1 List of analysis papers and short summaries

Legend of Failure Modes

I	Implementation Variations
B	Baseline Issues
DI	Data, Internal validity issues (ie. methodological errors, etc.)
O	Adaptive Overfitting
DE	Data, External validity issues (ie. spurious correlations, data misalignment, etc.)
M	Metrics misalignmennt
H	Comparison to Human Performance
G	General critiques of benchmarking

1.1 NLP (Translation, Question answering, Natural language Inference)

DE *Can Small and Synthetic Benchmarks Drive Modeling Innovation? A Retrospective Study of Question Answering Modeling Approaches [39]* Explores whether synthetic benchmarks could have driven architectural modeling progress in natural language (instead of SQuAD) and finds agreement between the two types of benchmarks in multiple cases.

H *Putting human assessments of machine translation systems in order [40].* Authors identify the unawknowledged design decisions that bias the assessment of human annotators that use the relative ranking method to evaluate model performance on 25 translation tasks from the annual Workshop on Machine Translation (WMT) in 2010 and 2011. In particular, the order in which candidate translations are presented is shown to bias human judgement and thus evaluation outcomes.

H *Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation [77]* A prior claim that a Chinese to English machine translation system achieved human parity falls through when translationese is removed from the picture. Further, expert annotators, in this case, professional translators, are better able to tell between machine and human translations.

I *Do Transformer Modifications Transfer Across Implementations and Applications? [51]* The authors find that most proposed modifications to the transformer architecture do not significantly improve performance across a variety of benchmarks. They suggest this is because modifications are specific to implementations and applications, and fail to transfer beyond their original niche.

DE, O *The Effect of Natural Distribution Shift on Question Answering Models [48]* Explores a variety of naturally occuring distribution shifts for language models, such as collecting data from various online source domains, and finds these changes in the distribution can have a large impact on model performance. The authors also find no signs of overfitting from test set re-use on the popular SQuAD benchmark.

DE,DI *What Will it Take to Fix Benchmarking in Natural Language Understanding? [7].* Survey paper with natural language processing that asserts learning problems should be well constructed, have adequate statistical power, and be representative of the task they aim to solve.

M *Translationese in Machine Translation Evaluation [29].* The authors show that using "reverse"-direction sentences, which were translated from language A to language B but used for a B-to-A dataset, inflate human evaluation scores. They also examine prior claims of model-human parity and find evaluation problems such as not using a large enough test set; a re-evaluation suggests that the machine system was outperformed by humans.

DI *The Effect of Translationese in Machine Translation Test Sets [87].* The inclusion of translationese, a translation artifact, in machine translation test sets inflates human evaluation scores for machine translation systems, and in some cases changes rankings of models.

DI *BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance [45].* Training the same NLP model architecture

(BERT) over a hundred different random seeds obtains consistent performance on MNLI, a natural language inference (NLI) dataset, but widely varying generalization performance, as measured on HANS, an NLI dataset that tests for biases learned on MNLI.

DE *Probing Neural Network Comprehension of Natural Language Arguments [52].* Finds that language models trained to solve a reasoning comprehension task exploit statistical cues within the dataset to achieve high performance.

M *Re-evaluating the role of bleu in machine translation research [9].* The authors highlight two situations where the use of BLEU fails to distinguish between translations which a human could tell apart and would rate differently. They find low correlation between BLEU scores and human judgements of adequacy and fluency.

B *A call for clarity in reporting bleu scores [58].* The most commonly used automatic metric (as opposed to human evaluation) in machine translation, BLEU, is not reported consistently: some papers preprocess text before scoring, and there are many parameters used by BLEU that aren't reported. The paper proposes a standarized tool for BLEU to solve these problems.

M *BLEU might be Guilty but References are not Innocent [25].* The authors show that improving reference translations improves correlation of BLEU with human judgement.

DI *With Little Power Comes Great Responsibility [10].* This paper describes the influence of statistical power in NLP experimental design and how small dataset size in GLUE make it difficult to distinguish between statistical noise and meaningful model improvements.

M, G *Beyond Accuracy: Behavioral Testing of NLP models with CheckList [64].* The authors propose an alternative evaluation paradigm to benchmarks, instead focusing on specific tests for known or anticipated failure modes for broadly relevant linguistic capabilities.

DE *Learning the Difference that Makes a Difference with Counterfactually-Augmented Data [33].* Crowdsourced perturbations of two NLP datasets cause model performance to drop, while learning on the regular and perturbed data improves on both domains and generalization to new domains. The two datasets are IMDB, a sentiment classification dataset [43], and SNLI [8], a natural language inference dataset.

B *Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping [16].* Finds that small factors such as random seed variance can have a huge impact on BERT performance, which is up to 7% on downstream tasks in the case of random seed variance.

M *On The Evaluation of Machine Translation Systems Trained With Back-Translation [19].* The authors show that BLEU fails to capture human preferences for models trained with back-translation.

DE *Evaluating NLP Models via Contrast Sets [26].* Benchmarks fail to address how NLP models perform in specific cases. The authors propose "contrast sets", where experts manually perturb test data points in a semantically meaningful way, to identify whether models are still able to output the correct answer. SOTA models perform worse on these contrast sets.

DI *The Curse of Performance Instability in Analysis Datasets: Consequences, Source, and Suggestions [90].* Analysis datasets are similar to standard benchmarks but specifically designed to test a linguistic capability or known failure mode. The authors find that model performance on benchmarks between random seeds is stable, but performance on analysis dataset can vary widely.

B *On the State of the Art of Evaluation in Neural Language Models [47].* The authors compare Recurrent Highway Networks (RHNs) against Long Short-Term Memory networks (LSTMs). They find that prior work demonstrating RHN superiority over LSTMs allocated more compute to RHNs, and find similar or competitive performance for LSTMs once this is controlled for.

DE *Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference [46].* The Multi-genre Natural Language Inference dataset (MNLI) contains several examples of syntactic heuristics, where the answer can be predicted by following a simple rule, such as always predicting "contradiction" when the premise contains "not". The authors construct a new dataset HANS which contains examples that both satisfy and violate the heuristics, and show that SOTA models perform extremely badly (e.g. 100% to 0% accuracy) on the portion of HANS which violates the heuristics the models have learned from MNLI.

I *Revisiting Few-sample BERT Fine-tuning [88].* Many previous papers have proposed solutions for stable BERT finetuning. The authors find that instability is caued by a bug in the ADAM implementation, and that fixing this bug reduces the advantage of propose finetuning methods.

B *It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners [69].* By reformulating SuperGLUE tasks into cloze-style questions, small language models can be can also be fine-tuned to have performace better than GPT-3.

DE *Robustness Gym: Unifying the NLP Evaluation Landscape [27].* Presents software developer tools that cover a range of different NLP metrics, datasets, etc. in order to help practicioners evaluate their models in various conditions.

G, M *Utility is in the Eye of the User: A Critique of NLP Leaderboards [21].* Authors highlight that leaderboards fail to measure (i.e., don't come with metrics for) factors beyond model performance, such as model size or inference speed or environmental impact.

DE *How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks [32].* Reading comprehension benchmarks require models to pick the answer to a question given a passage. Models do well on reading comprehension benchmarks even when the passage or question is withheld, suggesting that the datasets are poorly constructed because knowledge of the passage or question is irrelevant.

1.2 Computer vision (Image classification, object detection, Medical,General-purpose benchmarks, 3D shape reconstruction)

DE *AI for radiographic COVID-19 detection selects shortcuts over signal [14].* Finds that models trained for COVID-19 detection through radiographs exploit spurious correlates that do not hold when deployed in different environments.

DE *Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging [54].* Finds that models trained for detecting a pneumothorax from chest x-rays latch onto obvious dataset-level heuristics such as the presence of a chest drain instead of adequately solving the task.

H *Evaluating Machine Accuracy on ImageNet [70].* Humans are trained through documentation guidance and practice to classify objects in ImageNet and achieve comparable accuracy to modern machine learning models, though experience significantly less of a performance drop than models due to distribution shift. These labelers are about 3% to 9% better than the human performance levels reported from early 2015, indicating the variability of human baselines. Top-1 accuracy (a more natural task for humans) is almost perfectly linearly correlated with multi-label accuracy for the evaluated models, but humans fail more often for fine-grained categories (eg. differentiating dog breeds) while models fail more evenly across label categories.

B *Rethinking Few-Shot Image Classification: a Good Embedding Is All You Need? [76].* Finds that the simple baseline of training a linear model on top of a supervised classifier in the context of meta-learning tasks can outperform a variety of previous meta-learning appraoches such as MAML.

M *Are we done with ImageNet?[3]*
The authors collect multi-label annotations for ImageNet via a modified crowdsourcing process. The results show a slightly plateauing trend, indicating that models may have overfit to specifics of the ImageNet distribution.

DE *Measuring Robustness to Natural Distribution Shifts in Image Classification [73].* Finds that models trained on ImageNet fail to generalize well to other distributions with shifts in object pose, lighting, object composition, etc.

DE *In a forward direction: Analyzing distribution shifts in machine translation test sets over time [38].* A fixed machine translation model scores better on newer machine translation test sets than older test sets (the Workshop on Machine Translation releases a new test set every year). The observed increase in scores for any single model is attributable to changes made in dataset construction which progressively removed translationese, a problematic translation artifact.

DE *Transfusion: Understanding Transfer Learning for Medical Imaging [59].* This study looks at 2 medical image datasets: diabetic retinopathy prediction and chest x-ray prediction. They compare a few deep learning models and find that imagenet pretraining doesn't really help performance on these downstream datasets.

DE *CheXtransfer: Performance and Parameter Efficiency of ImageNet Models for Chest X-Ray Interpretation [34]*
The paper compares ImageNet performance of several CNN architectures to performance on X-ray classification. The authors find that X-ray classification performance has plateaued as a function of ImageNet performance, but ImageNet pre-training still helps on the X-ray dataset.

DE *Do Better ImageNet Models Transfer Better? [35]*
The authors evaluate ImageNet models on twelve other image classification datasets and find that better ImageNet models also perform better on the other datasetes, especially when the models are pre-trained on ImageNet.

DE *Is it Enough to Optimize CNN Architectures on ImageNet? [79].* The authors train 500 ImageNet architectures on 8 other image classification datasets from different domains and find that the correlation between ImageNet performance and dowstream dataset performance varies wildly, with even negative correlations for some.

DE *Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study [86].* Finds that models trained to diagnose pneumonia in chest radiographs in one hospital fail to generalize well to other hospital due to differences in data collection, equipment, patient populations, etc.

O *Does ImageNet Generalize to ImageNet? [62]*
The authors construct a new test set for ImageNet and find that overfitting from test set re-use did not occur despite a decade of competitive testing on this dataset. Instead, distribution shift led to a substantial drop in accuracy.

O *Does CIFAR-10 Generalize to CIFAR-10? [61]*
The authors construct a new test set for CIFAR-10 and find that overfitting from test set re-use did not occur despite a decade of competitive testing on this dataset. Instead, distribution shift led to a substantial drop in accuracy.

O *Cold Case: The Lost MNIST Digits* [83]
The authors construct a new test set for MNIST and find that overfitting from test set re-use did not occur despite two decades of competitive testing on this dataset.

B *A Baseline for Few-shot Image Classification [15].* Finds that a simple transductively-tuned baseline can outperform all more complex methods (MAML, MetaOpt, etc.) on few-shot learning tasks when controlling for all other factors of variation.

B *What Do Single-view 3D Reconstruction Networks Learn? [74].* Finds that simple baselines such as clustering and retrieval on top of the pretrained embedding space outperform recent deep methods for 3D reconstruction.

I *On Buggy Resizing Libraries and Surprising Subtleties in FID Calculation [56].* The widely used Frechet Inception Distance (FID) metric for evaluating generative models is not consistently reported. Differences between image processing libraries and choices in implementing FID cause meaningful differences in scores.

DE *From ImageNet to Image Classification: Contextualizing Progress on Benchmarks [78].* The authors provide multi-label annotations for ImageNet via a modified crowdsourcing process and study the impact of images with multiple labels on ImageNet accuracy metrics.

B *Overfitting in adversarially robust deep learning [65].* The paper shows that early stopping combined with a simple loss function is competitive with more complicated loss functions that were proposed for adversarially robust image classification.

DI *Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks [53].* Authors revealed significant label errors in mainstream datasets, such as an average error rate of 3.4% across the reviewed 10 datasets, including 6% of the ImageNet validation set.

1.3 Meta-learning / Architecture search

B *Random search and reproducibility for neural architecture search [36].* Given the same computational budget, random search with minor modifications (e.g. early stopping) outperforms state of the art neural architecture search methods.

B *Evaluating the Search Phase of Neural Architecture Search [85].* Finds that random search within the penn treebank and cifar10 dataset search spaces leads to similar performance as leading neural architecture search algorithms when given equal compute.

1.4 Generative models (GANs, generative language models)

M *HYPE: A Benchmark for Human eYe Perceptual Evaluation of Generative Models [89].* This paper introduces a human benchmark for evaluation of generative models, which scores if a human can tell a real image vs fake. The authors found that HYPE scores were not correlated with commonly used automated metrics such as FID.

B *Are GANs Created Equal? A Large-Scale Study [42].* Evaluates many GAN losses, fixing the backbone architecture, dataset, and other training details, and finds that most GAN models can reach similar performance given equal compute budget.

1.5 Optimization for deep learning

B *Hyperband: a novel bandit-based approach to hyperparameter optimization [37].* Finds that random search combined with early stopping outperforms more sophisticated Bayesian hyper-parameter optimization methods.

I *Decoupled Weight Decay Regularization [41].* The authors point out that L_2 regularization is distinct from weight decay regularization for adaptive gradient algorithms like Adam, even though the former is often substituted for the latter. They show that implementing weight decay regularization improves Adam's generalization performance.

B *A Large Batch Optimizer Reality Check: Traditional, Generic Optimizers Suffice Across Batch Sizes [50].* LARS and LAMB optimizers are designed to increase the speed of model training given large batch sizes. Traditional optimizers like Nesterov momentum and Adam perform comparably at large batch sizes, signifying that such interventions are not significant improvements when compared to an adequate baseline.

1.6 Learning on graphs

B *Combining Label Propagation and Simple Models Out-performs Graph Neural Networks [31].* Finds that incorporating a graph label propagation step with simple models outperforms more recent deep graph neural networks.

B *Benchmarking Graph Neural Networks [18].* This paper documents common pitfalls and problems in benchmarking graph neural networks.

B *Simplifying Graph Convolutional Networks [82].* Finds that a simple graph preprocessing step with an adjacency matrix combined with logistic regression outperforms more recent deep graph neural networks.

B *Pitfalls of Graph Neural Network Evaluation [71].* Graph neural network papers (GNN) fail to control for relevant factors when making comparisons. The authors of this paper attempt to evaluate four GNN architectures while controlling for everything except architectures: keeping the optimizers, initialization methods, compute budget, etc, the same. Performance turns out to be similar between different GNNs.

1.7 Tabular data & classical methods (ie. Medical (MIMIC))

B *Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? [23]* When evaluating the performance of 179 classifiers on the whole UCI repository (121 data sets) [17], authors found that random forest classifiers outperformed any other type, with these methods achieving over 90% accuracy in 84.3% of the data sets.

B *Evaluating Progress on Machine Learning for Longitudinal Electronic Healthcare Data [2].* Finds that on tabular data prediction tasks found on MIMIC-III, simple logistic regression achieves comparable performance to more sophisticated methods developed over the past three years.

1.8 Reinforcement Learning

DE, B *What Matters In On-Policy Reinforcement Learning? A Large-Scale Empirical Study [1].* Implements many RL algorithms and more than 50 code-level tricks and optimizations to consistently benchmark performance; one surprising finding is that policy initialization scheme plays a huge role in policy performance.

I *Implementation Matters in Deep Policy Gradients: A Case Study on PPO and TRPO [20].* The authors compare two deep policy gradient algorithms, Proximal Policy Optimization (PPO) and Trusted Region Policy Optimization (TRPO). They find that "code level optimizations", algorithmic modification described as auxillary details or undescribed altogether, are responsible for most of PPO's performance gain over TRPO and significantly affect algorithmic behaviour.

B *Deep reinforcement learning that matters [30].* Evaluation of reinforcement learning (RL) algorithms suffers from several issues: varying the random seed varies algorithm performance enough to change performance rankings; many under-reported hyperparameters greatly affect algorithm performance; different implementations of the same algorithm perform differently.

B *Simple random search provides a competitive approach to reinforcement learning [44].* A lightweight modification of random search achieves similar reward as SOTA reinforcement learning methods on MuJoCo Gym tasks while requiring fewer samples.

B *Towards Generalization and Simplicity in Continuous Control [60].* Simple methods using policies with linear and RBF parameterizations can solve many continuous control benchmarks, including MuJoCo Gym tasks. Further, the authors highlight that policies learned on the benchmarks are trajectory-centric: when these policies are perturbed, they fail to recover.

1.9 Visual question answering

DE *Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering [28].* Finds that original VQA dataset is not balanced in terms of label distribution for certain questions, making achieving high performance relatively easy.

DE, T *On the Value of Out-of-Distribution Testing: An Example of Goodhart's Law [75].* Critiques the use of VQA-CP as a valid OOD dataset for VQA tasks, since VQA-CP inverts the VQA label distribution, and many robust methods explicitly rely on this fact.

1.10 Information retrieval

B *Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models [84].* Examines many information-retrieval papers from 2005-2019, and find that no approach (both neural or non-neural) comes close to the 2004 best.

1.11 Metric Learning

B *A Metric Learning Reality Check [49].* Authors benchmark several deep metric learning algorithms on three datasets under identical training conditions and find that papers have drastically overstated improvements over classic methods.

B *Unbiased Evaluation of Deep Metric Learning Algorithms [22].* Authors benchmark several deep metric learning algorithms on three datasets under identical training conditions and find that older methods perform significantly better than previously believed.

B *Revisiting Training Strategies and Generalization Performance in Deep Metric Learning [67].* Authors benchmark several deep metric learning algorithms on three datasets under identical training conditions and find that generally, performance between criteria is much more similar than literature indicates.

1.12 Recommender Systems

B *A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research [11].* This is a recommender systems reproducibility experiment, where they compare the proposed methods to a range of baselines on the datasets the original papers used; 11 of the 12 methods were outperformed by simple baselines on the datasets the respective paper had identified.

B *On the Difficulty of Evaluating Baselines: A Study on Recommender Systems[63].* This is a recommender systems reproducibility experiment on the MovieLens-10M benchmark; finds that a well-tuned vanilla matrix factorization baseline significantly outperforms more recent methods reported in the literature.

1.13 Semi-supervised /Unsupervised representation learning

B *Realistic Evaluation of Deep Semi-Supervised Learning Algorithms [55].* This work is a standardized evaluation of semi-supervised learning algorithms on SVHN and CIFAR-10; they find that prior work underestimated the performance of fully supervised learning in the small-n regime and that ImageNet pre-training + fine-tuning with few samples does better than any of the semi-supervised methods they benchmarked.

1.14 General / other

DE *Machine Learning that Matters [80].* A scientist hoping to use machine learning for practical applications gets frustrated with the inadequate quality of the UCI repository [17] and the benchmark culture it perpetuates. She advocates instead for a more systems-level perspective on machine learning development and evaluation.

H *Performance vs. competence in human–machine comparisons [24].* There is a difference between possessing human-level ability (“competence”), and the superficial demonstration of a skill (“performance”). Many models perform better than human counterparts on a given learning problem but do not achieve this performance in a human-like way, and thus fail to demonstrate competence when tested for that skill outside the scope of the initial learning problem.

DE,DI *A Flawed Dataset for Symbolic Equation Verification [13].* A synthetic dataset for equation verification is heavily critiqued for the lack of rigor in how it is generated, the correctness of the axioms presented and the relevance of the task represented.

DI *“Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI [68].* Authors interview 53 ML practitioners in 6 countries and conclude that data work remains under-valued as a research topic of interest, even though data labelling consists of 25-60% of the cost of model development. They identify that data issues compound on each other in “data cascades”, contributing to critical failures in model deployment within high stakes scenarios.

M *Accounting for variance in machine learning benchmarks [6].* Several sources of variation in the dataset and implementation of machine learning models can obscure our understanding of their performance (eg. data sampling, data augmentation, parameter initialization, and hyperparameters choices). This paper recommends randomization and more robust trial reporting in order to appropriately and consistently address these issues.

G *Pitfalls in Machine Learning Research: Reexamining the Development Cycle [4].* This paper comments on the challenges throughout the model development lifecycle that contributes to failures in machine learning deployment. Algorithmic design, data collection, and evaluation practices are named as concrete areas of concern - authors recommend interventions such as third party assessment, statistical testing and data audits.

B *Can You Trust Your Model’s Uncertainty? [72]* This survey of uncertainty estimation methods shows that ensemble methods consistently outperform the rest.

O *The Ladder: A Reliable Leaderboard for Machine Learning Competitions [5]*
The paper introduces a specific attack on competition leaderboards demonstrating that overfitting from test set re-use is easily possible. The paper also describes a mechanism to protect from overfitting.

O *A Meta-Analysis of Overfitting in Machine Learning [66]*
The authors survey more than 100 classification competitions on Kaggle and find little to no overfitting from test set re-use.

DE *Underspecification Presents Challenges for Credibility in Modern Machine Learning [12].* Machine learning models that are identically trained and developed fail in different ways once deployed - this is partially due to the "underspecification" of the learning problem, in which features of the problem in the training domain are unaccounted for with respect to their influence on performance in the deployment domain.

DE,B *In the wild: From ml models to pragmatic ml systems [81]* Implicit assumptions in the experimental setup of few-shot and continual learning tasks obscure a clear understanding of performance measurement. The FLUID framework re-introduces certain experimental design considerations that need to be explicitly designed for real world model deployment.

DE,DI *Data and it's (dis)contents [57].* The culture around dataset development, use, and distribution demonstrates a lack of cautious attention paid to this critical aspect of broader machine learning development.

References

[1] 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.

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

[3] 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.

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

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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`.

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

[16] 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.

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

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

[19] 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.

[20] 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.

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

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

[23] 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.

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

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

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

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

[39] 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.

[40] 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.

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

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

[48] 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.

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

[50] 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.

[51] 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.

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

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

[54] 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.

[55] 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.

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

[57] 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.

[58] 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.

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

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

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

[62] 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.

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

[64] 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.

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

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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`.

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

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

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

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

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

[81] 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.

[82] 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.

[83] 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`.

[84] 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.

[85] 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.

[86] 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.

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

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

[89] 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.

[90] 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.