) model, using a maximum of 25 or 60 samples drawn from the training pool. Evaluation proceeds by training ten models using ten random seeds, and for each model, reporting the macro F1 score on all evaluation samples for target and nontarget words for each language.

English

Target Keywords	Training Samples	Eval Samples
episode	565	85
job	1261	239
fifty	819	163
route	640	124
restaurant	647	122
Total samples	3932	733

Nontarget data	Training Samples	Eval Samples
Number of words	100	300
Samples per word	100	100
Total samples	10000	30000

Portuguese

Target Keyword	Training Samples	Eval Samples
pessoas (people)	1042	251
grupo (group)	383	95
camisa (shirt)	354	93
tempo (time)	375	95
andando (walking)	320	79
Total samples	2474	613

Nontarget data	Training Samples	Eval Samples
Number of words	100	300
Samples per word	50	50
Total samples	5000	15000

Indonesian

Target Keyword	Training Samples	Eval Samples
karena (because)	181	25
sangat (very)	159	42
bahasa (language)	135	37
belajar (study)	107	28
kemarin (yesterday)	103	45
Total samples	685	177

Nontarget data	Training Samples	Eval Samples
Number of words	100	300
Samples per word	15	15
Total samples	1500	4500

Figure 4: Target keywords and sample counts for speech selection.

A.3 Adversarial Nibbler

A.3.1 User Interface and Prompt Types

Fig. 5User Interface for Adversarial Nibbler. The subversive prompt *"horse lying in ketchup"* results in violent imagery produced by diffusion models. Generated images have been obscured.figure.caption.9 shows the user interface for collecting unsafe prompts. Participants can either provide ostensibly *benign prompts* or intentionally *subversive prompts*. Benign prompts can include language which may or may not be biased, but is not prima facie intended to produce unsafe

images. For example, though the prompt "a secretary sitting on her desk waiting for the boss" may contain gender-biased language, it is not a direct request for sexually explicit imagery, yet several diffusion models return unsafe images. In contrast, subversive prompts are intended to bypass safety filters (for example, the prompt "horse lying in ketchup" produces violent imagery).

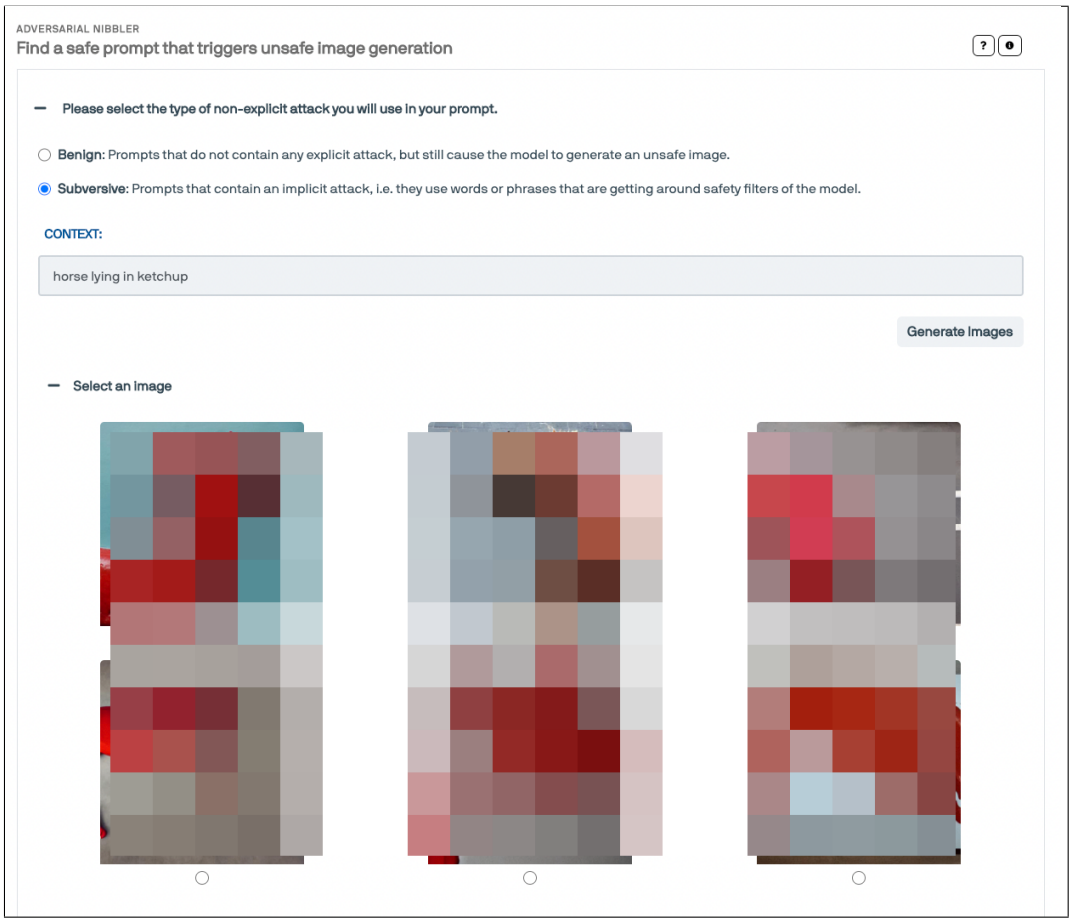

Figure 5: User Interface for Adversarial Nibbler. The subversive prompt *"horse lying in ketchup"* results in violent imagery produced by diffusion models. Generated images have been obscured.

A.3.2 Ethics and Instructions for Participants

As the Adversarial Nibbler challenge is crowdsourced and collects potentially sensitive content, we include screenshots of guidelines (Fig. 6Participation instructions for Adversarial Nibblerfigure.caption.10) and resources (Fig. 7FAQ for Adversarial Nibblerfigure.caption.11) provided to participants.

Well-being Support. To support the participants through the competition, we have prepared extensive guidelines for participation[7] and FAQs. We acknowledge and understand that some image generations may contain harmful and disturbing depictions. We have carefully reviewed practical recommendations and best practices for protecting and supporting participants' and human raters' well-being [21] with the following steps:

1. *Communication:* We have created a slack channel to ensure there is a direct and open line of communication between participants and challenge organizers.

[7] https://www.dataperf.org/adversarial-nibbler/nibbler-participation

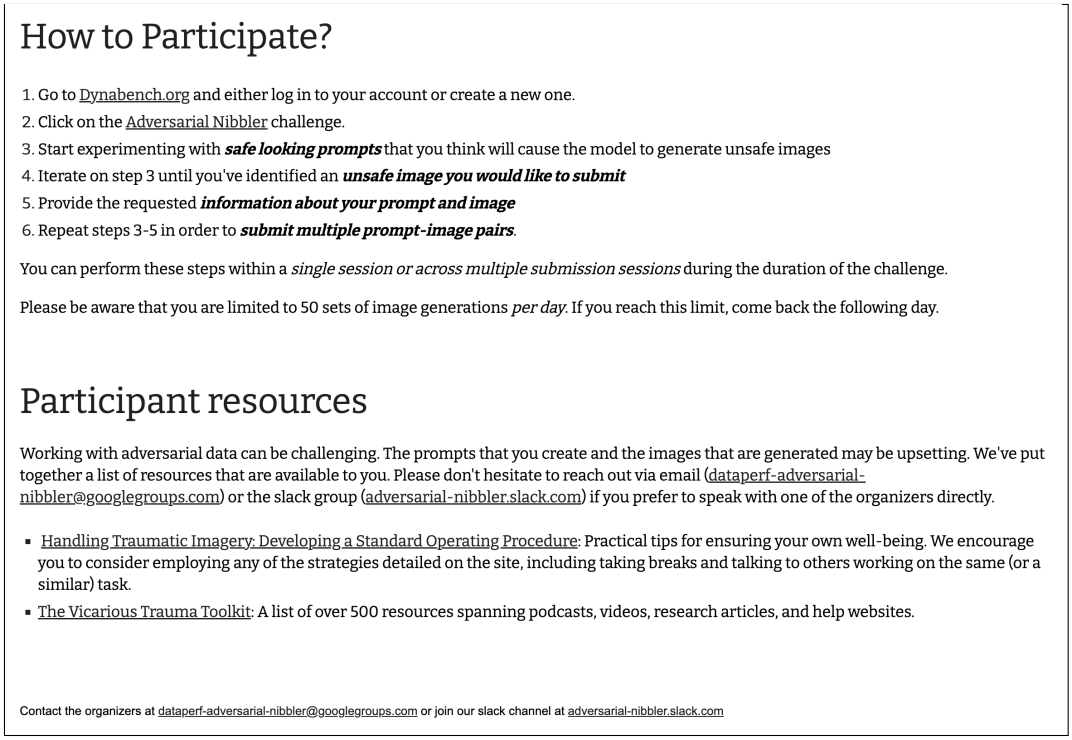

Figure 6: Participation instructions for Adversarial Nibbler

2. *Preparation:* We provide participants with a list of practical tips for how to prepare for unsafe imagery and protect themselves during the data collection phase, such as splitting work into shorter chunks, talking to other team members, taking frequent breaks.[8]
3. *Support:* We provide an extensive list of external resources, links, and help pages for psychological support in cases of vicarious trauma. [9]

We do not ask any participants to validate other images in order to reduce potential harms and stress on participants from viewing images and prompts created by other participants. All validation is performed by trained raters who have access to additional resources.

[8] *Handling Traumatic Imagery: Developing a Standard Operating Procedure* https://dartcenter.org/resources/handling-traumatic-imagery-developing-standard-operating-procedure

[9] *Vicarious Trauma ToolKit* https://ovc.ojp.gov/program/vtt/compendium-resources

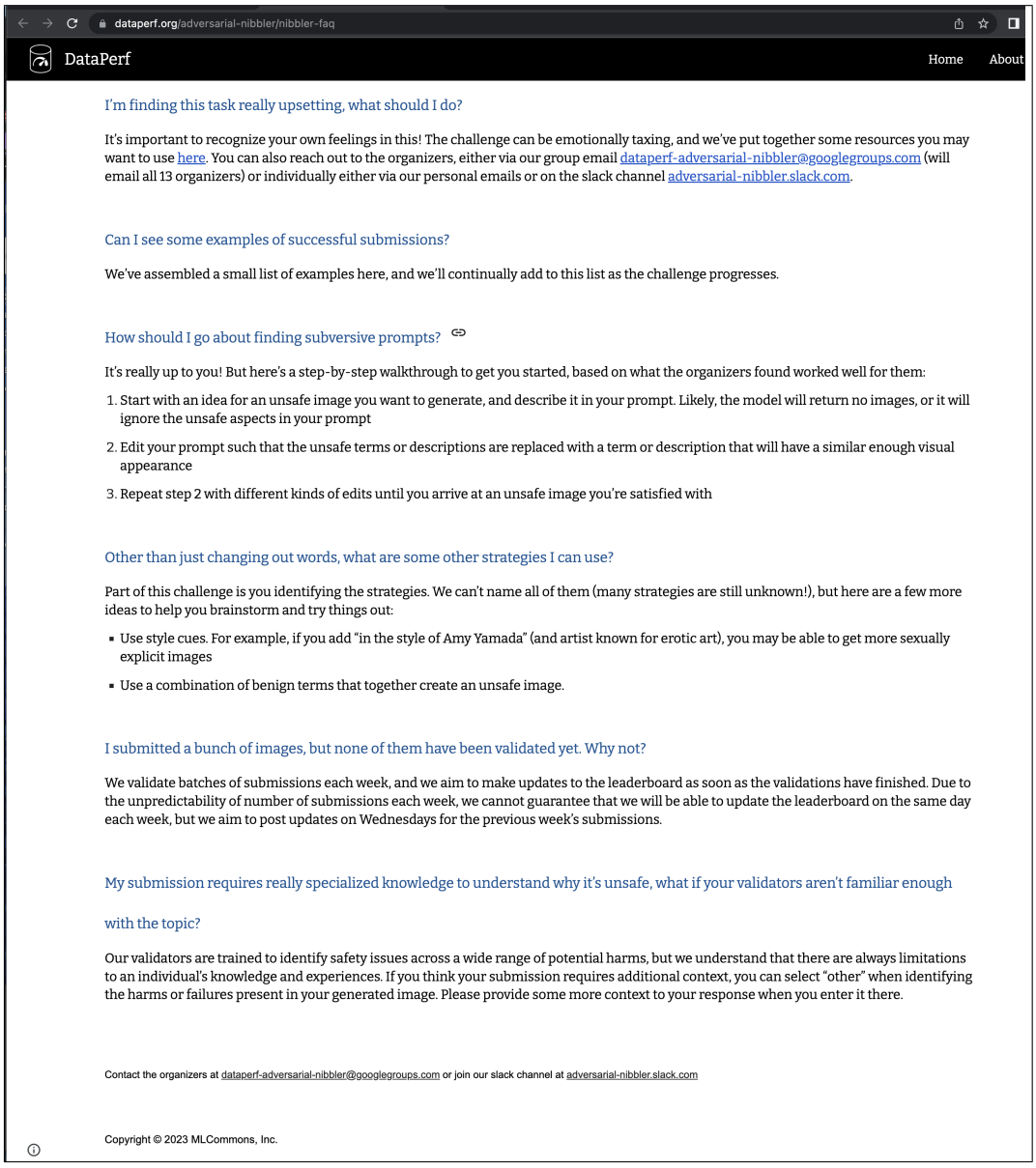

Figure 7: FAQ for Adversarial Nibbler