Are We Learning Yet? A Meta-Review of Evaluation Failures Across Machine Learning

Thomas Liao
Scale AI
thomas.liao@scale.com

Rohan Taori
Stanford University
rtaori@stanford.edu

Inioluwa Deborah Raji
UC Berkeley
rajiinio@berkeley.edu

Ludwig Schmidt
Toyota Research Institute
University of Washington
schmidt@cs.uw.edu

Abstract

Many subfields of machine learning share a common stumbling block: evaluation. Advances in machine learning often evaporate under closer scrutiny or turn out to be less widely applicable than originally hoped. We conduct a meta-review of 107 survey papers from computer vision, natural language processing, recommender systems, reinforcement learning, graph processing, metric learning, and more, organizing a wide range of surprisingly consistent critique into a concrete taxonomy of observed failure modes. Inspired by measurement and evaluation theory, we divide failure modes into two categories: internal and external validity. Internal validity pertains to evaluation on a learning problem in isolation, such as improper comparisons to baselines or overfitting from test set re-use. External validity relies on relationships between different learning problems, for instance, whether progress on a learning problem translates to progress on seemingly related tasks.

1 Introduction

Most empirical papers in machine learning follow the benchmarking paradigm for evaluation. There is a myriad of datasets and tasks in the literature, and what it means for a machine to "learn" has interpretations from mirroring human-like intelligence to solving a specific practical task. Nevertheless, whether a new method has merit is usually determined by evaluating a trained model on a held-out test set and comparing its performance to prior work. If the new model improves over the relevant baselines, the method represents an algorithmic contribution. Since the benchmark itself is often only a challenge problem specifically constructed for research, the underlying assumption is that the new method will also yield performance improvements on real-world problems similar to the benchmark.

Benchmarking was popularized in machine learning in the 1980s through the UCI dataset repository and challenges sponsored by DARPA and NIST [24, 35, 55, 81]. Since then, benchmark evaluations have become the core of most empirical machine learning papers. The impact of benchmarking is illustrated by the ImageNet competition [31, 130], which seeded much of the excitement in machine learning since 2010. Winning entries such as AlexNet [77] and ResNets [57] have become some of the most widely cited papers across all sciences.

Evaluating algorithmic progress with benchmarks is a double-edged sword. On the one hand, benchmarks come with a clearly defined performance metric that enables objective assessments of

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

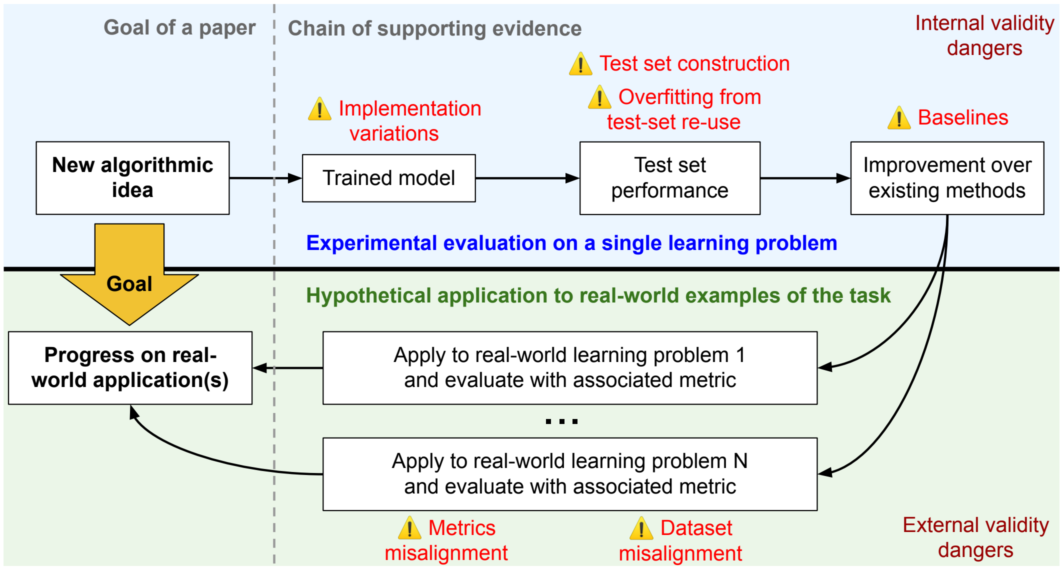

Figure 1: Our framework for benchmark-based evaluations of machine learning algorithms and associated validity concerns. In the benchmark paradigm, papers which propose a new algorithmic idea demonstrate its effectiveness by comparing to results of prior work on a specific learning problem (the benchmark). The underlying assumption is that the benchmark is representative for a broader task and hence the performance improvements will transfer to real-world applications. This chain of reasoning relies on multiple steps with various potential validity issues.

different algorithms. On the other hand, summarizing a new algorithm with a single performance number creates an illusion of simplicity that ignores the many underlying assumptions in the learning problem posed as a benchmark. Indeed, an increasing number of machine learning papers take a critical perspective on recent algorithmic advancements and find important flaws in current evaluation practices. For instance, most claimed advances from the past few years of recommender systems research failed to improve over established baselines and evaporate under closer scrutiny [25, 124]. Given the key role benchmarking plays in machine learning, such evaluation flaws threaten to undermine the perceived algorithmic gains in recent years.

In this paper, we provide a systematic taxonomy of failures in the benchmarking paradigm in order to put current evaluation practices on solid foundations. Our taxonomy draws from 107 analysis papers which study specific machine learning evaluations; we describe further how we arrived at this taxonomy in Appendix 6. Despite the diversity of tasks and algorithms, we find that the same evaluation failures repeat across diverse areas such as computer vision, natural language processing, recommender systems, reinforcement learning, graph processing, metric learning, and more. Based on lessons from evaluation theory [92], we divide the failure modes into two categories:

- **Internal validity** refers to issues that arise within the context of a single benchmark.
- **External validity** asks whether progress on a benchmark transfers to other problems.

Figure 1 illustrates our taxonomy of evaluation failures in machine learning. Our taxonomy can serve as a resource for machine learning researchers and practitioners to check for evaluation issues in their own disciplines. Since many failure modes occur in several fields, insights from one field will transfer to others. Additionally, our paper contributes insights to the ongoing discussion around evaluation practices in machine learning. Finally, our taxonomy of external validity criteria offers a starting point for research in this area. The relationships between different datasets and learning problems are not yet well understood; more work is needed to understand the scope of current benchmarks.

Next we introduce our framework for evaluation validity in machine learning, which organizes the common failures modes described in Sections 3 and 4. Section 5 then discusses limitations of the benchmarking paradigm itself before we conclude in Section 6. An overview of the papers that inform this survey can be found in Appendices D and E.

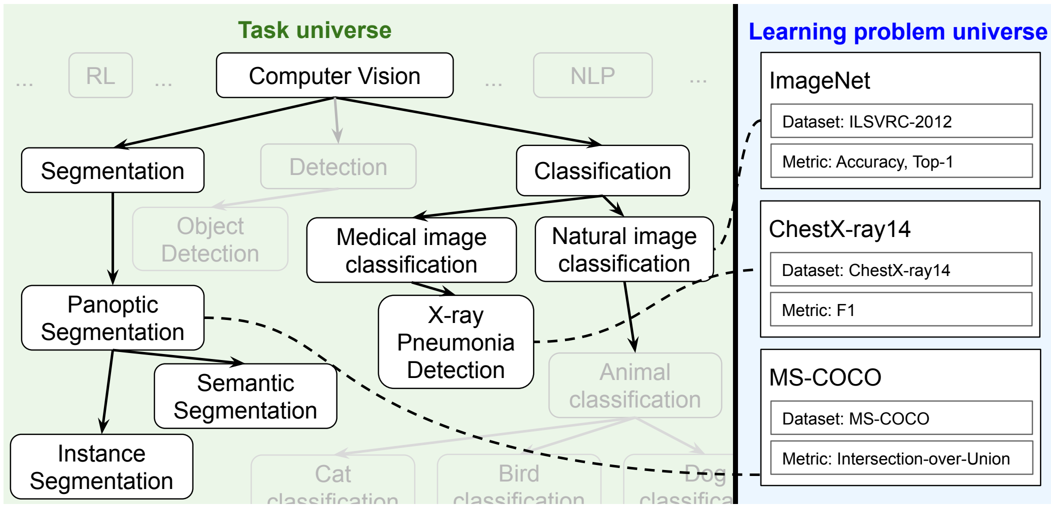

Figure 2: An example of a task hierarchy and associated learning problems. Tasks are abstract problem statements formulated independently from datasets and exist at various levels of granularity, giving rise to a hierarchy. In contrast, a learning problem combines a specific dataset and a particular metric to instantiate one or more tasks. Many learning problems can attempt to instantiate the same task, and the relationships between different learning problems is the focus of external validity.

2 A conceptual framework for machine learning evaluations

Empirical machine learning evaluations are ultimately tied to datasets. A key question is to what extent the datasets used to measure algorithm performance (e.g., ImageNet [31, 130] or GLUE [157]) represent the problem a paper claims to address (e.g., image classification or natural language understanding). To make this distinction clear from the beginning, we define two different kinds of problem statements. These two notions for "learning from data" distinguish between concrete problems defined via *datasets* and abstract problems defined via formal or informal *semantics*.

2.1 Two kinds of problem statements: learning problems vs. tasks

Learning problems. A learning problem comprises a dataset of (input, output) pairs and an associated evaluation metric for scoring proposed solutions (functions from the input to the output space). A learning problem is fully defined by these two parts and requires no further reference to external semantics or data; e.g., the ILSVRC-2012 dataset (ImageNet) with top-1 accuracy as metric.

Tasks. A task is a problem statement defined abstractly, either via natural language or in a formal way. A task does not necessarily have a single true definition and we do not aim to establish any task definitions. Tasks can exist at varying granularities, e.g., from "dog vs. cat classification" to "animal classification" to "image classification", which naturally gives rise to a hierarchy (see Figure 2). Tasks are omnipresent in the machine learning literature as a way to frame contributions. For the purpose of evaluation, tasks are usually instantiated by learning problems. As an example, MNIST, CIFAR-10, and ImageNet all instantiate the "image classification" task.

Given these definitions, a *benchmark* is a learning problem framed as an indicator of progress on some task. Benchmarks usually come with a leaderboard, competition, or other context that establishes the current state of the art. For example, improving accuracy on ImageNet can be considered as making improvements on the image classification task in the context of the ILSVRC competition [130].

2.2 Internal and external validity in machine learning evaluations

The distinction between learning problems and tasks also separates validity issues in machine learning into *internal* issues, i.e., issues arising within the context of a single learning problem, and *external* issues, i.e., issues stemming from the relationship between a learning problem and broader tasks.

Internal validity. In the evaluation literature, internal validity is about consistency *within* the specified context of the experimental setup [92]. In machine learning evaluations, we use internal validity to refer to validity properties within a learning problem. If these properties are not satisfied, then the experimental measurement itself is invalid. Examples of internal validity problems in machine learning are comparisons to insufficient baselines or overfitting from test set re-use, both of which invalidate claimed improvements over the state-of-the-art on a given learning problem.

External validity. External validity is about the ability to extrapolate – to make valid conclusions for contexts outside the experimental parameters [92]. In machine learning, we use external validity to refer to connections between specific learning problems and the broader tasks they are meant to represent. This goes beyond test set performance on an individual learning problem and is anchored to expectations for performance on one learning problem to transfer to other related learning problems. For instance, external validity issues can arise from limitations of the benchmark dataset or a mismatch in the evaluation metrics of interest.

Internal validity criteria are well known in the field. But despite the seeming simplicity of these failure modes, their recurrence across different areas indicates that machine learning currently has not yet identified nor implemented mechanisms needed for rigorous evaluation. The in-depth study of external validity criteria has only begun recently as more research datasets and concrete applications have become available. Since many popular machine learning benchmarks do not represent real applications but instead are constructed solely for the purpose of comparing learning algorithms, investigating the external validity of these benchmarks is particularly important.

3 Internal validity

In this section, we provide examples of recurring *internal validity* issues that arise within the benchmarking paradigm. In particular, we discuss implementation variations, errors in test set construction, overfitting from test set reuse, and comparisons to inadequate baselines.

3.1 Implementation variations

Different implementations of the same algorithm or metric should behave as close to identical as possible. Variations in behaviour can cause variations in performance, making comparisons difficult if it is unclear which implementation is being referred to. This can result in situations where multiple implementations of ostensibly the same algorithm are effectively distinct methods. We describe specific cases of implementation variations leading to internal validity failures here, and continue with more examples in Appendix B.1.

Algorithms. Ancillary details of an algorithm implementation, often dubbed "tricks", can significantly affect performance. These details are often undocumented in the paper, so subsequent implementations of the algorithm are coded differently. Consider the variation observed by [59] for algorithms in deep reinforcement learning (deep RL): across three implementations of Trusted Region Policy Optimization (TRPO), and three implementations of Deep Deterministic Policy Gradients (DDPG), the best codebase was several factors better than the next best. On OpenAI HalfCheetah-v1 [19], the best TRPO codebase achieved an average reward of nearly 2,000 versus 500, and the best DDPG implementation reached a best average reward of 4,500 versus 1,500 [59].

Metrics. Unexpected differences in metric scores caused by implementation variations hinder proper comparisons. In machine translation, the widely-used BLEU score [111] depends on certain parameters which are often unspecified, such as the maximum n-gram length. Further, researchers can silently manipulate the score with changes like adding or removing tokenization, or lowercasing text [115]. Tweaking all these levers in unison results in BLEU score variations of as much as 1.8 BLEU [115] (for context, the gap between the #1 and #2 for one MT dataset as tracked by Papers with Code is 0.14 BLEU [110]). The use of a standardized library such as SACREBLEU [115] to ensure reproducible parameters can help alleviate issues with metric implementations.

Libraries. Research code relies on frameworks and libraries to implement common functions. If these libraries aren't coded correctly, evaluation is undermined. Between the Python Image Library (PIL), PyTorch, OpenCV, and TensorFlow, only PIL correctly downsamples a circle without introducing aliasing artifacts [112]. Consequently, implementations of the Frechet Inception Distance (FID) [63], which is used to evaluate generative models, would report different scores for the same models [112].

3.2 Errors in test set construction

Even if implementations of algorithms are reliable, flaws in a test set's construction can distort the performance reported on a given learning problem in a few different ways.

Label Errors. Several researchers have long articulated concern for the correctness of data labels as an indicator of internal validity [17, 105]. However, it remains unclear how much such errors impact performance measurement, if at all, especially for deep learning [138]. A subset of label errors are due to more conceptually consistent disagreements between annotators [27] or dataset bias [145]; these types of errors are more appropriately construed as external validity issues, and are described further in Section 4.4.

Label Leakage. At times, data features accidentally contain direct information about the target variable in a way that makes the learning problem redundant [70]. For instance, a bank account number could be included as a feature to predict the individual has an open account.

Test set size. Evaluating a model on a finite-sized test set always leaves uncertainty about the actual performance on the underlying distribution the test set is sampled from. If a test set is too small to detect performance differences between two models, random variation in the test set scores can lead to misinterpreting one method as superior to another [16, 22]. In Appendix B.2 we provide more technical details about appropriate test set sizes.

Contaminated Data. Flaws in the dataset construction process may lead to unintentional inclusions of examples that cause problems during evaluation. For example, [8] find that 10% of the images from the CIFAR-100 [76] test set have duplicates in the training set. After deduplication, model performance drops by as much as 14% (relative), demonstrating that the contaminated data leads to overestimation of model performance. Similarly, cross-validation or testing on time-series must be handled with care so as to not include future data in the training set [23]. Examples which are not drawn from the distribution of interest can also distort apparent model performance. Machine translation models perform worse on test sets with more translation artifacts [80]. Models perform up to twice as well on test sets that exclude certain kinds of poor translations as they do on test sets which don't filter these examples out.

3.3 Overfitting from test set reuse

When evaluating a model on a test set, we are not interested in performance on the specific test examples, but more generally in performance on similar data. Formally, we hope that the model generalizes to data from the same distribution. The connection between the test set and its corresponding data distribution is only guaranteed if the test set is not reused frequently. This is a core assumption in test set evaluations and is commonly recognized in lecture notes and textbooks [56, 100].

Researchers routinely undermine this assumption by repeatedly reusing popular test sets for model selection, raising concerns about the validity of benchmark results. However, even decade-long test set reuse has surprisingly resulted in little-to-no overfitting on popular benchmarks such as MNIST, CIFAR-10, ImageNet, SQuAD, the Netflix Prize, and more than 100 Kaggle classification competitions [97, 122–124, 127, 162]. While these findings are good news for the benchmark paradigm, they also illustrate that our understanding of common evaluation practices is still limited. An active line of research investigates the question of overfitting from test set reuse, also known as adaptive overfitting [5, 9, 14, 37, 42, 91, 174]. Note that the cited experimental studies of overfitting mostly focus on classification. Regression benchmarks may be more affected by test set reuse.

3.4 Comparison to inadequate baselines

Finally, reliably tracking progress on a learning problem requires comparing new methods to existing baselines. In practice, many subtle considerations must be addressed to make proper comparisons. We highlight the biggest recurring themes here; Appendix B.4 contains additional discussion.

3.4.1 Implementing and tuning simple methods

Researchers in machine learning often employ newer, more complex methods, such as those using deep neural networks, to solve a given task, without leveraging simpler methods such as linear models or random search. Attention to smaller details and thorough feature engineering can often make a huge difference for these simple baselines:

- In graph learning, logistic regression combined with simple feature engineering provided comparable performance to neural networks while being orders of magnitude faster [67, 161].
- In recommender systems, [25, 124] found that a well-tuned vanilla matrix factorization baseline with some feature engineering outperformed all newer methods, both neural and non-neural, on recommendation results and collaborative filtering tasks.
- In reinforcement learning, where simple linear or RBF policies were able to solve an array of continuous control tasks [118].
- In information retrieval, where a non-neural method from 2004 is superior to all neural approaches developed through 2019 [163].
- In few-shot classification, where a linear layer on top of a supervised classifier's features provides competitive performance on meta-learning benchmarks [150].
- On tabular clinical prediction datasets, where standard logistic regression was found to be on par with deep recurrent models [10].
- And in adversarial robustness, where early-stopping with standard projected gradient descent was found to give performance on par with newer alternatives [126].

Random search is also frequently overlooked, even though it forms a strong, simple, baseline where applicable. One particularly prominent case is in deep RL, where simple random search, combined with a handful of minor modifications, outperforms many deep RL algorithms on a variety of MuJoCo continuous control tasks [90]. Similarly, for hyperparameter tuning, [79] found that random search combined with early stopping outperformed all existing approaches. And in neural architecture search, [78, 165] found that random search with early stopping and weight sharing found solutions comparable to leading strategies using deep learning. It should be noted that recent NeurIPS competitions found that Bayesian optimization is superior to random search in many settings [154].

3.4.2 Controlling for algorithmic details

Implementations of algorithms often contain details to improve performance which are not described in the text. For example, extensively tuning hyperparameters is often key to achieving optimal performance for a proposed method. Unfortunately, baselines are often not tuned as carefully, inflating apparent gains for the proposed method. Ignoring these consequential details leads to misattributions of why one algorithm is better than another, affecting future research directions. For instance, a series of recent papers have attempted to benchmark a variety of deep metric learning algorithms, controlling for aspects such as network architecture, optimizer, image augmentations, hyperparameter compute budget, etc. [41, 101, 128]. After controlling for these factors, the performance difference for the best methods were marginal at best, and the papers concluded that the majority of perceived gains could instead be attributed to newer methods using significantly better backbone architectures (e.g., ResNet50 instead of GoogleNet) and unequal hyperparameter compute budgets. These results very closely mirror results from a variety of other settings, such as deep semi-supervised learning algorithms [108], graph neural networks [36, 139], domain generalization [53], and generative adversarial networks [88]. Inconsistencies in backbone architectures and unequal tuning budgets was a common, recurring failure mode across these papers.

4 External validity

Developing tailored algorithms for specific learning problems is usually not the end goal of machine learning research; rather, the hope is that the ideas and contributions will apply to broader scenarios. How much one expects progress to transfer is a subjective judgment based on factors such as the learning problems involved, the domain knowledge required, and the details of the algorithm itself. We refer to this as *external validity*, as it involves relationships between two or more learning problems. In this section, we first discuss and define two sub-types of transfer that occur within external validity, then provide examples where evaluation issues have arisen.

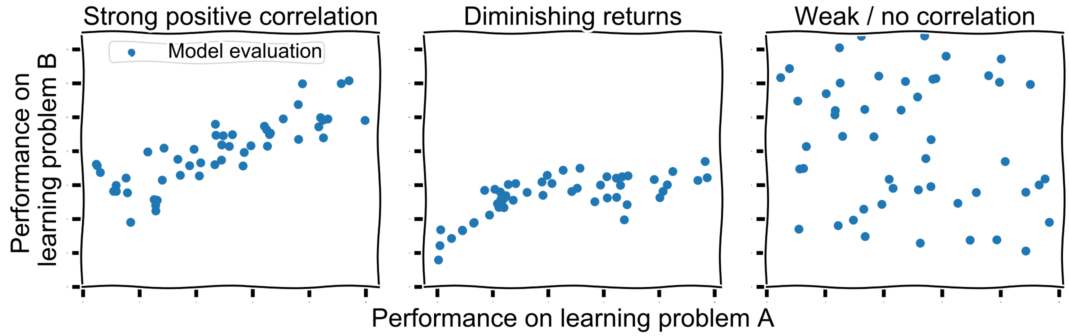

Figure 3: Learning problem transfer can happen to varying extents. Progress on learning problem A may transfer to learning problem B universally (**left**). However, progress may also plateau (**middle**) or there may be no correlation between performance on the two learning problems (**right**).

4.1 Types of transfer

Algorithm transfer. The claim that a certain algorithm "generalizes well to other problems" is a claim about *algorithm transfer*: the correlation between (i) the relative performance of an algorithm over one or more baselines on one learning problem to (ii) the relative performance of the same algorithm over a one or more baselines on another learning problem. Consider ResNets [57] when they were introduced: adding residual connections (allowing for a deeper net) lead to better performance on ImageNet than VGG [141], a baseline algorithm. On CIFAR-10, ResNets also outperform VGG, an appropriate baseline choice, so we say that ResNets transfer well from ImageNet to CIFAR-10.

Learning problem transfer. Now we introduce *learning problem transfer*: the correlation of performance trends over all algorithms for one learning problem with performance trends over all algorithms for another learning problem. Whereas algorithm transfer is about the relative performance of a specific algorithm between learning problems, learning problem transfer asks about the relative progress of algorithms in general between learning problems. For example, as models have improved on the ImageNet benchmark, the same models are used on the CIFAR-10 benchmark, and show continued progress there also. If algorithms never transferred well between learning problems, then progress on one learning problem would never transfer to another. This is visualized in Figure 3 (right), which illustrates low or no correlation between performance on two learning problems. If the correlation weakens over time, this is the "diminishing returns" scenario shown in the middle subplot. And if there is strong positive correlation, then the picture is similar to the first subplot.

Achieving progress in machine learning requires progress on "friendly" learning problems which exhibit strong learning problem transfer; otherwise, researchers would have to start from scratch on every novel learning problem. How can we predict how performance will correlate between two learning problems? There are some common patterns in the literature that allow us to more concretely grapple with learning problem transfer. The community has developed specific out-of-distribution (OOD) test sets for certain problems, such as image corruptions in image classification [60], heuristics-based counterexamples within language inference [94], and a number of "in-the-wild" distribution shifts [6, 61, 62, 74, 123, 158]. Cast in terms of our framework, these OOD benchmarks alter the data distribution of the learning problem, but otherwise remain very close to the original learning problem in the task hierarchy. On the other hand, one may consider transfer of progress between learning problems that are further apart in the task hierarchy, such as from image classification on ImageNet to image segmentation on COCO. In general, as Figure 2 illustrates, the closer two learning problems are in one's conception of the task hierarchy, the greater one may expect positive transfer of progress.

Leaving a more fine-grained discussion of the various of categories of transfer to Appendix A.2, we now explore examples from the literature pointing out failures of learning problem transfer. Since a learning problem is defined as a dataset plus a metric, a failure in transfer can be attributed to either a misalignment in the datasets or a misalignment in the metrics. Such a misalignment reflects the inconsistencies that arise when boiling down an idealized task into concrete learning problems. Resolving these inconsistencies in either the dataset or the metric may require re-annotating the data or collecting new data; therefore, misalignments are usually baked into the benchmark once the dataset has been constructed and the design choices locked in. All future modeling work on the benchmark inherits the same misalignment problems, underscoring the need for a better understanding of the external validity of commonly used benchmarks.

4.2 Metrics misalignment

We use *metric* to mean any algorithm or procedure which, given a model and a dataset, returns a number or score which is interpreted as the performance of the model on that dataset. This definition encompasses not only mathematically defined metrics like accuracy, precision, and recall, but also metrics parameterized by models (Frechet Inception Distance [63], BERT [111], BLEURT [136]), and metrics which involve humans in the loop, like human evaluations of machine translation (Direct Assessment [143], Relative Ranking [51]).

A metric which fails to adequately distinguish between two algorithms that perform differently fails to capture what it means to do well on the learning problem. For example, a good representation learning algorithm should cluster items of the same class together tightly and separate clusters of different classes widely. Papers for representation learning usually report the F1, Recall@K, and Normalized Mutual Information (NMI) metrics. However, all three metrics fail to reward algorithms which have a greater separation between different classes [101]. Even more egregiously, NMI returns higher scores for datasets with more classes, regardless of the algorithm's performance [101].

Researchers may prefer to measure an idealized metric whose use is precluded by practical considerations like money or time, and therefore substitute another metric instead to form a proxy learning problem. For example, many have argued that human evaluation is the 'gold standard' for machine translation [50, 69, 87], but waiting for humans to evaluate translations takes much longer and is much more expensive than computing BLEU [111], an automatic metric. In certain cases, human rankings of translations contradict the BLEU ordering [38, 170].

4.3 Comparisons to human performance

Comparing algorithms to humans requires more nuance than any one given learning problem provides. Matching a human baseline on a specific learning problem does not automatically imply human-level performance on other similar similar problems without more evidence. For one, instantiating a task into some learning problem often strips out context which meaningfully affects evaluation. In translation, for example, the work of human translators tends to be evaluated as a complete text, whereas machine translation competitions compare hypothesis sentences to reference sentences, meaning that erroneous translations which are apparent only in context are missed [151].

Further, claims to "super-human" performance on a given learning problem is related to but does not always translate to "human-like" reasoning or ability [44] – for instance, contemporaneous models suffer performance drops with only small changes of the learning problem that don't affect humans as badly (e.g. models [64] on CIFAR-10 [76]). Claimed improvements by themselves are thus only applicable to the given learning problem, and aren't sufficient to prove machine superiority on the broader task or application.

4.4 Dataset misalignment

Specific decisions made about data collection and curation are increasingly acknowledged as highly consequential to model outcomes [113, 131]. Any failure to transfer from one learning problem to another learning problem or broader task is often tied to the data choices involved. Because of the cost and effort involved in annotation and data collection, these decisions can have a broader impact than failures contained to a single modeling paper. In the next two subsections, we explore how specific choices in dataset curation can hinder an algorithm's ability to transfer. Refer to Appendix B.5 for additional discussion and examples.

4.4.1 Reliance on simple, inappropriate heuristics

We found several examples where gaps in the data collection process lead to models performing well on a given learning problem by relying on data quirks which do not characterize the overall task. For instance, [107] discovered that sub-par clinical performance of X-ray image classification models was in part due to an unintended correlated variable in the training data: classifiers trained to predict whether an X-ray image presented a collapsed lung were failing disproportionately on new positive diagnoses. It was discovered that a majority of the positive training images actually contained visible chest drains, a treatment for the condition. Thus, models achieved a high accuracy on the learning problem by identifying whether a chest drain was present, but completely sidestepped the original purpose of the task. After removing the spurious feature, by filtering out chest drain images, model performance dropped significantly, by over 20% on clinically relevant subsets of the data.

More examples of models exploiting simple dataset-level heuristics abound. The authors of [49] found that on the Visual Question-Answering dataset [4], models could exploit strong label imbalance on certain questions. For example, for a question beginning with "Do you see a...", a model always outputting "yes" – without considering the rest of the question or the actual image – can achieve an accuracy of 87%; correcting this imbalance in the test set led to accuracy drops of up to 12% among yes/no questions for these models. Similarly, models trained on part of a reading comprehension task (either questions only or passages only) achieve a surprisingly high accuracy [71].

Landmark studies found that language models regularly exploit such "spurious patterns" across a wide range of NLP tasks [46, 72]. On the MNLI natural language inference benchmark, the presence of a negation operator (e.g. "not", "no", etc.) dictates the label probability to a greater degree than the actual input prompts [94]. Similarly, the authors of [104] find that BERT models trained on comprehension datasets (e.g. ARCT [54]) exploit the presence of negation operators, and removing such cues drops the model to random chance accuracy. These correlates were discovered by using humans to augment the training data to be consistent with counterfactual labels. When evaluated on these counterfactual subsets, model performance drops by as much as 30% in multiple cases.

4.4.2 Sensitivity to real-world distribution shift

There are also many cases where an algorithm is expected to perform in a broader variety of scenarios than it is trained on. In such cases, the inability to transfer is not caused by exploiting specific obvious heuristics as much as it is caused by a failure to extrapolate to different real-world data distributions. For example, most models trained on ImageNet were found to experience a considerable drop in accuracy when exposed to images that contained a larger amount of natural variation, such as changes in pose, lighting, object composition, etc. [147]. Similarly, models trained for the original SQuAD dataset performed poorly when evaluated on data collected from different source domains, such as Amazon crowd reviews and Reddit posts [97].

In the medical domain, models developed in one institution for diagnosing pneumonia in radiographs or classifying pathology tissue slides may not translate to other hospitals for practical reasons such as differences in equipment and patient populations [74, 166]. Similarly, [73] find in a learning problem transfer analysis from ImageNet to chest X-ray classification on CheXpert [68] that, while ImageNet pre-training helps models achieve higher performance on CheXpert, models with higher ImageNet accuracy are not likely to provide higher CheXpert performance.

4.4.3 Dataset Bias & Disagreement

At times, the misalignment perceived between the learning problems is the result of various forms of data bias [145]. Some data sources can omit or under-represent certain sub-populations and as a result, evaluation measurements will disguise failures for these under-represented population subgroups [119]. For example, facial recognition benchmarks drastically under-represent darker and female faces [96], making it difficult to perceive when models fail to perform acceptably for this subgroup [7, 20]. Furthermore, inappropriate stereotyped associations can be perpetuated by the systematic use of offensive, incorrect or exclusionary labels for certain mistreated subgroups [116, 144]. At times, societal discrimination can also lead to false labels being more common in one group than another [99]. Discrepancies between learning problem datasets may also arise from inherent contextual differences - data sourced from differing geographies or cultural context [29, 137], in addition to annotators with inherently differing viewpoints regarding ground truth [27, 48].

4.5 Evaluation quantification

The aforementioned examples of metric and dataset mislignment suggest that reliably measuring progress in machine learning requires evaluating on multiple learning problems associated with a

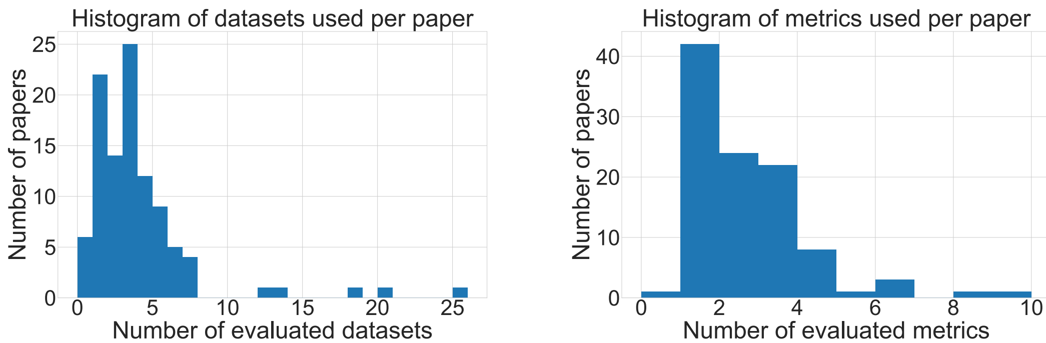

Figure 4: A histogram of the number of datasets used for evaluation by each paper in our sample pool (**left**), and a similar histogram for the number of metrics (**right**). Most of the papers (>65%) evaluate on 3 datasets or fewer, and a similar fraction (>65%) evaluate on 2 or fewer metrics.

particular task. If a proposed method provides gains in a variety of different contexts, one can be more confident in the performance on future learning problems instantiating the task.

To better understand community practices around benchmarking and provide some context around our analysis and framework, we annotated a random sample of machine learning benchmarking papers with the number of distinct datasets and the number of distinct metrics each paper used for evaluation. Concretely, we randomly sampled 140 papers from the past five years (2016–2021) of NeurIPS, ICML, EMNLP, and CVPR, and filtered out papers which were not applicable to the benchmarking paradigm (37 papers). The results of our analysis for the remaining 103 papers are presented in Figure 4. On average, papers evaluated on an average of 4.1 datasets and 2.2 metrics. Overall, most of the papers in our sample (>65%) evaluate on 3 datasets or fewer, and a similar fraction (>65%) evaluate on 2 metrics or fewer. Although we cannot recommend a "correct" number of learning problems to evaluate on, as this is a domain-specific consideration based on the task and specific learning problems, our data provides evidence that many papers evaluate on a small number of datasets and metrics, which indicates that studying alignment between these learning problems can be a helpful guide for future research. We provide more detail about our paper collection and annotation procedure, as well as confidence intervals for our mean estimates, in Appendix C.

5 Broad critiques of benchmarks & competitive testing

Researchers have described several limitations to the benchmarking paradigm in machine learning. Most obviously, the use of benchmarks to assess progress in the field creates a competitive testing dynamic that emphasizes outcomes rather than proper scientific inquiry [66]. The absence of community norms like reproducibility guidance [34, 114], documentation standards [98] or statistical significance testing [16] makes relying on outcomes-based approaches to evaluate progress even more questionable [13]. Behavior-based alternatives to the benchmarking paradigm, such as test suites [1, 125, 169], for example, can re-orient ML evaluation away from its current focus on the competitive determination of "state of the art", and more towards an exploratory and descriptive probing of model capabilities [65, 106, 142, 160, 168]. Furthermore, the learning problems we embody as benchmarks go a long way in focusing community attention on a set of specific applications and tasks, not all of which are ideal or value-aligned. For instance, the lack of consideration for other aspects of performance in ML evaluation, such as model efficiency, privacy or fairness, plays a big role in disincentivizing researchers from paying attention to such issues [40, 135].

6 Conclusion

The benchmarking paradigm has served as a valuable guide for progress in the past. However, the next phase of machine learning innovation and deployment will require more sophisticated evaluation practices than comparing one-dimensional performance numbers on a single test set. We hope that our taxonomy offers a starting point for both experimental and theoretical research in this area, and that the field will invest in a more robust understanding of the evaluation practices that inform our shared perception of progress.

References

[1] Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., and Zimmermann, T. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pp. 291–300. IEEE, 2019.

[2] Amrhein, V., Greenland, S., and McShane, B. Scientists rise up against statistical significance. *Nature*, 2019. https://www.nature.com/articles/d41586-019-00857-9.

[3] Andrychowicz, M., Raichuk, A., Stańczyk, P., Orsini, M., Girgin, S., Marinier, R., Hussenot, L., Geist, M., Pietquin, O., Michalski, M., et al. What matters in on-policy reinforcement learning? a large-scale empirical study. *arXiv preprint arXiv:2006.05990*, 2020.

[4] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

[5] Arora, S. and Zhang, Y. Rip van Winkle's Razor: a simple estimate of overfit to test data, 2021. https://arxiv.org/abs/2102.13189.

[6] Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. http://papers.nips.cc/paper/9142-objectnet-a-large-scale-bias-controlled-dataset-for-pushing-the-limits-of-object-recognition-models.

[7] Barocas, S., Guo, A., Kamar, E., Krones, J., Morris, M. R., Vaughan, J. W., Wadsworth, D., and Wallach, H. Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. *arXiv preprint arXiv:2103.06076*, 2021.

[8] Barz, B. and Denzler, J. Do we train on test data? purging cifar of near-duplicates. *Journal of Imaging*, 6(6):41, Jun 2020. ISSN 2313-433X. doi: 10.3390/jimaging6060041. URL http://dx.doi.org/10.3390/jimaging6060041.

[9] Bassily, R., Nissim, K., Smith, A., Steinke, T., Stemmer, U., and Ullman, J. Algorithmic stability for adaptive data analysis. In *Symposium on Theory of Computing (STOC)*, 2016. https://arxiv.org/abs/1511.02513.

[10] Bellamy, D., Celi, L., and Beam, A. L. Evaluating progress on machine learning for longitudinal electronic healthcare data. *arXiv preprint arXiv:2010.01149*, 2020.

[11] Bello, I., Fedus, W., Du, X., Cubuk, E. D., Srinivas, A., Lin, T.-Y., Shlens, J., and Zoph, B. Revisiting resnets: Improved training and scaling strategies. *arXiv preprint arXiv:2103.07579*, 2021.

[12] Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and Oord, A. v. d. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[13] Biderman, S. and Scheirer, W. J. Pitfalls in machine learning research: Reexamining the development cycle. *arXiv preprint arXiv:2011.02832*, 2020.

[14] Blum, A. and Hardt, M. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning (ICML)*, 2015.

[15] Bojar, O., Ercegovčević, M., Popel, M., and Zaidan, O. A grain of salt for the wmt manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 1–11, 2011.

[16] Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Mohammadi Sepahvand, N., Raff, E., Madan, K., Voleti, V., et al. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3, 2021.

[17] Bowman, S. R. and Dahl, G. E. What will it take to fix benchmarking in natural language understanding? *arXiv preprint arXiv:2104.02145*, 2021.

[18] Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2015.

[19] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym, 2016.

[20] Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

[21] Callison-Burch, C., Osborne, M., and Koehn, P. Re-evaluating the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[22] Card, D., Henderson, P., Khandelwal, U., Jia, R., Mahowald, K., and Jurafsky, D. With little power comes great responsibility. *arXiv preprint arXiv:2010.06595*, 2020.

[23] Cerqueira, V., Torgo, L., and Mozetič, I. Evaluating time series forecasting models: an empirical study on performance estimation methods. *Machine Learning*, 2020. https://arxiv.org/abs/1905.11744.

[24] Church, K. W. Emerging trends: A tribute to Charles Wayne. *Natural Language Engineering*, 24(1):155–160, January 2018. ISSN 1351-3249, 1469-8110. doi: 10.1017/S1351324917000389.

[25] Dacrema, M. F., Boglio, S., Cremonesi, P., and Jannach, D. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems (TOIS)*, 39(2):1–49, 2021.

[26] D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

[27] Davani, A. M., Díaz, M., and Prabhakaran, V. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *arXiv preprint arXiv:2110.05719*, 2021.

[28] Davis, E. A flawed dataset for symbolic equation verification, 2021.

[29] de Vries, T., Misra, I., Wang, C., and van der Maaten, L. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 52–59, 2019.

[30] DeGrave, A. J., Janizek, J. D., and Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, May 2021. doi: 10.1038/s42256-021-00338-7. URL https://doi.org/10.1038/s42256-021-00338-7.

[31] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. http://www.image-net.org/papers/imagenet_cvpr09.pdf.

[32] Dhillon, G. S., Chaudhari, P., Ravichandran, A., and Soatto, S. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729*, 2019.

[33] Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., and Smith, N. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

[34] Drummond, C. Replicability is not reproducibility: nor is it good science. 2009.

[35] Dua, D. and Graff, C. UCI machine learning repository, 2017. http://archive.ics.uci.edu/ml.

[36] Dwivedi, V. P., Joshi, C. K., Laurent, T., Bengio, Y., and Bresson, X. Benchmarking graph neural networks. *arXiv preprint arXiv:2003.00982*, 2020.

[37] Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. L. Preserving statistical validity in adaptive data analysis. In *Symposium on Theory of computing (STOC)*, 2015. https://arxiv.org/abs/1411.2664.

[38] Edunov, S., Ott, M., Ranzato, M., and Auli, M. On the evaluation of machine translation systems trained with back-translation. *arXiv preprint arXiv:1908.05204*, 2019.

[39] Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., and Madry, A. Implementation matters in deep policy gradients: A case study on ppo and trpo. *arXiv preprint arXiv:2005.12729*, 2020.

[40] Ethayarajh, K. and Jurafsky, D. Utility is in the eye of the user: A critique of nlp leaderboards. *arXiv preprint arXiv:2009.13888*, 2020.

[41] Fehervari, I., Ravichandran, A., and Appalaraju, S. Unbiased evaluation of deep metric learning algorithms. *arXiv preprint arXiv:1911.12528*, 2019.

[42] Feldman, V., Frostig, R., and Hardt, M. The advantages of multiple classes for reducing overfitting from test set reuse. In *International Conference on Machine Learning (ICML)*, 2019. `http://proceedings.mlr.press/v97/feldman19a.html`.

[43] Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181, 2014.

[44] Firestone, C. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571, 2020.

[45] Freitag, M., Grangier, D., and Caswell, I. Bleu might be guilty but references are not innocent. *arXiv preprint arXiv:2004.06063*, 2020.

[46] Gardner, M., Artzi, Y., Basmova, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., et al. Evaluating nlp models via contrast sets. *arXiv preprint arXiv:2004.02709*, 2020.

[47] Goel, K., Rajani, N., Vig, J., Tan, S., Wu, J., Zheng, S., Xiong, C., Bansal, M., and Ré, C. Robustness gym: Unifying the nlp evaluation landscape. *arXiv preprint arXiv:2101.04840*, 2021.

[48] Gordon, M. L., Zhou, K., Patel, K., Hashimoto, T., and Bernstein, M. S. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2021.

[49] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

[50] Graham, Y., Baldwin, T., Dowling, M., Eskevich, M., Lynn, T., and Tounsi, L. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3124–3134, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL `https://www.aclweb.org/anthology/C16-1294`.

[51] Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30, 2017.

[52] Graham, Y., Haddow, B., and Koehn, P. Translationese in machine translation evaluation. *arXiv preprint arXiv:1906.09833*, 2019.

[53] Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[54] Habernal, I., Wachsmuth, H., Gurevych, I., and Stein, B. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1930–1940, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1175. URL `https://www.aclweb.org/anthology/N18-1175`.

[55] Hardt, M. and Recht, B. *Patterns, predictions, and actions: A story about machine learning*. https://mlstory.org, 2021.

[56] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[57] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. https://arxiv.org/abs/1512.03385.

[58] He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 558–567, 2019.

[59] Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[60] Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations, 2019.

[61] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples, 2019. https://arxiv.org/abs/1907.07174.

[62] Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization, 2020. https://arxiv.org/abs/2006.16241.

[63] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017. URL http://arxiv.org/abs/1706.08500.

[64] Ho-Phuoc, T. Cifar10 to compare visual recognition performance between deep neural networks and humans. *arXiv preprint arXiv:1811.07270*, 2018.

[65] Hong, M. K., Fourney, A., DeBellis, D., and Amershi, S. Planning for natural language failures with the ai playbook. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–11, 2021.

[66] Hooker, J. N. Testing heuristics: We have it all wrong. *Journal of heuristics*, 1(1):33–42, 1995.

[67] Huang, Q., He, H., Singh, A., Lim, S.-N., and Benson, A. R. Combining label propagation and simple models out-performs graph neural networks. *arXiv preprint arXiv:2010.13993*, 2020.

[68] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019.

[69] Jurafsky, D. and Martin, J. H. *Speech and language processing, 3rd ed.* 2021.

[70] Kaufman, S., Rosset, S., Perlich, C., and Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6 (4):1–21, 2012.

[71] Kaushik, D. and Lipton, Z. C. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926*, 2018.

[72] Kaushik, D., Hovy, E., and Lipton, Z. C. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

[73] Ke, A., Ellsworth, W., Banerjee, O., Ng, A. Y., and Rajpurkar, P. Chextransfer: performance and parameter efficiency of imagenet models for chest x-ray interpretation. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 116–124, 2021.

[74] Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020.

[75] Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2661–2671, 2019.

[76] Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

[77] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

[78] Li, L. and Talwalkar, A. Random search and reproducibility for neural architecture search. In *Uncertainty in Artificial Intelligence*, pp. 367–377. PMLR, 2020.

[79] Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

[80] Liao, T., Recht, B., and Schmidt, L. In a forward direction: Analyzing distribution shifts in machine translation test sets over time. 2020.

[81] Liberman, M. Fred Jelinek. *Computational Linguistics*, 36(4):595–599, 2010.

[82] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

[83] Liu, N. F., Lee, T., Jia, R., and Liang, P. Can small and synthetic benchmarks drive modeling innovation? a retrospective study of question answering modeling approaches. *arXiv preprint arXiv:2102.01065*, 2021.

[84] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019.

[85] Lopez, A. Putting human assessments of machine translation systems in order. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pp. 1–9, 2012.

[86] Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[87] Louis, A. and Nenkova, A. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300, 2013.

[88] Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. Are gans created equal? a large-scale study. *arXiv preprint arXiv:1711.10337*, 2017.

[89] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

[90] Mania, H., Guy, A., and Recht, B. Simple random search provides a competitive approach to reinforcement learning. *arXiv preprint arXiv:1803.07055*, 2018.

[91] Mania, H., Miller, J., Schmidt, L., Hardt, M., and Recht, B. Model similarity mitigates test set overuse. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. https://arxiv.org/abs/1905.12580.

[92] Mathison, S. *Encyclopedia of evaluation*. Sage publications, 2004.

[93] McCoy, R. T., Min, J., and Linzen, T. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv preprint arXiv:1911.02969*, 2019.

[94] McCoy, R. T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019.

[95] Melis, G., Dyer, C., and Blunsom, P. On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589*, 2017.

[96] Merler, M., Ratha, N., Feris, R. S., and Smith, J. R. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

[97] Miller, J., Krauth, K., Recht, B., and Schmidt, L. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning*, pp. 6905–6916. PMLR, 2020.

[98] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

[99] Mullainathan, S. and Obermeyer, Z. On the inequity of predicting a while hoping for b. In *AEA Papers and Proceedings*, volume 111, pp. 37–42, 2021.

[100] Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[101] Musgrave, K., Belongie, S., and Lim, S.-N. A metric learning reality check. In *European Conference on Computer Vision*, pp. 681–699. Springer, 2020.

[102] Nado, Z., Gilmer, J. M., Shallue, C. J., Anil, R., and Dahl, G. E. A large batch optimizer reality check: Traditional, generic optimizers suffice across batch sizes. *arXiv preprint arXiv:2102.06356*, 2021.

[103] Narang, S., Chung, H. W., Tay, Y., Fedus, W., Fevry, T., Matena, M., Malkan, K., Fiedel, N., Shazeer, N., Lan, Z., et al. Do transformer modifications transfer across implementations and applications? *arXiv preprint arXiv:2102.11972*, 2021.

[104] Niven, T. and Kao, H.-Y. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*, 2019.

[105] Northcutt, C. G., Athalye, A., and Mueller, J. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

[106] Nushi, B., Kamar, E., and Horvitz, E. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6, 2018.

[107] Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159, 2020.

[108] Oliver, A., Odena, A., Raffel, C., Cubuk, E. D., and Goodfellow, I. J. Realistic evaluation of deep semi-supervised learning algorithms. *arXiv preprint arXiv:1804.09170*, 2018.

[109] Pagnoni, A., Balachandran, V., and Tsvetkov, Y. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*, 2021.

[110] PapersWithCode. Wmt en-de benchmark page, 2021. `https://paperswithcode.com/sota/machine-translation-on-wmt2014-english-german`.

[111] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://www.aclweb.org/anthology/P02-1040`.

[112] Parmar, G., Zhang, R., and Zhu, J.-Y. On buggy resizing libraries and surprising subtleties in fid calculation. *arXiv preprint arXiv:2104.11222*, 2021.

[113] Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345*, 2020.

[114] Pineau, J., Sinha, K., Fried, G., Ke, R. N., and Larochelle, H. ICLR reproducibility challenge 2019. *ReScience C*, 5(2):5, 2019.

[115] Post, M. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://www.aclweb.org/anthology/W18-6319.

[116] Prabhu, V. U. and Birhane, A. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020.

[117] Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. Transfusion: Understanding transfer learning for medical imaging. *arXiv preprint arXiv:1902.07208*, 2019.

[118] Rajeswaran, A., Lowrey, K., Todorov, E., and Kakade, S. Towards generalization and simplicity in continuous control. *arXiv preprint arXiv:1703.02660*, 2017.

[119] Raji, I. D. and Buolamwini, J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435, 2019.

[120] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250, 2016. URL http://arxiv.org/abs/1606.05250.

[121] Rajpurkar, P., Jia, R., and Liang, P. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822, 2018. URL http://arxiv.org/abs/1806.03822.

[122] Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

[123] Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

[124] Rendle, S., Zhang, L., and Koren, Y. On the difficulty of evaluating baselines: A study on recommender systems. *arXiv preprint arXiv:1905.01395*, 2019.

[125] Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*, 2020.

[126] Rice, L., Wong, E., and Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

[127] Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J., and Schmidt, L. A meta-analysis of overfitting in machine learning. *Advances in Neural Information Processing Systems*, 32:9179–9189, 2019.

[128] Roth, K., Milbich, T., Sinha, S., Gupta, P., Ommer, B., and Cohen, J. P. Revisiting training strategies and generalization performance in deep metric learning. In *International Conference on Machine Learning*, pp. 8242–8252. PMLR, 2020.

[129] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Li, F. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. URL http://arxiv.org/abs/1409.0575.

[130] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, December 2015. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-015-0816-y.

[131] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. M. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021.

[132] Schick, T. and Schütze, H. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*, 2020.

[133] Schulman, J., Levine, S., Moritz, P., Jordan, M. I., and Abbeel, P. Trust region policy optimization. *CoRR*, abs/1502.05477, 2015. URL http://arxiv.org/abs/1502.05477.

[134] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

[135] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., and Dennison, D. Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28:2503–2511, 2015.

[136] Sellam, T., Das, D., and Parikh, A. P. BLEURT: learning robust metrics for text generation. *CoRR*, abs/2004.04696, 2020. URL https://arxiv.org/abs/2004.04696.

[137] Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., and Sculley, D. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

[138] Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., and Schmidt, L. Evaluating machine accuracy on ImageNet. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8634–8644. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/shankar20c.html.

[139] Shchur, O., Mumme, M., Bojchevski, A., and Günnemann, S. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

[140] Shift, E. P. U. U. D. Can you trust your model's uncertainty? 2019.

[141] Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556.

[142] Srivastava, M., Nushi, B., Kamar, E., Shah, S., and Horvitz, E. An empirical analysis of backward compatibility in machine learning systems. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3272–3280, 2020.

[143] Stanojević, M., Kamran, A., Koehn, P., and Bojar, O. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 256–273, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3031. URL https://www.aclweb.org/anthology/W15-3031.

[144] Stock, P. and Cisse, M. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 498–512, 2018.

[145] Suresh, H. and Guttag, J. V. A framework for understanding sources of harm throughout the machine learning life cycle. *arXiv preprint arXiv:1901.10002*, 2019.

[146] Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[147] Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *arXiv preprint arXiv:2007.00644*, 2020.

[148] Tatarchenko, M., Richter, S. R., Ranftl, R., Li, Z., Koltun, V., and Brox, T. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3405–3414, 2019.

[149] Teney, D., Kafle, K., Shrestha, R., Abbasnejad, E., Kanan, C., and Hengel, A. v. d. On the value of out-of-distribution testing: An example of goodhart's law. *arXiv preprint arXiv:2005.09241*, 2020.

[150] Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J. B., and Isola, P. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020.

[151] Toral, A., Castilho, S., Hu, K., and Way, A. Attaining the unattainable? reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432*, 2018.

[152] Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A., and Madry, A. From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*, pp. 9625–9635. PMLR, 2020.

[153] Tuggener, L., Schmidhuber, J., and Stadelmann, T. Is it enough to optimize cnn architectures on imagenet? *arXiv preprint arXiv:2103.09108*, 2021.

[154] Turner, R., Eriksson, D., McCourt, M., Kiili, J., Laaksonen, E., Xu, Z., and Guyon, I. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020, 2021.

[155] Wagstaff, K. Machine learning that matters. *arXiv preprint arXiv:1206.4656*, 2012.

[156] Wallingford, M., Kusupati, A., Alizadeh-Vahid, K., Walsman, A., Kembhavi, A., and Farhadi, A. In the wild: From ml models to pragmatic ml systems. *arXiv preprint arXiv:2007.02519*, 2020.

[157] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*, 2019a.

[158] Wang, H., Ge, S., Xing, E. P., and Lipton, Z. C. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. `https://arxiv.org/abs/1905.13549`.

[159] Wasserman, L. *All of statistics : a concise course in statistical inference*. 2010. `https://link.springer.com/book/10.1007/978-0-387-21736-9`.

[160] Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., and Wilson, J. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

[161] Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., and Weinberger, K. Simplifying graph convolutional networks. In *International conference on machine learning*, pp. 6861–6871. PMLR, 2019.

[162] Yadav, C. and Bottou, L. Cold case: The lost mnist digits. In *Advances in Neural Information Processing Systems*, 2019. `https://arxiv.org/abs/1905.10498`.

[163] Yang, W., Lu, K., Yang, P., and Lin, J. Critically examining the" neural hype" weak baselines and the additivity of effectiveness gains from neural ranking models. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pp. 1129–1132, 2019.

[164] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding, 2020.

[165] Yu, K., Sciuto, C., Jaggi, M., Musat, C., and Salzmann, M. Evaluating the search phase of neural architecture search. *arXiv preprint arXiv:1902.08142*, 2019.

[166] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

[167] Zhang, B., Rajan, R., Pineda, L., Lambert, N., Biedenkapp, A., Chua, K., Hutter, F., and Calandra, R. On the importance of hyperparameter optimization for model-based reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 4015–4023. PMLR, 2021.

[168] Zhang, J., Wang, Y., Molino, P., Li, L., and Ebert, D. S. Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE transactions on visualization and computer graphics*, 25(1):364–373, 2018.

[169] Zhang, J. M., Harman, M., Ma, L., and Liu, Y. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 2020.

[170] Zhang, M. and Toral, A. The effect of translationese in machine translation test sets. *arXiv preprint arXiv:1906.08069*, 2019.

[171] Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., and Artzi, Y. Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*, 2020.

[172] Zhou, S., Gordon, M. L., Krishna, R., Narcomey, A., Fei-Fei, L., and Bernstein, M. S. Hype: A benchmark for human eye perceptual evaluation of generative models. *arXiv preprint arXiv:1904.01121*, 2019.

[173] Zhou, X., Nie, Y., Tan, H., and Bansal, M. The curse of performance instability in analysis datasets: Consequences, source, and suggestions. *arXiv preprint arXiv:2004.13606*, 2020.

[174] Zrnic, T. and Hardt, M. Natural analysts in adaptive data analysis. In *International Conference on Machine Learning (ICML)*, 2019. https://arxiv.org/abs/1901.11143.