EXPERIMENTAL STANDARDS FOR DEEP LEARNING RESEARCH: A NATURAL LANGUAGE PROCESSING PERSPECTIVE

Dennis Ulmer[♣] **Elisa Bassignana**[♣] **Max Müller-Eberstein**[♣] **Daniel Varab**[♣]
Mike Zhang[♣] **Christian Hardmeier**[♣] **Barbara Plank**[♣♢]
[♣]Department of Computer Science, IT University of Copenhagen, Denmark
[♢]Center for Information and Language Processing (CIS), LMU Munich, Germany
dennis.ulmer@mailbox.org

ABSTRACT

The field of Deep Learning (DL) has undergone explosive growth during the last decade, with a substantial impact on Natural Language Processing (NLP) as well. Yet, as with other fields employing DL techniques, there has been a lack of common experimental standards compared to more established disciplines. Starting from fundamental scientific principles, we distill ongoing discussions on experimental standards in DL into a single, widely-applicable methodology. Following these best practices is crucial to strengthening experimental evidence, improve reproducibility and enable scientific progress. These standards are further collected in a public repository to help them transparently adapt to future needs.

1 INTRODUCTION

The field of Artificial Intelligence (AI) and Machine Learning (ML) has seen immense growth over the span of the last 20 years. Figure 2 shows how, according to Zhang et al. (2021), the number of peer-reviewed papers has increased twelve-fold compared to the year 2000. At the same time, interest in Deep Learning (DL) has increased substantially as well, demonstrated via Google Trends in the same figure. While such progress is remarkable, rapid growth comes at a cost: Akin to concerns in other disciplines (John et al., 2012; Jensen et al., 2021), several authors have noted issues with reproducibility (Gundersen & Kjensmo, 2018; Belz et al., 2021) and a lack of significance testing (Marie et al., 2021) or published results not carrying over to different experimental setups, for instance in NLP (Narang et al., 2021; Gehrmann et al., 2022), Reinforcement Learning (Henderson et al., 2018; Agarwal et al., 2021), and optimization (Schmidt et al., 2021a). Others have questioned commonly-accepted procedures (Gorman & Bedrick, 2019; Søgaard et al., 2021; Bouthillier et al., 2021; van der Goot, 2021) as well as the (negative) impacts of research on society (Hovy & Spruit, 2016; Mohamed et al., 2020; Bender et al., 2021; Birhane et al., 2021) and environment (Strubell et al., 2019; Schwartz et al., 2020; Henderson et al., 2020). These problems have not gone unnoticed—many of the works mentioned here have proposed a cornucopia of solutions. In a quickly-moving publication environment however, keeping track and implementing these proposals becomes challenging. In this work, we weave many of them together into a cohesive methodology for gathering stronger experimental evidence, that can be implemented with reasonable effort.

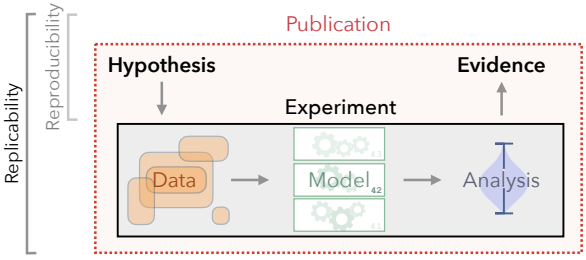

Figure 1: **Visualization of the Scientific Process in Deep Learning.** Uncertainty is introduced at each experimental step, influencing the resulting evidence as well as the documentation required for either reproducibility or replicability.

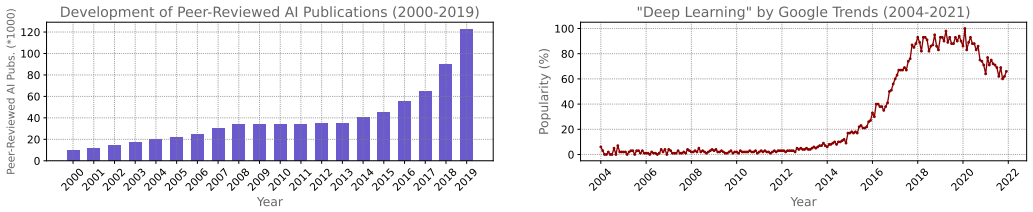

Figure 2: **Development of AI and DL.** Shown is the development of AI and DL measured by the number of peer-reviewed publications between 2000-2019 Zhang et al. (2021) on the left and relative interest measured by the search engine Google between 2004-2021 Google Trends (2022) on the right. Both experience an explosive growth around 2014.

Based on the scientific method (Section 2), we divide the empirical research process—obtaining evidence from data via modeling—into four steps, which are depicted in Figure 1: *Data* (Section 3), including dataset creation and usage, *Codebase & Models* (Section 4), *Experiments & Analysis* (Section 5) and *Publication* (Section 6). For each step, we survey contemporary findings and summarize them into actionable practices for empirical research. While written mostly from the perspective of NLP researchers, we expect many of these insights will be useful for practitioners of other adjacent sub-fields of ML and DL.

Contributions ① We survey and summarize a wide array of proposals regarding the improvement of the experimental (and publishing) pipeline into a single accessible methodology applicable for a wide and diverse readership. At the end of every section, we provide a summary with the most important points, marked with $\diamond$ to indicate that they should be seen as a minimal requirement to ensure reproducibility, and $\star$ for additional recommended actions. ② We create, point to, or supply useful resources to support everyday research activities and improve reproducibility in the field. We furthermore provide examples and case studies illustrating these methods in Appendix A. We also provide an additional list of resources in Appendix C. The same collection as well as checklists derived from the actionable points at the end of sections are also maintained in an open-source repository,[1] and we invite the research community to discuss, modify and extend these resources. ③ We discuss current trends and their implications, hoping to initiate a more widespread conversation about them in the ML community to facilitate common standards and improve the quality of research in general.

2 Preliminaries

In order for our proposed methodology to remain as broadly applicable as possible, it must be built on the scientific principles for generating strong evidence for the general advancement of knowledge. We therefore define terms which are crucial to this process.

The Scientific Method In science – and ML – knowledge can be obtained through several ways, for instance theory building, qualitative methods, and empirical research (Kuhn, 1970; Simon, 1995). For our purposes, we focus on the latter aspect, in which (exploratory) analyses lead to falsifiable hypotheses that can be tested and iterated upon (Popper, 1934).[2] This process requires that *anyone* must be able to back or dispute these hypotheses in the light of new evidence.

In the following, we focus on the evaluation of hypotheses as well as how to ensure the *replicability* and *reproducibility* of the experiments which gave rise to the original empirical evidence. In computational literature, one term requires access to the original code in order to re-run experiments exactly, while the other requires sufficient information in order to reproduce the original findings even in the absence of code and original data. Strikingly, these central terms already lack agreed-upon definitions, mainly regarding which term defines which level of detail (Peng, 2011; Fokkens

[1] `https://github.com/Kaleidophon/experimental-standards-deep-learning-research`

[2] We also see that such hypothesis-driven science is not always applicable or even possible (Carroll, 2019). Nevertheless, it creates a strong common denominator that encompasses most kinds of empirical ML research.

et al., 2013; Liberman, 2015; Cohen et al., 2018). As the underlying ideas are equivalent, we follow the prevailing definitions in the DL and NLP communities in the following sections (Drummond, 2009; Dodge & Smith, 2020).

Replicability Within DL, we take replicability to mean the exact replication of prior reported evidence. In a computational experimental environment, access to the same data, code and tooling should be sufficient to generate prior results, however many environmental factors, such as hardware differences, make it difficult to achieve exact replication in practice. Nonetheless, we regard experiments to be replicable if a practitioner is able to re-run them to produce the same evidence within a small margin of error dependent on the environment, without the need to approximate or guess experimental details. All controllable factors in an experiment, including data, models and code, must therefore be readily available without the need for re-implementation.

Reproducibility In contrast, we take reproducibility to mean the availability of all necessary and sufficient information such that an experiment's findings can independently be reaffirmed when the same research question is asked. As discussed later, the availability of all components as required for replicability is rare—even in a computational setting. Findings generated this way are nonetheless valuable for the scientific community if their underlying hypotheses can be evaluated by anyone with access to the publication. An experiment then is reproducible if enough information is provided to find the original evidence even without the tooling for replicating a metric's exact value (e.g., values differ, but follow the same patterns across experiments with an equivalent setup to the original).

We assume that the practitioner aims to follow these principles in order to find answers to a well-motivated research question by gathering the strongest possible evidence for or against their hypotheses. The following methods therefore aim to reduce uncertainty in each step of the experimental pipeline in order to ensure reproducibility and/or replicability (visualized in Figure 1).

3 DATA

Frequently, it is claimed that a model solves a particular cognitive task, however in reality it merely scores higher than others on some specific dataset according to some predefined metric (Schlangen, 2021). Of course, the broader goal is to improve systems more generally by using individual datasets as proxies. Admitting that our experiments cover only a small slice of the real-world sample space will help more transparently measure progress towards this goal. In light of these limitations and as there will always be private or otherwise unavailable datasets which violate replicability, a practitioner must ask themselves: *Which key information about the data must be known in order to reproduce an experiment's findings?* In this section we define requirements for putting this question into practice during dataset creation and usage such that anyone can draw the appropriate conclusions from a published experiment.

Choice of Dataset The choice of dataset will arise from the need to answer a specific research question within the limits of the available resources. Such answers typically come in the form of comparisons between different experimental setups while using the equivalent data and evaluation metrics. Using a publicly available, well-documented dataset will likely yield more comparable work, and thus stronger evidence. In absence of public data, creating a new dataset according to guidelines which closely follow prior work can also allow for useful comparisons. Should the research question be entirely unexplored, creating a new dataset will be necessary. In any case, the data itself must contain the information necessary to generate evidence for the researcher's hypothesis. For example, a model for a classification task will not be learnable unless there are distinguishing characteristics between data points and consistent labels for evaluation. Therefore, an exploratory data analysis is recommended for assessing data quality and anticipating problems with the research setup. Simple baseline methods such as regression analyses or simply manually verifying random samples of the data may provide indications regarding the suitability and difficulty of the task and associated dataset (Caswell et al., 2021).

Metadata At a higher level, data sheets and statements (Gebru et al., 2020; Bender & Friedman, 2018) aim to standardize metadata for dataset authorship in order to inform future users about assumptions and potential biases during data collection and annotation. Such documentation is im-

portant, as biases can be introduced at all levels, including the research design Hovy & Prabhumoye (2021). Simultaneously, they encourage reflection by authors on whether their process adheres to their own predetermined guidelines (Waseem et al., 2021). Generally, higher-level documentation should aim to capture the dataset's *representativeness* with respect to the global population. This is especially crucial for "high-stakes" environments in which subpopulations may be disadvantaged due to biases during data collection and annotation (He et al., 2019). Even in lower-stake scenarios a model trained on only a subset of the global data distribution can have inconsistent behaviour when applied to a different target data distribution (D'Amour et al., 2020; Koh et al., 2020). For instance, language or even domain differences can have a noticeable impact on model performance (White & Cotterell, 2021; Ramesh Kashyap et al., 2021). Increased data diversity can improve the generalization ability of models to new domains, e.g., more languages in NLP (Benjamin, 2018). However, diversity itself can be difficult to quantify (Gong et al., 2019) and global data coverage is likely unachievable, highlighting the importance of documenting representativeness through metadata in order to ensure reproducibility—even in absence of the original data. For replicability using the original data, further considerations include long-term storage in addition to versioning, as to ensure equal comparisons in future work (see Appendix A.1 for case studies).

Instance Annotation At the level of data instances, the most important aim is high data quality. This entails both the labeling and the collection process as it is not enough to supply a large amount of data, and expect an ML algorithm to learn desired characteristics—the data must be accurate and relevant for the task to enable effective learning (Pustejovsky & Stubbs, 2012; Tseng et al., 2020) and reliable evaluation Bowman & Dahl (2021); Basile et al. (2021). Since most datasets involve human annotation, a careful annotation design is crucial Pustejovsky & Stubbs (2012); Paun et al. (2022). In NLP, language ambiguity poses inherent challenges and disagreement is genuine Basile et al. (2021); Specia (2021); Uma et al. (2021). As such insights into the annotation process are valuable, yet often inaccessible, we recommend to release datasets with raw annotations prior to aggregation into categorical labels, and complement data with insights like statistics on inter-annotator coding Paun et al. (2022), e.g., over time (Braggaar & van der Goot, 2021). When creating new datasets such statistics strengthen the reproducibility of future findings, as they transparently communicate the inherent variability in the world instead of obscuring it.

Pre-processing The goal of a dataset is to allow for quantitative comparisons of different hypotheses. Given a well-constructed or well-chosen dataset, the first step of an experimental setup will be the process by which a model takes in the data. This must be well documented or replicated—most easily by publishing the associated code—as perceivably tiny pre-processing choices can lead to huge accuracy discrepancies (Pedersen, 2008; Fokkens et al., 2013). In NLP, this primarily involves decisions such as sentence segmentation, tokenization and normalization. In general, the data setup pipeline should ensure that a model "observes" the same kind of data across comparisons. Next, the dataset must be split into representative subsamples which should only be used for their intended purpose, i.e., model training, tuning and evaluation (see Section 5). In order to support claims about the generality of the results, it is necessary to use a test split without overlap with other splits. Alternatively, a tuning/test set could consist of data that is completely foreign to the original dataset (Ye et al., 2021), ideally even multiple sets Bouthillier et al. (2021). It should be noted that even separate test splits are prone to overfitting if they have been in use for a longer period of time, as more people aim to beat a particular benchmark (Gorman & Bedrick, 2019). If a large variety of resources are not available, it is also possible to construct challenging test sets from existing data (Ribeiro et al., 2020; Kiela et al., 2021; Søgaard et al., 2021). Finally, the metrics by which models are evaluated should be consistent across experiments and thus benefit from standardized evaluation code (Dehghani et al., 2021). For some tasks, metrics may be driven by community standards and are well-defined (e.g., classification accuracy). In other cases, approximations must stand in for human judgment (e.g., in machine translation). In either case—but especially in the latter—dataset authors should inform users about desirable performance characteristics and recommended metrics.

Appropriate Conclusions The results a model achieves on a given data setup should first and foremost be taken as just that. Appropriate, broader conclusions can be drawn using this evidence provided that biases or incompleteness of the data are addressed (e.g., results only being applicable to a subpopulation). Even with statistical tests for the significance of comparisons, properties such as the size of the dataset and the distributional characteristics of the evaluation metric may influence

the statistical power of any evidence gained from experiments (Card et al., 2020). It is therefore important to keep in mind that in order to claim the reliability of the obtained evidence, for example, larger performance differences are necessary on less data than what might suffice for a large dataset, or across multiple comparisons (see Section 5). Finally, a practitioner should be aware that a model's ability to achieve high scores on a certain dataset may not be directly attributable to its capability of simulating a cognitive ability, but rather due to spurious correlations in the input (Ilyas et al., 2019; Schlangen, 2021; Nagarajan et al., 2021). By for instance only exposing models to a subset of features that should be inadequate to solve the task, we can sometimes detect when they take unexpected shortcuts (Fokkens et al., 2013; Zhou et al., 2015). Communicating the limits of the data helps future work in reproducing prior findings more accurately.

Best Practices: **Data**

- ◇ Consider dataset and experiment limitations when drawing conclusions (Schlangen, 2021);
- ◇ Document task adequacy, representativeness and pre-processing (Bender & Friedman, 2018);
- ◇ Split the data such as to avoid spurious correlations (Gorman & Bedrick, 2019);
- ⋆ Perform exploratory data analyses to ensure task adequacy (Caswell et al., 2021);
- ⋆ Publish the dataset accessibly & indicate changes;
- ⋆ Claim significance considering the dataset's statistical power (Card et al., 2020).

4 Codebase & Models

The ML and NLP community has historically taken pride in promoting open access to papers, data, code, and documentation, but some authors have also noted room for improvement Wieling et al. (2018); Belz et al. (2020). A common practice has been to open-source all components of the experimental procedure in a repository (e.g., `git`, or a simple `zip` file). We, as a community, expect such a repository to contain model implementations, pre-processing code, evaluation scripts, and detailed documentation on how to obtain claimed results using these components. It is important to note that the benefit of such a repository is its ability to enable *replication*. In particular, a comprehensive code base directly enables replicability (generating prior reported evidence with the exact same environment and code), while an incomplete code base pushes research towards reproducibility (i.e., needing to re-implement code to obtain similar results). In short: *The more we document our methodology, the better.* In practice, such documentation is often communicated through a `README` file, in which user-oriented information such as hardware assumptions, software prerequisites, instructions, help, or descriptions of how to run the software are described. In Appendix B, we propose minimal requirements for such a `README` file and give pointers on files and code structure. In DL, such data can be large and impractical to share. However, because results rely heavily on data, it is essential to carefully consider how one can share the data with researchers in the future. Repositories for long-term data storage backed by public institutions should be preferred (e.g., LINDAT/CLARIN by Váradi et al., 2008, more examples in Appendix C). Yet more than often, practitioners cannot distribute data due to privacy, legal, or storage reasons. In such cases, practitioners must instead carefully consider how to distribute data and tools to allow future research to produce accurate replications of the original data Zong et al. (2020).

Hyperparameter Search Usually, a common part of the ML pipeline is to perform some sort of hyperparameter search. The corresponding tuning strategies still remains an open area of research (see Bischl et al., 2021 for a comprehensive overview), but the following rules of thumb exist: If there are very few parameters that can be searched exhaustively under the computation budget, grid search or Bayesian optimization can be applied. Otherwise, random search is preferred, as it explores the search space more efficiently (Bergstra & Bengio, 2012). More advanced methods like Bayesian Optimization (Snoek et al., 2012) and bandit search-based approaches (Li et al., 2017) can be used as well if applicable (Bischl et al., 2021). In any case, the following information should be reported: Hyperparameters that were searched per model including their options and ranges, the final hyperparameter settings used, number of trials, settings of the search procedure if applicable. As tuning of hyperparameters is typically performed using specific parts of the dataset, it is important

to note that any modeling decisions based on them automatically invalidates their use as *test* data, since reported results are not unseen anymore.

Models The recent surge of large Transformer-based models has had a large impact on DL and NLP Vaswani et al. (2017); Devlin et al. (2019); Dosovitskiy et al. (2021); Chen et al. (2021). These and many other contemporary models, however, have very large computational and memory footprints. To avoid retraining models, and more importantly, to allow for replicability, it is recommended to save and share model weights. This may face similar challenges as those of datasets (namely, large file sizes), but it remains an impactful consideration. In most cases, simply sharing the best or most interesting model could suffice. It should be emphasized that distributing model weights should always complement a well-documented repository as libraries and hosting sites might not be supported in the future.

Model Evaluation With respect to models and tasks, the exact evaluation procedure can differ greatly. It is important to either reference the exact evaluation script used (including parameters, citation and version, if applicable) or at least include the evaluation script in the code base. Moreover, to ease error or post-hoc analyses, we highly recommend saving model predictions in separate files whenever possible, and making them available at publication Card et al. (2020); Gehrmann et al. (2022). This could for instance be done using plain `.txt` or `.csv` files.

Model Cards Apart from quantitative evaluation and optimal hyperparameters, Mitchell et al. (2019) propose model cards: A type of standardized documentation, as a step towards responsible ML and AI technology, accompanying trained ML models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic and intersectional groups that are relevant to the intended application domains. They can be reported in the paper or project. For example, we refer to Mitchell et al. (2019); Menon et al. (2020) that actively used a model card.

Best Practices: **Codebase & Models**

- ◇ Publish a code repository with documentation and licensing to distribute for replicability;
- ◇ Report all details about hyperparameter search and model training;
- ◇ Specify the hyperparameters for replicability;
- ⋆ Use model cards;
- ⋆ Publish models, predictions and evaluation scripts.

5 Experiments & Analysis

Experiments and their analyses constitute the core of most scientific works, and empirical evidence is valued especially highly in ML research (Birhane et al., 2021). As such, special care should be put into designing and executing them. We outlined in the introduction how issues with replicability and significance of results in the ML literature have been raised by several authors (Gundersen & Kjensmo, 2018; Henderson et al., 2018; Narang et al., 2021; Schmidt et al., 2021a). Therefore, we discuss the most common issues and counter-strategies at different stages of an experiment.

Model Training For model training, it is advisable to set a random seed for replicability, and train multiple initializations per model in order to obtain a sufficient sample size for later statistical tests. Commonly used values are three to five runs, however this should be adapted based on the observed variance: Using bootstrap power analysis, existing model scores are raised by a constant compared to the original sample using a significance test in a bootstrapping procedure (Yuan & Hayashi, 2003; Tufféry, 2011; Henderson et al., 2018). If the percentage of significant results is low, we should collect more scores.[3] Bouthillier et al. (2021) further recommend to vary as many sources of randomness in the training procedure as possible (i.e., data shuffling, data splits etc.) to obtain a

[3]We are aware that this poses some tension with the hardware requirements of many modern DL architectures, which is why we dedicate part of the discussion in Section 7 to this question.

closer approximation of the true model performance. Nevertheless, the question of what conclusions can be drawn from these outcomes can be harder than it might superficially seem, precisely due to the mentioned sources of statistical uncertainty. A common solution is the use of statistical hypothesis testing, which we portray here along with criticisms and alternatives.

Significance Testing Especially with deep neural networks, even with a fixed set of hyperparameters, performance can be influenced by a number of (stochastic) factors such as the random seed (Dror et al., 2019) or even the choice of hardware or framework (Leventi-Peetz & Östreich, 2022). As such, multiple factors have to be taken into account when drawing conclusions from experimental results. First of all, the size of the dataset should support sufficiently powered statistical analyses (see Section 3). Secondly, an appropriate significance test should be chosen. We give a few rules of thumb based on Dror et al. (2018): When the distribution of scores is known, for instance a normal distribution for the Student's t-test, a *parametric* test should be chosen. Parametric tests are designed with a specific distribution for the test statistic in mind, and have strong statistical power (i.e. a lower Type II error). The underlying assumptions can sometimes be hard to verify (see Dror et al. (2018) §3.1), thus when in doubt *non-parametric* tests can be used. This category features tests like the Bootstrap, employed in case of a small sample size or the Wilcoxon signed-rank test (Wilcoxon, 1992) when many observations are available. Depending on the application, the usage of specialized tests might furthermore be desirable (Dror et al., 2019; Agarwal et al., 2021). Due to spatial constraints, we here refer to Dror et al. (2018); Raschka (2018) for a general introduction to the topic and Azer et al. (2020) for an overview over Bayesian significance tests. In Appendix A.4, we also list a number of resources, such as Bayesian significance tests by Azer et al. (2020), an implementation of the test by Dror et al. (2019) by Ulmer (2021) and a test framework that is adapted for deep reinforcement learning by Agarwal et al. (2021). With the necessary tools at hand, we can now return to carefully answer the original research questions. Azer et al. (2020) provide a guide on how to adequately word insights when a statistical test was used, and Greenland et al. (2016) list common pitfalls and misinterpretations of results. We also want to draw attention to the fact that comparisons between multiple models and/or datasets, *require* an adjustment of the confidence level, for instance using the Bonferroni correction (Bonferroni, 1936), which is a safe and conservative choice and easily implemented for most tests (Dror et al., 2017; Ulmer, 2021).

Critiques & Alternatives Although statistical hypothesis testing is an established tool in many disciplines, its (mis)use has received criticism for decades (Berger & Sellke, 1987; Demšar, 2008; Ziliak & McCloskey, 2008). For instance, Wasserstein et al. (2019) recommend not to frame the p-value as a gatekeeper between a dichotomous "significant" and "not significant"—something that has been argued to reinforce a publication bias, i.e., a favoring of positive results (Locascio, 2017)—but instead report it as a continuous value, interpreting any results with the appropriate scepticism and uncertainty.[4] In addition to statistical significance, another approach advocates for reporting *effect size* (Berger & Sellke, 1987; Lin et al., 2013), so for instance the mean difference, or the absolute or relative gain in performance for a model compared to a baseline. The effect size can be modeled using Bayesian analysis (Kruschke, 2013; Benavoli et al., 2017), which better fit the uncertainty surrounding experimental results, but requires the specification of a plausible model[5] and potentially the usage of Markov Chain Monte Carlo sampling (Brooks et al., 2011; Gelman et al., 2013). Benavoli et al. (2017) give a tutorial for applications to ML and supply an implementation of their proposed methods in a software package (see Appendix C).

Reporting Results Lastly, report the number of runs/random seeds used, and, if appropriate, the significance threshold or confidence level of the statistical test or the underlying Bayesian model and chosen priors. Report all scores using mean and standard deviation and report p-values or comparable quantities as continuous values, not binary decisions. Using those results, evaluate the evidence for and against your initial hypotheses.

[4] Or, as Wasserstein et al. (2019) note: "*statistically significant*—don't say it and don't use it".

[5] Here, we are *not* referring to a neural network, but instead to a process generating experimental observations, specifying a prior and likelihood for model scores. Conclusions are drawn from the posterior distribution over parameters of interest (e.g., the mean performance), as demonstrated by Benavoli et al. (2017).

Best Practices: **Experiments & Analysis**

- ◇ Report mean & standard dev. over multiple runs;
- ◇ Perform significance testing or Bayesian analysis and motivate your choice of method;
- ◇ Carefully reflect on the amount of evidence regarding your initial hypotheses.

6 PUBLICATION

Subsequent to all the prior consideration, the publication step of a research project allows the findings to be spread across the scientific community. In this section, we discuss some additional trends in the DL field that researchers should consider when publishing their work.

Citation Control Frequently, researchers cite non-archival versions of papers without noticing that the paper has been published already. The published version of a paper is peer-reviewed, increasing the probability that any mistakes or ambiguities have been resolved. In Appendix C we suggest tools to verify the version of any cited papers.

Hardware Requirements The paper should report the computing infrastructure used. At minimum, the specifics about the CPU and GPU. This is for indicating the amount of compute necessary for the project, but also for the sake of replicability issues due to the non-deterministic nature of the GPU (Jean-Paul et al., 2019; Wei et al., 2020). Moreover, Dodge et al. (2019) demonstrate that test performance scores alone are insufficient for claiming the dominance of a model over another, and argue for reporting additional performance details on validation data as a function of computation budget, which can also estimate the amount of computation required to obtain a given accuracy.

Environmental Impact The growth of computational resources required for DL over the last decade has led to financial and carbon footprint discussions in the AI community. Schwartz et al. (2020) introduce the distinction between *Red AI*—AI research that seek to obtain state-of-the-art results through the use of massive computational power—and *Green AI*—AI research that yields novel results without increasing computational cost. In the paper the authors propose to add *efficiency* as an evaluation criterion alongside accuracy measures. Strubell et al. (2019) discuss the problem from a more NLP-specific perspective: They quantify the approximate financial and environmental costs of training a variety of widely used models for NLP (e.g., BERT, GPT-2). In conclusion, to reduce costs and improve equity, they propose (1) *Reporting training time and sensitivity to hyperparameters*, (2) *Equitable access to computation resources*, and (3) *Prioritizing computationally efficient hardware and algorithms* (Appendix C includes a tool for CO_2 estimation of computational models).

Social Impact The widespread of DL studies and their increasing use of human-produced data (e.g., from social media and personal devices) means the outcome of experiments and applications have direct effects on the lives of individuals. Waseem et al. (2021) argue that addressing and mitigating biases in ML is near-impossible as subjectivity is inescapable and thus converging in a universal truth may further harm already marginalized social groups. As a follow-up, the authors argue for a reflection on the consequences the imaginary objectivity of ML has on political choices. From the NLP perspective, Hovy & Spruit (2016) analyze and discuss the social impact research may have beyond the more explored privacy issues. They make an ethical analysis of NLP on social justice, i.e., equal opportunities for individuals and groups, and underline three problems of the mutual relationship between language, society and individuals: exclusion, over-generalization and overexposure.

Ethical Considerations There has been effort on the development of concrete ethical guidelines for researchers within the ACM Code of Ethics and Professional Conduct (Association for Computing Machinery, 2022). The Code lists seven principles stating how fundamental ethical principles apply to the conduct of a computing professional (like DL and NLP practitioners) and is based on two main ideas: computing professionals' actions change the world and the public good is always the primary consideration. Mohammad (2021) discusses the importance of going beyond

individual models and datasets, back to the ethics of the task itself. As a practical recommendation, he presents *Ethics Sheets for AI Tasks* as tools to document ethical considerations *before* building datasets and developing systems. In addition, researchers are invited to collect the ethical considerations of the paper in a cohesive narrative, and elaborate them in a paragraph, usually in the Introduction/Motivation, Data, Evaluation, Error Analysis or Limitations section (Mohammad, 2020; Hardmeier et al., 2021).

Best Practices: **Publication**

- ◇ Avoid citing pre-prints (if applicable);
- ◇ Describe the computational requirements;
- ◇ Consider the potential ethical & social impact;
- ⋆ Consider the environmental impact and prioritize computational efficiency;
- ⋆ Include an Ethics and/or Bias Statement.

7 Discussion

Several of the cited references in this work have either highlighted a deficiency in methodological standards, or a necessity to update them. We want to dedicate this last section to discuss structural issues regarding the implementation of our recommendations in a decidedly opinionated way.

Compute Requirements Specifically with regard to statistical significance in Section 5, there is a stark tension between the hardware requirements of modern methods (Sevilla et al., 2022) and the computational budget of the average researcher as well as the uncertainty under which experimental results are interpreted. Significance tests require many runs to produce reliable results: Neural network performance may fluctuate wildly,[6] and thus pose daunting computational costs, which but the best-funded research labs can afford (Hooker, 2021). Under these circumstances, it becomes difficult to judge whether the results obtained via larger models and datasets *actually* constitute substantial progress or just statistical flukes. At the same time, such experiments can create environmental concerns (Strubell et al., 2019; Schwartz et al., 2020).[7] Therefore, the community must decide collectively whether these factors, including impeded reproducibility and weakened empirical evidence, constitute a worthy price for the knowledge obtained from training large neural networks.

Incentives in Publishing As demonstrated by Figure 2, DL has gained traction as an empirical field of research. At such a point, more rigorous standards are necessary to maintain high levels of scholarship. Unfortunately, we see this process lagging behind, illustrated by repeated calls for improvement (Henderson et al., 2018; Gundersen & Kjensmo, 2018; Agarwal et al., 2021; Narang et al., 2021). Why is that so? We speculate that many problems are intertwined with the incentives set by the current publishing environment: The career of students and researchers often hinges on published papers and incoming citations. As of now, better experimental standards are not aligned with this goal, since they often do not increase the acceptance probability: The more details are provided for replicability purposes, the more potential points of criticism are exposed to reviewers. Under these circumstances, the quality of published works decreases, replication suffers and peer review load increases. In this vein, Chu & Evans (2021) show how an increased amount of papers actually leads to *slowed* progress in a field, making it harder for new, promising ideas to break through. It also creates adverse incentives for actors to "rig the benchmark lottery" (Dehghani et al., 2021), since achieving state-of-the-art results remains an import requirement for publishing (Birhane et al., 2021).

Culture Change How can we change this trend? **As researchers**, we can start implementing a lot of the recommendations in this work in order to drive bottom-up change by reaching a critical

[6] E.g., BERT models with different seeds (Sellam et al., 2021) sometimes perform quite differently compared to the originally released model (Devlin et al., 2019).

[7] E.g., GPT-3 was estimated to have cost ca. 12M USD (Turner, 2020) or 188,702 kWh of energy to train (Anthony et al., 2020).

mass (Centola et al., 2018). **As reviewers**, we can shift focus from results to more rigorous methodologies (Rogers & Augenstein, 2021) and allow more critiques of past works and meta-reviews to be published (Birhane et al., 2021; Lampinen et al., 2021). **As a community**, we can change the incentives around research and experiment with new initiatives. Rogers & Augenstein (2020) and Su (2021) give recommendations on how to improve the peer-review process by better paper-reviewer matching and paper scoring.[8] Other attempts are currently undertaken to encourage reproducibility and reproduction of past works.[9] Other ideas change the publishing process more fundamentally, for instance by splitting it into two steps: The first part, where authors are judged solely on the merit of their research question and methodology; and the second one, during which the analysis of their results is evaluated (Locascio, 2017). This aims to reduce publication bias and puts more scrutiny on the experimental methodology. In a similar vein, van Miltenburg et al. (2021) recommend a procedure similar to clinical studies, where whole research projects are pre-registered, i.e., specifying the parameters of research before carrying out any experiments (Nosek et al., 2018). The implications of these ideas are not only positive, however, as a slowing rate of publishing might disadvantage junior researchers (Chu & Evans, 2021).

Limitations This work comes with two main limitations: On the one hand, it can only take a snapshot of an ongoing discussion. On the other hand, it was written from an NLP perspective, and thus some of the listed suggestions might not apply to other subfields employing DL methods. With these limitations in mind, we invite members of the community to contribute to our open-source repository.

8 Conclusion

Being able to (re-)produce empirical findings is critical for scientific progress, particularly in fast-growing fields Manning (2015). To reduce the risks of a reproducibility crisis and unreliable research findings Ioannidis (2005), experimental rigor is imperative. Being aware of possible harmful implications and to avoid them is therefore important. Every step carries possible biases Hovy & Prabhumoye (2021); Waseem et al. (2021). While necessarily incomplete, this paper aims at providing a rich toolbox of actionable recommendations *for each research step*, and a reflection and summary of the ongoing broader discussion. With concrete best practices to raise awareness and call for uptake, we hope to aid researchers in their empirical endeavors.

Acknowledgments

We would like to thank Giovanni Cinà, Rotem Dror, Miryam de Lhoneux, and Tanja Samardžić for their feedback on this draft. Furthermore, we would like to express our gratitude to the NLPnorth group in general for frequent discussions and feedback on this work.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 34, 2021.

Lasse F Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint arXiv:2007.03051*, 2020.

Marianna Apidianaki, Saif Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat. Proceedings of the 12th international workshop on semantic evaluation. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018.

[8] Faggion (2016) for instance has argued to also make the reviewing process itself more transparent. We can see developments in this direction in the adoption of OpenReview.

[9] See for instance the reproducibility certication of the TMLR journal (TMLR, 2022), NAACL 2022 reproducibility badges (Association for Computational Linguistics, 2022).

Association for Computational Linguistics. Reproducibility criteria. https://2022.naacl.org/calls/papers/#reproducibility-criteria, 2022. Accessed: 2022-02-09.

Association for Computing Machinery. Acm code of ethics and professional conduct. https://www.acm.org/code-of-ethics, 2022. Accessed: 2022-02-10.

Erfan Sadeqi Azer, Daniel Khashabi, Ashish Sabharwal, and Dan Roth. Not all claims are created equal: Choosing the right statistical approach to assess hypotheses. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 5715–5725. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.506. URL https://doi.org/10.18653/v1/2020.acl-main.506.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. We need to consider disagreement in evaluation. In *1st Workshop on Benchmarking: Past, Present and Future*, pp. 15–21. Association for Computational Linguistics, 2021.

Christine Basta, Marta R Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp. 33–39, 2019.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. ReproGen: Proposal for a shared task on reproducibility of human evaluations in NLG. In *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 232–236, Dublin, Ireland, December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.inlg-1.29.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. A systematic review of reproducibility research in natural language processing. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 381–393. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.eacl-main.29/.

Alessio Benavoli, Giorgio Corani, Francesca Mangili, Marco Zaffalon, and Fabrizio Ruggeri. A bayesian wilcoxon signed-rank test based on the dirichlet process. In *International conference on machine learning*, pp. 1026–1034. PMLR, 2014.

Alessio Benavoli, Giorgio Corani, Janez Demsar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *J. Mach. Learn. Res.*, 18:77:1–77:36, 2017. URL http://jmlr.org/papers/v18/16-305.html.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:58–04, 12 2018. doi: 10.1162/tacla00041. URL http://dx.doi.org/10.1162/tacla00041.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021.

Martin Benjamin. Hard numbers: Language exclusion in computational linguistics and natural language processing. In *Proceedings of the LREC 2018 Workshop "CCURL2018–Sustaining Knowledge Diversity in the Digital Age*, pp. 13–18, 2018.

James O Berger and Thomas Sellke. Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American statistical Association*, 82(397):112–122, 1987.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590*, 2021.

Bernd Bischl, Martin Binder, Michel Lang, Tobias Pielok, Jakob Richter, Stefan Coors, Janek Thomas, Theresa Ullmann, Marc Becker, Anne-Laure Boulesteix, et al. Hyperparameter optimization: Foundations, algorithms, best practices and open challenges. *arXiv preprint arXiv:2107.05847*, 2021.

Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3, 2021.

Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.385. URL https://aclanthology.org/2021.naacl-main.385.

Anouck Braggaar and Rob van der Goot. Challenges in annotating and parsing spoken, code-switched, Frisian-Dutch data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pp. 50–58, Kyiv, Ukraine, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.adaptnlp-1.6.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 9263–9274. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.745. URL https://doi.org/10.18653/v1/2020.emnlp-main.745.

Sean M Carroll. Beyond falsifiability: Normal science in a multiverse. *Why trust a theory*, pp. 300–314, 2019.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6588–6608, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. URL https://www.aclweb.org/anthology/2020.coling-main.579.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad,

Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a glance: An audit of web-crawled multilingual datasets, 2021.

Damon Centola, Joshua Becker, Devon Brackbill, and Andrea Baronchelli. Experimental evidence for tipping points in social convention. *Science*, 360(6393):1116–1119, 2018.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34, 2021.

Johan SG Chu and James A Evans. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41), 2021.

K Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany J Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E Hunter. Three dimensions of reproducibility in natural language processing. In *LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources and Evaluation*, volume 2018, pp. 156. NIH Public Access, 2018.

Giorgio Corani and Alessio Benavoli. A bayesian approach for comparing cross-validated algorithms on multiple data sets. *Machine Learning*, 100(2):285–304, 2015.

Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. 2021a.

Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2021b.

Alicia Curth, David Svensson, James Weatherall, and Mihaela van der Schaar. Really doing great at estimating cate? a critical look at ml benchmarking practices in treatment effect estimation. 2021.

Alexander D'Amour, Katherine A. Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yi-An Ma, Cory Y. McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *CoRR*, abs/2011.03395, 2020. URL `https://arxiv.org/abs/2011.03395`.

Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. 2021.

Janez Demšar. On the appropriateness of statistical tests in machine learning. In *Workshop on Evaluation Methods for Machine Learning in conjunction with ICML*, pp. 65. Citeseer, 2008.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Jesse Dodge and Noah A. Smith. Reproducibility at emnlp 2020. `https://2020.emnlp.org/blog/2020-05-20-reproducibility`, 2020.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1224. URL https://www.aclweb.org/anthology/D19-1224.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Trans. Assoc. Comput. Linguistics*, 5:471–486, 2017. URL https://transacl.org/ojs/index.php/tacl/article/view/1241.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 1383–1392. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance - how to properly compare deep neural models. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 2773–2785. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1266. URL https://doi.org/10.18653/v1/p19-1266.

Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart. Statistical significance testing for natural language processing. *Synthesis Lectures on Human Language Technologies*, 13(2): 1–116, 2020.

Chris Drummond. Replicability is not reproducibility: Nor is it good science. *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, 01 2009.

ELRA. The european language resources association (elra). http://www.elra.info/en/about/, 1995. Accessed: 2022-02-10.

CM Faggion. Improving the peer-review process from the perspective of an author and reviewer. *British dental journal*, 220(4):167–168, 2016.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1691–1701, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P13-1166.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Computing Research Repository*, arxiv:1803.09010, 2020. URL https://arxiv.org/abs/1803.09010. version 7.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint arXiv:2202.06935*, 2022.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. Bayesian data analysis. *Chapman Hall, London*, 2013.

Zhiqiang Gong, Ping Zhong, and Weidong Hu. Diversity in machine learning. *IEEE Access*, 7: 64323–64350, 2019.

Google Trends. "deep learning" by google trends. https://trends.google.com/trends/explore?date=all&q=Deep%20Learning, 2022. Accessed: 2022-02-09.

Kyle Gorman and Steven Bedrick. We need to talk about standard splits. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 2786–2791, 2019.

Sander Greenland, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31(4):337–350, 2016.

Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://www.aclweb.org/anthology/2020.acl-main.740.

Christian Hardmeier, Marta R. Costa-jussà, Kellie Webster, Will Radford, and Su Lin Blodgett. How to write a bias statement: Recommendations for submissions to the workshop on gender bias in NLP. *CoRR*, abs/2104.03026, 2021. URL https://arxiv.org/abs/2104.03026.

Jianxing He, Sally L Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. The practical implementation of artificial intelligence technologies in medicine. *Nature medicine*, 25(1):30–36, 2019.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 3207–3214. AAAI Press, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16669.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020.

Sara Hooker. The hardware lottery. *Communications of the ACM*, 64(12):58–65, 2021.

Dirk Hovy and Shrimai Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021.

Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 591–598, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2096. URL https://www.aclweb.org/anthology/P16-2096.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 125–136, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/e2c420d928d4bf8ce0ff2ec19b371514-Abstract.html.

John P. A. Ioannidis. Why most published research findings are false. *PLOS Medicine*, 2(8): null, 08 2005. doi: 10.1371/journal.pmed.0020124. URL https://doi.org/10.1371/journal.pmed.0020124.

Nathalie Japkowicz and Mohak Shah. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

S Jean-Paul, T Elseify, I Obeid, and J Picone. Issues in the reproducibility of deep learning results. In *2019 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–4. IEEE, 2019.

Theis Ingerslev Jensen, Bryan T Kelly, and Lasse Heje Pedersen. Is there a replication crisis in finance? Technical report, National Bureau of Economic Research, 2021.

Leslie K John, George Loewenstein, and Drazen Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5):524–532, 2012.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020. doi: 10.1162/tacl_a_00300. URL https://www.aclweb.org/anthology/2020.tacl-1.5.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. *CoRR*, abs/2012.07421, 2020. URL https://arxiv.org/abs/2012.07421.

John K Kruschke. Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1 (5):658–676, 2010.

John K Kruschke. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2):573, 2013.

John K Kruschke and Torrin M Liddell. Bayesian data analysis for newcomers. *Psychonomic bulletin & review*, 25(1):155–177, 2018.

Thomas S Kuhn. *The structure of scientific revolutions*, volume 111. Chicago University of Chicago Press, 1970.

Andrew Kyle Lampinen, Stephanie CY Chan, Adam Santoro, and Felix Hill. Publishing fast and slow: A path toward generalizability in psychology and ai. 2021.

A-M Leventi-Peetz and T Östreich. Deep learning reproducibility and explainable ai (xai). *arXiv preprint arXiv:2202.11452*, 2022.

Quentin Lhoest, Albert Villanova del Moral, Patrick von Platen, Thomas Wolf, Yacine Jernite, Abhishek Thakur, Lewis Tunstall, Suraj Patil, Mariama Drame, Julien Chaumond, Julien Plu, Joe Davison, Simon Brandeis, Teven Le Scao, Victor Sanh, Kevin Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Steven Liu, Nathan Raw, Sylvain Lesage, Théo Matussière, Lysandre Debut, Stas Bekman, and Clément Delangue. huggingface/datasets: 1.12.1, September 2021. URL https://doi.org/10.5281/zenodo.5510481.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

Mark Liberman. Replicability vs. reproducibility — or is it the other way around? https://languagelog.ldc.upenn.edu/nll/?p=21956, 2015. Accessed: 2022-02-21.

Mingfeng Lin, Henry C Lucas Jr, and Galit Shmueli. Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research*, 24(4):906–917, 2013.

Joseph J Locascio. Results blind science publishing. *Basic and applied social psychology*, 39(5): 239–246, 2017.

Eric L Manibardo, Ibai Laña, and Javier Del Ser. Deep learning for road traffic forecasting: Does it make a difference? *IEEE Transactions on Intelligent Transportation Systems*, 2021.

Christopher D. Manning. Last words: Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707, December 2015. doi: doi:10.1162/COLI_a_00239. URL https://www.aclweb.org/anthology/J15-4006.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 7297–7306. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.566. URL https://doi.org/10.18653/v1/2021.acl-long.566.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 2437–2445, 2020.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4):659–684, 2020.

Saif M. Mohammad. What is a research ethics statement and why does it matter?, 2020. URL http://www.saifmohammad.com/WebDocs/EthicsStatement-web.pdf.

Saif M. Mohammad. Ethics sheets for AI tasks. *CoRR*, abs/2107.01183, 2021. URL https://arxiv.org/abs/2107.01183.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=fSTD6NFIW_b.

Sharan Narang, Hyung Won Chung, Yi Tay, Liam Fedus, Thibault Févry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, Yanqi Zhou, Wei Li, Nan Ding, Jake Marcus, Adam Roberts, and Colin Raffel. Do transformer modifications transfer across implementations and applications? In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 5758–5773. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.465. URL https://doi.org/10.18653/v1/2021.emnlp-main.465.

Adrian Nilsson, Simon Smith, Gregor Ulm, Emil Gustavsson, and Mats Jirstrand. A performance evaluation of federated learning algorithms. In *Proceedings of the second workshop on distributed infrastructures for deep learning*, pp. 1–8, 2018.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4034–4043, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.497.

Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

Silviu Paun, Ron Artstein, and Massimo Poesio. Statistical methods for annotation analysis. *Synthesis Lectures on Human Language Technologies*, 15(1):1–217, 2022.

Ted Pedersen. Last words: Empiricism is not a matter of faith. *Computational Linguistics*, 34 (3):465–470, 2008. doi: 10.1162/coli.2008.34.3.465. URL https://www.aclweb.org/anthology/J08-3010.

Roger D Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 46–54, 2020.

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. DaN+: Danish nested named entities and lexical normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6649–6662, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

Karl Popper. *Karl Popper: Logik der Forschung*. Mohr Siebeck, Tübingen, Germany, 1934.

James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly Media, Inc., 2012.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, and Roger Zimmermann. Domain divergences: A survey and empirical analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1830–1849, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.147. URL https://aclanthology.org/2021.naacl-main.147.

Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

Kenneth Reitz and Tanya Schlusser. *The Hitchhiker's guide to Python: best practices for development*. " O'Reilly Media, Inc.", 2016.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://www.aclweb.org/anthology/2020.acl-main.442.

Stefan Riezler and Michael Hagmann. Validity, reliability, and significance. 2021.

Stefan Riezler and John T Maxwell III. On some pitfalls in automatic evaluation and significance testing for mt. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 57–64, 2005.

Anna Rogers and Isabelle Augenstein. What can we do to improve peer review in nlp? In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pp. 1256–1262. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.112. URL https://doi.org/10.18653/v1/2020.findings-emnlp.112.

Anna Rogers and Isabelle Augenstein. How to review for acl rolling review? https://aclrollingreview.org/reviewertutorial, 2021. Accessed: 2022-02-21.

David Schlangen. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 670–674, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.85. URL https://aclanthology.org/2021.acl-short.85.

Robin M Schmidt, Frank Schneider, and Philipp Hennig. Descending through a crowded valley-benchmarking deep learning optimizers. In *International Conference on Machine Learning*, pp. 9367–9376. PMLR, 2021a.

Victor Schmidt, Kamal Goyal, Aditya Joshi, Boris Feld, Liam Conell, Nikolas Laskaris, Doug Blank, Jonathan Wilson, Sorelle Friedler, and Sasha Luccioni. CodeCarbon: Estimate and Track Carbon Emissions from Machine Learning Computing. 2021b. doi: 10.5281/zenodo.4658424.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *Commun. ACM*, 63(12): 54–63, 2020. doi: 10.1145/3381831. URL https://doi.org/10.1145/3381831.

Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, et al. The multiberts: Bert reproductions for robustness analysis. *arXiv preprint arXiv:2106.16163*, 2021.

Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. Compute trends across three eras of machine learning. *arXiv:2202.05924 [cs]*, Feb 2022. arXiv: 2202.05924.

Anastasia Shimorina, Yannick Parmentier, and Claire Gardent. An error analysis framework for shallow surface realization. *Transactions of the Association for Computational Linguistics*, 9: 429–446, 2021.

Herbert A Simon. Artificial intelligence: an empirical science. *Artificial Intelligence*, 77(1):95–127, 1995.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pp. 2960–2968, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. We need to talk about random splits. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 1823–1832. Association for Computational Linguistics, 2021. URL https://www.aclweb.org/anthology/2021.eacl-main.156/.

Lucia Specia. Disagreement in human evaluation: blame the task not the annotators. NoDaLiDa keynote, 2021. URL https://nodalida2021.github.io/invited_speakers.html.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650, Florence, Italy, July 2019. Association for Computational

Linguistics. doi: 10.18653/v1/P19-1355. URL https://www.aclweb.org/anthology/P19-1355.

Weijie Su. You are the best reviewer of your own papers: An owner-assisted scoring mechanism. *Advances in Neural Information Processing Systems*, 34, 2021.

TMLR. Submission guidelines and editorial policies. https://jmlr.org/tmlr/editorial-policies.html, 2022. Accessed: 2022-02-09.

Tina Tseng, Amanda Stent, and Domenic Maida. Best practices for managing data annotation projects. *CoRR*, abs/2009.11654, 2020. URL https://arxiv.org/abs/2009.11654.

Stéphane Tufféry. *Data mining and statistics for decision making*. John Wiley & Sons, 2011.

Elliot Turner. Twitter post (@eturner303): Reading the openai gpt-3 paper. https://twitter.com/eturner303/status/1266264358771757057, 2020. Accessed: 2022-02-09.

Dennis Ulmer. deep-significance: Easy and better significance testing for deep neural networks, 2021.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72: 1385–1470, 2021.

Rob van der Goot. We need to talk about train-dev-test splits. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 4485–4494. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.emnlp-main.368.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in nlp, 2021.

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. Preregistering NLP research. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 613–623. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.51. URL https://doi.org/10.18653/v1/2021.naacl-main.51.

Daniel Varab and Natalie Schluter. DaNewsroom: A large-scale Danish summarisation dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 6731–6739, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4.

Daniel Varab and Natalie Schluter. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10150–10161, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

Tamás Váradi, Peter Wittenburg, Steven Krauwer, Martin Wynne, and Kimmo Koskenniemi. Clarin: Common language resources and technology infrastructure. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. Disembodied machine learning: On the illusion of objectivity in nlp, 2021.

Ronald L Wasserstein, Allen L Schirm, and Nicole A Lazar. Moving to a world beyond "p <0.05", 2019.

Junyi Wei, Yicheng Zhang, Zhe Zhou, Zhou Li, and Mohammad Abdullah Al Faruque. Leaky dnn: Stealing deep-learning model secret with gpu context-switching side-channel. In *2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pp. 125–137. IEEE, 2020.

Jennifer C. White and Ryan Cotterell. Examining the inductive bias of neural language models with artificial languages. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 454–463. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.38. URL https://doi.org/10.18653/v1/2021.acl-long.38.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. Reproducibility in computational linguistics: are we willing to share? *Computational Linguistics*, 44(4):641–649, 2018.

Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pp. 196–202. Springer, 1992.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhenguo Li. Ood-bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. *arXiv preprint arXiv:2106.03721*, 2021.

Hongkun Yu, Chen Chen, Xianzhi Du, Yeqing Li, Abdullah Rashwan, Le Hou, Pengchong Jin, Fan Yang, Frederick Liu, Jaeyoun Kim, and Jing Li. TensorFlow Model Garden. https://github.com/tensorflow/models, 2020.

Ke-Hai Yuan and Kentaro Hayashi. Bootstrap approach to inference and power analysis based on three test statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology*, 56(1):93–110, 2003.

Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, et al. The ai index 2021 annual report. *arXiv preprint arXiv:2103.06312*, 2021.

Mike Zhang and Barbara Plank. Cartography active learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pp. 395–406. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.36. URL https://doi.org/10.18653/v1/2021.findings-emnlp.36.

Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015.

Steve Ziliak and Deirdre Nansen McCloskey. *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press, 2008.

Shi Zong, Ashutosh Baheti, Wei Xu, and Alan Ritter. Extracting COVID-19 events from twitter. *CoRR*, abs/2006.02567, 2020. URL https://arxiv.org/abs/2006.02567.

A CASE STUDIES & FURTHER READING

The implementation of the methods we advocate for in our work can be challenging. This is why we dedicate this appendix to listing further resources and pointing to examples that illustrate their intended use.

A.1 DATA

Data Statement Following Bender & Friedman (2018), the long form data statement should outline CURATION RATIONALE, LANGUAGE VARIETY, SPEAKER DEMOGRAPHIC, ANNOTATOR DEMOGRAPHIC, SPEECH SITUATION, TEXT CHARACTERISTICS and a PROVENANCE APPENDIX. A good example of a long form data statement can be found in Appendix B in Plank et al. (2020), where each of the former mentioned topics are outlined. For example, with respect to ANNOTATOR DEMOGRAPHIC, they mention "three students and one faculty (age range: 25-40), gender: male and female. White European. Native language: Danish, German. Socioeconomic status: higher-education student and university faculty." This is a concise explanation of the annotators involved in their project.

Data Quality Text corpora today are building blocks for many downstream NLP applications like question answering and summarization. In the work of Caswell et al. (2021), they audit the quality of quality of 205 language-specific corpora released within major public datasets. At least 15 of these 205 corpora have no usable text, and a large fraction contains less than 50% sentences of acceptable quality. The tacit recommendation is looking at samples of any dataset before using it or releasing it to the public. A good example is Varab & Schluter (2020; 2021), who filter out low-quality news articles from their summarization dataset with empty summaries or bodies, removing duplicates, and removing summaries that are long than them main body of text. More wide varieties of data filtering can be applied, like filtering on length-ratio, LangID, and TF-IDF wordlists Caswell et al. (2020). Note that there is no easy solution—data cleaning is not a trivial task Caswell et al. (2021).

Universal Dependencies Nivre et al. (2020) aims to annotate syntactic dependencies in addition to part-of-speech tags, morphological features etc. for as many languages as possible within a consistent set of guidelines. The dataset which consists of treebanks contributed by various authors is updated in a regular half-yearly cycle and is hosted on the long-term storage LINDAT/CLARIN repository (Váradi et al., 2008). Each release is clearly versioned such that fair comparisons can be made even while guidelines are continuously adapted. Maintenance of the project is conducted on a public `git` repository, such that changes to both the data and the guidelines can be followed transparently. This allows for contributors to suggest changes via pull requests.

A.2 MODELS

There are several libraries that allow for model hosting or distribution of model weights for "mature" models. `HuggingFace` Wolf et al. (2020) is an example of hosting models for distribution. It is an easy-to-use library for practitioners in the field. Other examples of model distribution is `Keras Applications`[10] or `TensorFlow Model Garden` Yu et al. (2020). Other ways of distributing models is setting hyperlinks in the repository (e.g., Joshi et al., 2020), to load the models from the checkpoints they have been saved to. A common denominator of all the aforementioned libraries is to list relevant model performances (designated metrics per task), the model size (in bytes), model parameters (e.g., in millions), and inference time (e.g., any time variable).

A.3 CODEBASE

At the code-level, there are several examples of codebases with strong documentation and clean project structure. We define documentation and project structure in Appendix B. Here, we give examples going from smaller projects to larger Python projects:

[10] `https://keras.io/api/applications/`

The codebase of CateNETS Curth & van der Schaar (2021a;b); Curth et al. (2021)[11] shows a clear project structure. This includes unit tests, versioning of the library, and licensing. In addition, there are specific files for each published work to replicate the results.

Not all projects require a `pip` installation or unit tests. For example—similar to the previous project—MaChAmp van der Goot et al. (2021)[12] shows detailed documentation, including several reproducible experiments shown in the paper (including files with model scores) and a clear project structure. Here, one possible complication lies in possible dependency issues once the repository grows, with unit tests as a mitigation strategy.

AdapterHub Pfeiffer et al. (2020)[13] demonstrates the realization of a large-scale project. This includes tutorials, configurations, and hosting of technical documentation (`https://docs.adapterhub.ml/`), as well as a dedicated website for the library itself.

A.4 Experimental Analysis

Statistical Hypothesis Testing A general introduction to significance testing in NLP is given by Dror et al. (2018); Raschka (2018); Azer et al. (2020). Furthermore, Dror et al. (2020) and Riezler & Hagmann (2021) provide textbooks around hypothesis testing in an NLP context. Japkowicz & Shah (2011) describe the usage of statistical test for general, classical ML classification algorithms. When it comes to usage, Zhang & Plank (2021) describe the statistical test used with all parameter and results alongside performance metrics. Shimorina et al. (2021) report p-values alongside test statistics for the Spearman's ρ test, using the Bonferroni correction due to multiple comparisons. Apidianaki et al. (2018) transparently report the p-values of a approximate randomization test (Riezler & Maxwell III, 2005) between all the competitors in an argument reasoning comprehension shared task and interpret them with the appropriate degree of carefulness.

Bayesian analysis Bayesian Data Analysis has a long history of application across many scientific disciplines. Popular textbooks about the topic are given by Kruschke (2010); Gelman et al. (2013) with a more gentle introduction by Kruschke & Liddell (2018). Benavoli et al. (2017) supply an in-depth tutorial for Bayesian Analysis for Machine Learning, by using a Bayesian signed ranked test (Benavoli et al., 2014), an extension of the frequentist Wilcoxon signed rank test and a Bayesian hierarchical correlated t-test (Corani & Benavoli, 2015). Applications can be found for instance by Nilsson et al. (2018), who use the Bayesian correlated t-test (Corani & Benavoli, 2015) to investigate the posterior distribution over the performance difference to compare different federated learning algorithms. To evaluate deep neural networks on road traffic forecasting, Manibardo et al. (2021) employ Bayesian analysis and plot Monte Carlo samples from the posterior distribution between pairs of models. The plots include ROPEs, i.e., regions of practical equivalence, where the judgement about the superiority of a model is suspended.

A.5 Publication Considerations

Replicability Gururangan et al. (2020) report in detail all the computational requirements for their adaptation techniques in a dedicated sub-section. Additionally, following the suggestions by Dodge et al. (2019), the authors report their results on the development set in the appendix.

Environmental Impact By introducing MultiBERTs Sellam et al. (2021), the authors include in their paper an *Environmental Statement*. In the paragraph they estimate the computational cost of their experiments in terms of hours, and consequential tons of CO2e. They release the trained models publicly with the aim to allow subsequent studies by other researchers without the computational cost of training MultiBERTs to be incurred.

Social and Ethical Impact Brown et al. (2020) present GPT-3 and include a whole section on the *Broader Impacts* language models like GPT-3 have. Despite improving the quality of text generation, they also have potentially harmful applications. Specifically, the authors discuss the potential for

[11] `https://github.com/AliciaCurth/CATENets`
[12] `https://github.com/machamp-nlp/machamp`
[13] `https://github.com/Adapter-Hub/adapter-transformers`

deliberate misuse of language models, and the potential issues of bias, fairness and representation (focusing on the gender, race and religion dimensions).

The work of Hardmeier et al. (2021) assists the researcher in writing a bias statement, by recommending to provide explicit statements of why the system's behaviors described as "bias" are harmful, in what ways, and to whom, then to reason on them. In addition, they provide an example of a bias statement from Basta et al. (2019).

B CONTENTS OF CODEBASE

The README First, the initial section of the `README` would consist of the name of the repository—to what paper or project is this code base tied to? Including a hyperlink to the paper or project itself. Second, developers also indicate the structure of the repository—what and where are the files, folders, code, et cetera in the project and how would they be used.

Empirical work requires the installation of libraries or software. It is important to install the right versions of the libraries to maintain replicability, and indicate the correct version of the specific package. In Python, a common practice is to make use of virtual environments in combination with a `requirements.txt` file. The main purpose of a virtual environment is to create an isolated environment for code projects. Each project can have its own dependencies (libraries) regardless of what dependencies every other project has to avoid clashes between libraries. For example, this file can be created by piping the output of `pip freeze` to a `requirements.txt` file. For further examples of virtual environment tools, we refer to Table 1 (Appendix C).

To ensure replicability, the practitioner writes a description on how to re-run all experiments that are depicted in a paper to get the same results. For example, these are evaluation scores or graphical plots. This can come in the form of a bash script, that indicates all the commands necessary.[14] Similarly, one can also indicate all commands in the `README`. To give credit to each others work, the last section of the `README` is usually reserved for credits, acknowledgments, and the citation. The citation is preferably provided in `BibTeX` format.

Project Structure From the Python programming language perspective, there are several references for initializing an adequate Python project structure.[15] This includes a `README`, LICENSE, `setup.py`, `requirements.txt`, and unit tests. To quote *The Hitchhiker's Guide to Python* (Reitz & Schlusser, 2016) on the meaning of "structure":

> "By 'structure' we mean the decisions you make concerning how your project best meets its objective. We need to consider how to best leverage Python's features to create clean, effective code. In practical terms, 'structure' means making clean code whose logic and dependencies are clear as well as how the files and folders are organized in the filesystem."

This includes decisions on where functions should go into which modules. Also on how data flows through the project. What features and functions can be grouped together or even isolated? In a broader sense, to answer the question on how the finished product should look like.

C RESOURCES

An overview over all mentioned resources in the paper is given in Table 1.

[14] See for instance `https://robvanderg.github.io/blog/repro.htm`

[15] Some examples: `https://docs.python-guide.org/writing/structure/` and `https://coderefinery.github.io/reproducible-research/02-organizing-projects/`

Table 1: Overview over mentioned resources.

Name	Description	Link
Anonymous Github	Website to anonymize a Github repository.	https://anonymous.4open.science
baycomp (Benavoli et al., 2017)	Implementation of Bayesian tests for the comparison of classifiers.	https://github.com/janezd/baycomp
BitBucket	A website and cloud-based service that helps developers store and manage their code, as well as track and control changes to their code.	https://bitbucket.org/
Conda	Open Source package management system and environment management system.	https://docs.conda.io/
codecarbon (Schmidt et al., 2021b)	Python package estimating and tracking carbon emission of various kind of computer programs.	https://github.com/mlco2/codecarbon
dbpl	Computer science bibliography to find correct versions of papers.	https://dblp.org/
deep-significance (Ulmer, 2021)	Python package implementing the ASO test by Dror et al. (2019) and other utilities	https://github.com/Kaleidophon/deep-significance
European Language Resources Association ELRA (1995)	Public institution for language and evaluation resources	http://catalogue.elra.info/en-us/
GitHub	A website and cloud-based service that helps developers store and manage their code, as well as track and control changes to their de.	https://github.com/
Google Scholar	Scientific publication search engine.	https://scholar.google.com/
Hugging Face Datasets (Lhoest et al., 2021)	Hub to store and share (NLP) datasets	https://huggingface.co/datasets
HyBayes (Azer et al., 2020)	Python package implementing a variety of frequentist and Bayesian significance tests	https://github.com/allenai/HyBayes
LINDAT/CLARIN Váradi et al. (2008)	Open access to language resources and other data and services for the support of research in digital humanities and social sciences	https://lindat.cz/
ONNX	Open format built to represent Machine Learning models.	https://onnx.ai/
Pipenv	Virtual environment for managing Python packages	https://pipenv.pypa.io/
Protocal buffers	Data structure for model predictions	https://developers.google.com/protocol-buffers/
rebiber	Python tool to check and normalize the bib entries to the official published versions of the cited papers.	https://github.com/yuchenlin/rebiber
Semantic Scholar	Scientific publication search engine.	https://www.semanticscholar.org/
Virtualenv	Tool to create isolated Python environments.	https://virtualenv.pypa.io/
Zenodo	General-purpose open-access repository for research papers, datasets, research software, reports, and any other research related digital artifacts	https://zenodo.org/