

Experimental Standards for Deep Learning in Natural Language Processing Research

Dennis Ulmer^② Elisa Bagnagnana^② Max Müller-Eberstein^② Daniel Varab^②
Mike Zhang^② Rob van der Goot^② Christian Hardmeier^② Barbara Plank^{②▲♫}

^②Department of Computer Science, IT University of Copenhagen, Denmark

[▲]Center for Information and Language Processing (CIS), LMU Munich, Germany

[♫]Munich Center for Machine Learning (MCML), 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, compared to more established disciplines, a lack of common experimental standards remains an open challenge to the field at large. Starting from fundamental scientific principles, we distill ongoing discussions on **experimental standards in NLP into a single, widely-applicable methodology**. Following these best practices is crucial to strengthen experimental evidence, improve reproducibility and support scientific progress. These standards are further collected in a public repository to help them transparently adapt to future needs.

1 Introduction

Spurred by the advances in Machine Learning (ML) and Deep Learning (DL), the field of Natural Language Processing (NLP) has seen immense growth over the span of the last ten years, as illustrated by the number of publications in Figure 2. 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 major **obstacles with reproducibility** (Gundersen and 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 text generation (Gehrmann et al., 2022) and with respect to new model architectures (Narang et al., 2021). Others have questioned commonly-accepted procedures (Gorman and 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 and 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

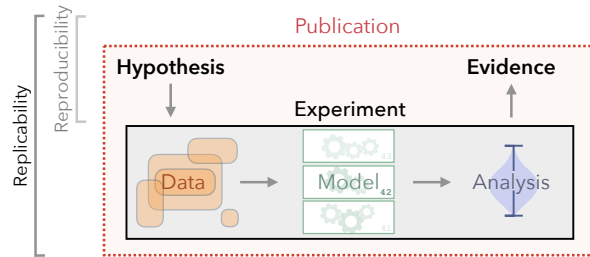


Figure 1: **Visualization of the Scientific Process in Deep Learning.** Uncertainty is introduced at each step, influencing the resulting evidence as well as the documentation required for reproducibility or replicability.

have not gone unnoticed—many of the mentioned works have proposed a **cornucopia** of solutions. In a quickly-moving environment however, keeping track and implementing these proposals becomes challenging. In this work, we weave these open issues together into a cohesive methodology for gathering stronger experimental evidence, that can be implemented with reasonable effort.

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. Using insights from adjacent sub-fields of ML / DL, we extract useful insights to help overcome current challenges with replicability in NLP.

Contributions ① We survey and summarize a wide array of proposals regarding the improvement of the experimental (and publishing) pipeline in NLP research 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 **♦ to indicate that they should be seen as a minimal require-**

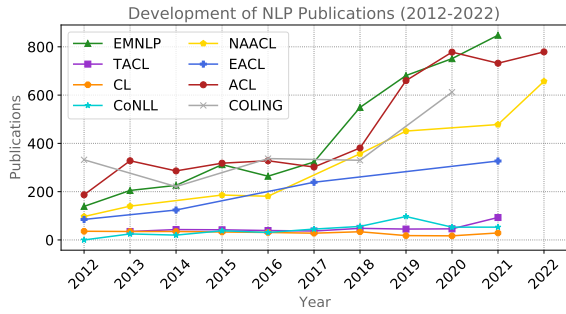


Figure 2: **Development of NLP publications.** Shown is the development of NLP measured by the number of peer-reviewed publications between 2012–2022 based on the data collected by Rei (2022).

ment, and \star for additional recommended actions.

② We create, point to, or supply useful resources to support everyday research activities and improve soundness of research 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 NLP community to facilitate common standards and improve the quality of research.

2 Preliminaries

Our proposed methodology must be built on the scientific principles for generating strong evidence for the general advancement of knowledge, as defined by the following terms:

The Scientific Method Knowledge can be obtained through several ways including theory building, qualitative methods, and empirical research (Kuhn, 1970; Simon, 1995). Here, 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.

¹<https://github.com/Kaleidophon/experimental-standards-deep-learning-research>

²While such hypothesis-driven science is not always applicable or possible (Carroll, 2019), it is a strong common denominator that encompasses most empirical ML research.

In the following, we focus on the **evidence-based evaluation of hypotheses** and how to ensure the **scientific soundness of the experiments** which gave rise to the original empirical evidence, with a focus on **replicability** and **reproducibility**. In computational literature, one term requires access to the original code and data 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 (see also Figure 1).³

Replicability Within DL, we take replicability to mean the exact replication of prior reported evidence. In a **computational environment**, access to the same data, code and tooling should be **sufficient to generate prior results**. However, many factors, such as hardware differences, make exact replication difficult to achieve. 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.

Reproducibility In comparison, we take **reproducibility** to mean the **availability of all necessary and sufficient information such that an experiment’s findings can be independently reaffirmed when the same research question is asked**. As discussed later, the **availability of all components** for replicability is rare—even in a computational setting. An experiment then is reproducible if anyone with access to the publication is able to re-identify the original evidence, i.e. exact results differing, but patterns across experiments being equivalent.

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

³Strikingly, these central terms already lack agreed-upon definitions (Peng, 2011; Fokkens et al., 2013; Liberman, 2015; Cohen et al., 2018), however we follow the prevailing definitions in the NLP community (Drummond, 2009; Dodge and Smith, 2020) as the underlying ideas are equivalent.

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 (Geburu et al., 2020; Bender and Friedman, 2018) aim to standardize metadata for dataset authorship in order to inform future users about assumptions and potential biases during all levels of data collection and annotation—including the

research design (Hovy and Prabhumoye, 2021). Simultaneously, they encourage reflection on whether the authors are adhering to their own 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; Sap et al., 2021). 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, domain differences have a noticeable impact on model performance (White and Cotterell, 2021; Ramesh Kashyap et al., 2021). Increased data diversity can improve the ability of models to generalize to new domains and languages (Benjamin, 2018), however diversity is difficult to quantify (Gong et al., 2019) and full coverage is unachievable. This highlights the importance of documenting representativeness in order to ensure reproducibility—even in absence of the original data. For replicability using the original data, further considerations include long-term storage and versioning, as to ensure equal comparisons in future work (see Appendix A.1 for case studies).

Instance Annotation Achieving high data quality entails that the data must be accurate and relevant for the task to enable effective learning (Pustejovsky and Stubbs, 2012; Tseng et al., 2020) and reliable evaluation (Bowman and Dahl, 2021; Basile et al., 2021). Since most datasets involve human annotation, a **careful annotation design** is crucial (Pustejovsky and Stubbs, 2012; Paun et al., 2022). **Ambiguity in natural language poses inherent challenges and disagreement is genuine** (Basile et al., 2021; Specia, 2021; Uma et al., 2021). As insights into the annotation process are valuable, yet often inaccessible, we recommend to release datasets with individual-coder annotations, as also put forward by Basile et al. (2021); Prabhakaran et al. (2021) and to complement data with insights like statistics on inter-annotator coding (Paun et al., 2022), e.g., over time (Braggaar and van der Goot, 2021), or coder uncertainty (Bassignana and Plank, 2022). When creating new datasets such information strengthens the reproducibility of future findings, as they transparently communicate the inherent variability instead of obscuring it.

Pre-processing 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). Typically, this 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, static test splits are prone to unconscious “overfitting” if they have been in use for a longer period of time, as people aim to beat a particular benchmark (Gorman and 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; Søgaaard 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 and Friedman, 2018);
- ◇ Split the data such as to avoid spurious correlations;
- ◇ Publish the dataset accessibly & indicate changes;
- ★ Perform exploratory data analyses to ensure task adequacy (Caswell et al., 2021);
- ★ Publish the dataset with individual-coder annotations, besides aggregation;
- ★ Claim significance considering the dataset’s statistical power (Card et al., 2020).

4 Codebase & Models

The NLP community has historically taken pride in promoting open access to papers, data, code, and documentation, but some have also noted room for improvement (Wieling et al., 2018; Belz et al., 2020). One practice has been to open-source all components of the experimental procedure in a repository, consisting of all code, necessary scripts, and detailed documentation. The benefit of such a repository is in its ability to enable direct *replication*. In particular, a comprehensive code base directly enables replicability. In practice, such documentation is often communicated through a README file, in which user-oriented information is described.⁴ In DL, full datasets can be large

⁴In Appendix B, we propose minimal requirements for a README file and give pointers on files and code structure.

and impractical to share. Due to their importance however, it is essential to carefully consider how one can share the data with researchers in the future. Therefore, repositories for long-term data storage backed by public institutions should be preferred (e.g., LINDAT/CLARIN by [Váradí et al., 2008](#), more examples in [Appendix C](#)). Nevertheless, practitioners often can not 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 Hyperparameter tuning strategies remain an open area of research (e.g., [Bischi et al., 2021](#)), but are central to the replication of contemporary models. The following rules of thumb exist: Grid search or Bayesian optimization can be applied if few parameters can be searched exhaustively under the computation budget. Otherwise, random search is preferred, as it explores the search space more efficiently ([Bergstra and Bengio, 2012](#)). 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 ([Bischi et al., 2021](#)). To avoid unnecessary guesswork, the following information is expected: Hyperparameters that were searched per model (including options and ranges), the final hyperparameter settings used, number of trials, and settings of the search procedure if applicable. As tuning of hyperparameters is typically performed using specific parts of the dataset, it is essential to note that any modeling decisions based on them automatically invalidate their use as *test* data.

Models Contemporary models (e.g., [Vaswani et al., 2017](#); [Devlin et al., 2019](#); [Dosovitskiy et al., 2021](#); [Chen et al., 2021](#)) 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 The exact model and task evaluation procedure can differ significantly (e.g. [Post, 2018](#)). It is important to either reference the exact evaluation script used (including parameters, citation, and version, if applicable) or include the evaluation script in the codebase. Moreover, to ease error or post-hoc analyses, we highly recommend saving model predictions whenever possible and making them available at publication ([Card et al., 2020](#); [Gehrmann et al., 2022](#)).

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, 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, and can help to collect important information for reproducibility, such as preprocessing and evaluation results. We refer to [Mitchell et al. \(2019\)](#); [Menon et al. \(2020\)](#) for examples of model cards.

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;
- ◇ Publish model predictions and evaluation scripts.;
- ★ Use model cards;
- ★ Publish models;

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](#)). 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. The number of runs should be adapted based on the observed variance: Using for instance 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 and Hayashi, 2003; Tufféry, 2011; Henderson et al., 2018). If the percentage of significant results is low, we should collect more scores.⁵ 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 closer approximation of the true model performance. Nevertheless, any drawn conclusion are still surrounded by a degree of statistical uncertainty, which can be combated by the use of statistical hypothesis testing.

Significance Testing Using deep neural networks, a number of (stochastic) factors such as the random seed (Dror et al., 2019) or even the choice of hardware (Yang et al., 2018) or framework (Leventi-Peetz and Östreich, 2022) can influence performance and need to be taken into account. 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 plenty observations are available. Depending on the application, the usage of specialized tests might furthermore be desirable (Dror et al., 2019; Agarwal et al., 2021). 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 et al., 2022). 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. Due to spa-

⁵The resulting tensions with modern DL hardware requirements are discussed in Section 7.

tial constraints, we here refer to Appendix A.4 for a number of easy-to-use software packages and further reading on the topic.

Critiques & Alternatives Although statistical hypothesis testing is an established tool in many disciplines, its (mis)use has received criticism for decades (Berger and Sellke, 1987; Demšar, 2008; Ziliak and McCloskey, 2008). For instance, Wasserstein et al. (2019) criticize the *p*-value as reinforcing publication bias through the dichotomy of “significant” and “not significant”, i.e., by favoring positive results (Locascio, 2017). Instead, Wasserstein et al. (2019) propose to report it as a continuous value and with the appropriate scepticism.⁶ In addition to statistical significance, another approach advocates for reporting *effect size* (Berger and 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 statistical model producing the observations⁷ 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) and guidelines for reporting details are given by Kruschke (2021), including for instance the choice of model and priors.

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

In this section, we discuss some additional trends in the DL field that researchers should consider when

⁶Or, as Wasserstein et al. (2019) note: “statistically significant—don’t say it and don’t use it”.

⁷Here, 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).

publishing their work, even though they might not directly be related to reproducibility & replicability.

Citation Control While frequently, researchers cite non-archival versions of papers, 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. [Hershcovich et al. \(2022\)](#) advocate for the usage of a *climate performance model card*, in which energy and emission statistics are being detailed. [Strubell et al. \(2019\)](#) approximate financial and environmental costs of training a variety of models (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 CO₂ 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 di-

rect effects on the lives of individuals. 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 ([Waseem et al., 2021](#); [Parmar et al., 2022](#)). As a follow-up, [Waseem et al., 2021](#) argue for a reflection on the consequences the imaginary objectivity of ML has on political choices. [Hovy and Spruit \(2016\)](#) analyze and discuss the social impact research may have beyond the more explored privacy issues. They make an ethical analysis 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

Since previous sections have emphasized the need to overhaul some experimental standards, we ded-

icate this last section to discuss some structural issues that might pose obstacles to this.

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. Only the best-funded research labs can afford the increasing computational costs to account for the statistical uncertainty of results and to reproduce prior works ([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](#)).⁸ 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](#), NLP 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 ([Gundersen and Kjensmo, 2018](#); [Narang et al., 2021](#)). Why is that so? We speculate that the reason for many of these problems are caused by adverse incentives set by the current publishing environment: As the career of researchers hinges on their publications and more rigorous experimental standards are often not required to get published, reproducing and reproducible works are not rewarded. Instead, actors are tempted to “rig the benchmark lottery” ([Dehghani et al., 2021](#)), since achieving state-of-the-art results remains important for publishing ([Birhane et al., 2021](#)). As of now, better experimental standards 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. This state of affairs might still seem like progress to some, but [Chu and 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. Further-

more, [Raff \(2022\)](#) shows that reproducible work can actually have a positive impact on a paper’s citation rates, and thus should be more embraced.

Culture Change How can we change this trend? **As researchers**, we can start implementing the recommendations in this work in order to drive bottom-up change and reach a critical mass ([Centola et al., 2018](#)). **As reviewers**, we can shift focus from results to more rigorous methodologies ([Rogers and Augenstein, 2021](#)), and allow more critiques and reproductions 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 and Augenstein \(2020\)](#) and [Su \(2021\)](#) give recommendations on how to improve the peer-review process by better paper-reviewer matching and paper scoring. Other attempts are currently undertaken to encourage 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 ([Loscascio, 2017](#)). 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 and Evans, 2021](#)).

8 Conclusion

Being able to (re-)produce empirical findings is critical for scientific progress, particularly in fast-growing fields like NLP ([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 and Prabhumoye, 2021](#); [Waseem et al., 2021](#)). This paper aims at providing a toolbox of actionable recommendations *for each research step*, and a reflection and summary of the ongoing broader

⁸E.g., GPT-3’s training was estimated to have cost ca. 12M USD ([Turner, 2020](#)) or 188,702 kWh ([Anthony et al., 2020](#)).

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

discussion. With concrete best practices to raise awareness and call for uptake, we hope to aid researchers in their empirical endeavors.

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, this work was aimed to primarily serve the NLP community, although other disciplines using DL might also profit from these guidelines. With these limitations in mind, we invite members of the community to contribute to our open-source repository.

Ethics Statement

We do not foresee any immediate negative ethical consequences in lieu with our work.

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. MZ and BP are supported by the Independent Research Fund Denmark (DFF) grant 9131-00019B. EB, MME, and BP are supported by the Independent Research Fund Denmark (DFF) Sapere Aude grant 9063-00077B. BP is supported by the ERC Consolidator Grant DIALECT 101043235.

References

- Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. 2021. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 34.
- Lasse F Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint arXiv:2007.03051*.
- Marianna Apidianaki, Saif Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat. 2018. Proceedings of the 12th international workshop on semantic evaluation. In *Proceedings of The 12th International Workshop on Semantic Evaluation*.
- Association for Computational Linguistics. 2022. Reproducibility criteria. <https://2022.naacl.org/calls/papers/#reproducibility-criteria>. Accessed: 2022-02-09.
- Association for Computing Machinery. 2022. Acm code of ethics and professional conduct. <https://www.acm.org/code-of-ethics>. Accessed: 2022-02-10.
- Erfan Sadeqi Azer, Daniel Khashabi, Ashish Sabharwal, and Dan Roth. 2020. Not all claims are created equal: Choosing the right statistical approach to assess hypotheses. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5715–5725. Association for Computational Linguistics.
- Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021. We need to consider disagreement in evaluation. In *1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21. Association for Computational Linguistics.
- Elisa Bassignana and Barbara Plank. 2022. CrossRE: A Cross-Domain Dataset for Relation Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics.
- Christine Basta, Marta R Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39.
- Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2020. ReproGen: Proposal for a shared task on reproducibility of human evaluations in NLG. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 232–236, Dublin, Ireland. Association for Computational Linguistics.
- Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 381–393. Association for Computational Linguistics.
- Alessio Benavoli, Giorgio Corani, Janez Demsar, and Marco Zaffalon. 2017. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *J. Mach. Learn. Res.*, 18:77:1–77:36.
- Alessio Benavoli, Giorgio Corani, Francesca Mangili, Marco Zaffalon, and Fabrizio Ruggeri. 2014. A bayesian wilcoxon signed-rank test based on the dirichlet process. In *International conference on machine learning*, pages 1026–1034. PMLR.

- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:58–04.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.
- Martin Benjamin. 2018. 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*, pages 13–18.
- James O Berger and Thomas Sellke. 1987. Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American statistical Association*, 82(397):112–122.
- James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).
- Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590*.
- Bernd Bischl, Martin Binder, Michel Lang, Tobias Pielok, Jakob Richter, Stefan Coors, Janek Thomas, Theresa Ullmann, Marc Becker, Anne-Laure Boulesteix, et al. 2021. Hyperparameter optimization: Foundations, algorithms, best practices and open challenges. *arXiv preprint arXiv:2107.05847*.
- Carlo Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62.
- Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. 2021. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3.
- Samuel R. Bowman and George Dahl. 2021. [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*, pages 4843–4855, Online. Association for Computational Linguistics.
- Anouck Braggaar and Rob van der Goot. 2021. [Challenges in annotating and parsing spoken, code-switched, Frisian-Dutch data](#). In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 50–58, Kyiv, Ukraine. Association for Computational Linguistics.
- Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. 2011. *Handbook of markov chain monte carlo*. CRC press.
- 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. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. [With little power comes great responsibility](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9263–9274. Association for Computational Linguistics.
- Sean M Carroll. 2019. Beyond falsifiability: Normal science in a multiverse. *Why trust a theory*, pages 300–314.
- Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. [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*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Alahsera 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, Iroko Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhlov, 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. 2021. [Quality at a glance: An audit of web-crawled multilingual datasets](#).
- Damon Centola, Joshua Becker, Devon Brackbill, and Andrea Baronchelli. 2018. Experimental evidence

- for tipping points in social convention. *Science*, 360(6393):1116–1119.
- Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34.
- Johan SG Chu and James A Evans. 2021. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41).
- K Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany J Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névél, Cyril Grouin, and Lawrence E Hunter. 2018. 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, page 156. NIH Public Access.
- Giorgio Corani and Alessio Benavoli. 2015. A bayesian approach for comparing cross-validated algorithms on multiple data sets. *Machine Learning*, 100(2):285–304.
- Alicia Curth, David Svensson, James Weatherall, and Mihaela van der Schaar. 2021. Really doing great at estimating cate? a critical look at ml benchmarking practices in treatment effect estimation.
- Alicia Curth and Mihaela van der Schaar. 2021a. Non-parametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR.
- Alicia Curth and Mihaela van der Schaar. 2021b. On inductive biases for heterogeneous treatment effect estimation.
- 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. 2020. Underspecification presents challenges for credibility in modern machine learning. *CoRR*, abs/2011.03395.
- Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The benchmark lottery.
- Janez Demšar. 2008. On the appropriateness of statistical tests in machine learning. In *Workshop on Evaluation Methods for Machine Learning in conjunction with ICML*, page 65. Citeseer.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.
- Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. [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)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.
- Jesse Dodge and Noah A. Smith. 2020. Reproducibility at emnlp 2020. <https://2020.emnlp.org/blog/2020-05-20-reproducibility>.
- 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. 2021. 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.
- Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. [Replicability analysis for natural language processing: Testing significance with multiple datasets](#). *Trans. Assoc. Comput. Linguistics*, 5:471–486.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. [The hitchhiker’s guide to testing statistical significance in natural language processing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1383–1392. Association for Computational Linguistics.
- Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart. 2020. Statistical significance testing for natural language processing. *Synthesis Lectures on Human Language Technologies*, 13(2):1–116.
- Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. [Deep dominance - how to properly compare deep neural models](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2773–2785. Association for Computational Linguistics.

- Chris Drummond. 2009. Replicability is not reproducibility: Nor is it good science. *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*.
- ELRA. 1995. The european language resources association (elra). <http://www.elra.info/en/about/>. Accessed: 2022-02-10.
- Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. [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)*, pages 1691–1701, Sofia, Bulgaria. Association for Computational Linguistics.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. [Datasheets for datasets](#). *Computing Research Repository*, arxiv:1803.09010. Version 7.
- Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint arXiv:2202.06935*.
- Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. Bayesian data analysis. *Chapman Hall, London*.
- Zhiqiang Gong, Ping Zhong, and Weidong Hu. 2019. Diversity in machine learning. *IEEE Access*, 7:64323–64350.
- Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2786–2791.
- Sander Greenland, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. 2016. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31(4):337–350.
- Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don’t stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Christian Hardmeier, Marta R. Costa-jussà, Kellie Webster, Will Radford, and Su Lin Blodgett. 2021. [How to write a bias statement: Recommendations for submissions to the workshop on gender bias in NLP](#). *CoRR*, abs/2104.03026.
- Jianxing He, Sally L Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. 2019. The practical implementation of artificial intelligence technologies in medicine. *Nature medicine*, 25(1):30–36.
- Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43.
- Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. [Deep reinforcement learning that matters](#). In *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*, pages 3207–3214. AAAI Press.
- Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. Towards climate awareness in nlp research. *arXiv preprint arXiv:2205.05071*.
- Sara Hooker. 2021. The hardware lottery. *Communications of the ACM*, 64(12):58–65.
- Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.
- Dirk Hovy and Shannon L. Spruit. 2016. [The social impact of natural language processing](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.
- Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. [Adversarial examples are not bugs, they are features](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 125–136.
- John P. A. Ioannidis. 2005. [Why most published research findings are false](#). *PLOS Medicine*, 2(8):null.
- Nathalie Japkowicz and Mohak Shah. 2011. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press.
- S Jean-Paul, T Elseify, I Obeid, and J Picone. 2019. Issues in the reproducibility of deep learning results. In *2019 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–4. IEEE.

- Theis Ingerslev Jensen, Bryan T Kelly, and Lasse Heje Pedersen. 2021. Is there a replication crisis in finance? Technical report, National Bureau of Economic Research.
- Leslie K John, George Loewenstein, and Drazen Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5):524–532.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. [SpanBERT: Improving pre-training by representing and predicting spans](#). *Transactions of the Association for Computational Linguistics*, 8:64–77.
- 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. 2021. [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*, pages 4110–4124, Online. Association for Computational Linguistics.
- Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2020. [WILDS: A benchmark of in-the-wild distribution shifts](#). *CoRR*, abs/2012.07421.
- John K Kruschke. 2010. Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(5):658–676.
- John K Kruschke. 2013. Bayesian estimation supercedes the t test. *Journal of Experimental Psychology: General*, 142(2):573.
- John K Kruschke. 2021. Bayesian analysis reporting guidelines. *Nature human behaviour*, 5(10):1282–1291.
- John K Kruschke and Torrin M Liddell. 2018. Bayesian data analysis for newcomers. *Psychonomic bulletin & review*, 25(1):155–177.
- Thomas S Kuhn. 1970. *The structure of scientific revolutions*, volume 111. Chicago University of Chicago Press.
- Andrew Kyle Lampinen, Stephanie CY Chan, Adam Santoro, and Felix Hill. 2021. Publishing fast and slow: A path toward generalizability in psychology and ai.
- A-M Leventi-Peetz and T Östreich. 2022. Deep learning reproducibility and explainable ai (xai). *arXiv preprint arXiv:2202.11452*.
- 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. 2021. [huggingface/datasets: 1.12.1](#).
- Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Roshtamizadeh, and Ameet Talwalkar. 2017. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816.
- Mark Liberman. 2015. Replicability vs. reproducibility — or is it the other way around? <https://languagelog.ldc.upenn.edu/n11/?p=21956>. Accessed: 2022-02-21.
- Mingfeng Lin, Henry C Lucas Jr, and Galit Shmueli. 2013. Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research*, 24(4):906–917.
- Joseph J Locascio. 2017. Results blind science publishing. *Basic and applied social psychology*, 39(5):239–246.
- Eric L Manibardo, Ibai Laña, and Javier Del Ser. 2021. Deep learning for road traffic forecasting: Does it make a difference? *IEEE Transactions on Intelligent Transportation Systems*.
- Christopher D. Manning. 2015. [Last words: Computational linguistics and deep learning](#). *Computational Linguistics*, 41(4):701–707.
- Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. [Scientific credibility of machine translation research: A meta-evaluation of 769 papers](#). In *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*, pages 7297–7306. Association for Computational Linguistics.
- Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. 2020. 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*, pages 2437–2445.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

- Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4):659–684.
- Saif M. Mohammad. 2020. [What is a research ethics statement and why does it matter?](#)
- Saif M. Mohammad. 2021. [Ethics sheets for AI tasks](#). *CoRR*, abs/2107.01183.
- Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. 2021. [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.
- 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. 2021. [Do transformer modifications transfer across implementations and applications?](#) In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5758–5773. Association for Computational Linguistics.
- Adrian Nilsson, Simon Smith, Gregor Ulm, Emil Gustavsson, and Mats Jirstrand. 2018. A performance evaluation of federated learning algorithms. In *Proceedings of the second workshop on distributed infrastructures for deep learning*, pages 1–8.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. [Universal Dependencies v2: An evergrowing multilingual treebank collection](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.
- Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606.
- Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2022. Don’t blame the annotator: Bias already starts in the annotation instructions. *arXiv preprint arXiv:2205.00415*.
- Silviu Paun, Ron Artstein, and Massimo Poesio. 2022. Statistical methods for annotation analysis. *Synthesis Lectures on Human Language Technologies*, 15(1):1–217.
- Ted Pedersen. 2008. [Last words: Empiricism is not a matter of faith](#). *Computational Linguistics*, 34(3):465–470.
- Roger D Peng. 2011. Reproducible research in computational science. *Science*, 334(6060):1226–1227.
- Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54.
- Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. DaN+: Danish nested named entities and lexical normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6649–6662, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Karl Popper. 1934. *Karl Popper: Logik der Forschung*. Mohr Siebeck, Tübingen, Germany.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. [On releasing annotator-level labels and information in datasets](#). In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O’Reilly Media, Inc.
- Edward Raff. 2022. Does the market of citations reward reproducible work? *arXiv preprint arXiv:2204.03829*.
- Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, and Roger Zimmermann. 2021. [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*, pages 1830–1849, Online. Association for Computational Linguistics.
- Sebastian Raschka. 2018. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.
- Marek Rei. 2022. [ML and nlp publications in 2021](#). <https://www.marekrei.com/blog/ml-and-nlp-publications-in-2021/>.
- Kenneth Reitz and Tanya Schlusser. 2016. *The Hitchhiker’s guide to Python: best practices for development*. " O’Reilly Media, Inc."

- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond accuracy: Behavioral testing of NLP models with CheckList](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
- Stefan Riezler and Michael Hagmann. 2021. Validity, reliability, and significance.
- Stefan Riezler and John T Maxwell III. 2005. 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*, pages 57–64.
- Anna Rogers and Isabelle Augenstein. 2020. [What can we do to improve peer review in nlp?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1256–1262. Association for Computational Linguistics.
- Anna Rogers and Isabelle Augenstein. 2021. How to review for acl rolling review? <https://aclrollingreview.org/reviewertutorial>. Accessed: 2022-02-21.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.
- David Schlangen. 2021. [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)*, pages 670–674, Online. Association for Computational Linguistics.
- Victor Schmidt, Kamal Goyal, Aditya Joshi, Boris Feld, Liam Conell, Nikolas Laskaris, Doug Blank, Jonathan Wilson, Sorelle Friedler, and Sasha Lucioni. 2021. [CodeCarbon: Estimate and Track Carbon Emissions from Machine Learning Computing](#).
- Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. [Green AI](#). *Commun. ACM*, 63(12):54–63.
- Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D’Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, et al. 2021. The multiberts: Bert reproductions for robustness analysis. *arXiv preprint arXiv:2106.16163*.
- Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. 2022. Compute trends across three eras of machine learning. *arXiv:2202.05924 [cs]*. ArXiv: 2202.05924.
- Anastasia Shimorina, Yannick Parmentier, and Claire Gardent. 2021. An error analysis framework for shallow surface realization. *Transactions of the Association for Computational Linguistics*, 9:429–446.
- Herbert A Simon. 1995. Artificial intelligence: an empirical science. *Artificial Intelligence*, 77(1):95–127.
- Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. 2012. [Practical bayesian optimization of machine learning algorithms](#). In *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*, pages 2960–2968.
- Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. [We need to talk about random splits](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1823–1832. Association for Computational Linguistics.
- Lucia Specia. 2021. [Disagreement in human evaluation: blame the task not the annotators](#). NoDaLiDa keynote.
- Emma Strubell, Ananya Ganesh, and Andrew McCalum. 2019. [Energy and policy considerations for deep learning in NLP](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.
- Weijie Su. 2021. You are the best reviewer of your own papers: An owner-assisted scoring mechanism. *Advances in Neural Information Processing Systems*, 34.
- TMLR. 2022. Submission guidelines and editorial policies. <https://jmlr.org/tmlr/editorial-policies.html>. Accessed: 2022-02-09.
- Tina Tseng, Amanda Stent, and Domenic Maida. 2020. [Best practices for managing data annotation projects](#). *CoRR*, abs/2009.11654.
- Stéphane Tufféry. 2011. *Data mining and statistics for decision making*. John Wiley & Sons.
- Elliot Turner. 2020. Twitter post (@eturner303): Reading the openai gpt-3 paper. <https://twitter.com/eturner303/status/1266264358771757057>. Accessed: 2022-02-09.
- Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2022. deep-significance-easy and meaningful statistical significance testing in the age of neural networks. *arXiv preprint arXiv:2204.06815*.

- Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Rob van der Goot. 2021. [We need to talk about train-dev-test splits](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4485–4494. Association for Computational Linguistics.
- Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. [Mas-sive choice, ample tasks \(MaChAmp\): A toolkit for multi-task learning in NLP](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.
- Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. 2021. [Preregistering NLP research](#). In *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*, pages 613–623. Association for Computational Linguistics.
- Daniel Varab and Natalie Schluter. 2020. DaNews-room: A large-scale Danish summarisation dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6731–6739, Marseille, France. European Language Resources Association.
- Daniel Varab and Natalie Schluter. 2021. [Mas-siveSumm: a very large-scale, very multilingual, news summarisation dataset](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tamás Váradi, Peter Wittenburg, Steven Krauwer, Martin Wynne, and Kimmo Koskenniemi. 2008. Clarin: Common language resources and technology infrastructure. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Zeeraak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. [Disembodied machine learning: On the illusion of objectivity in nlp](#).
- Ronald L Wasserstein, Allen L Schirm, and Nicole A Lazar. 2019. Moving to a world beyond “ $p < 0.05$ ”.
- Junyi Wei, Yicheng Zhang, Zhe Zhou, Zhou Li, and Mohammad Abdullah Al Faruque. 2020. 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)*, pages 125–137. IEEE.
- Jennifer C. White and Ryan Cotterell. 2021. [Examining the inductive bias of neural language models with artificial languages](#). In *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*, pages 454–463. Association for Computational Linguistics.
- Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Reproducibility in computational linguistics: are we willing to share? *Computational Linguistics*, 44(4):641–649.
- Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer.
- 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. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Jie Yang, Shuailong Liang, and Yue Zhang. 2018. [Design challenges and misconceptions in neural sequence labeling](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3879–3889, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhenguo Li. 2021. Ood-bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. *arXiv preprint arXiv:2106.03721*.
- Hongkun Yu, Chen Chen, Xianzhi Du, Yeqing Li, Abdullah Rashwan, Le Hou, Pengchong Jin, Fan Yang, Frederick Liu, Jaeyoun Kim, and Jing Li. 2020. TensorFlow Model Garden. <https://github.com/tensorflow/models>.
- Ke-Hai Yuan and Kentaro Hayashi. 2003. 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.

Mike Zhang and Barbara Plank. 2021. [Cartography active learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 395–406. Association for Computational Linguistics.

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

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

Shi Zong, Ashutosh Baheti, Wei Xu, and Alan Ritter. 2020. [Extracting COVID-19 events from twitter](#). *CoRR*, 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 and 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 and 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 their 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 Applications¹⁰ 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

¹⁰<https://keras.io/api/applications/>

project structure in [Appendix B](#). Here, we give examples going from smaller projects to larger Python projects:

The codebase of CateNETS (Curth and van der Schaar, 2021b,a; 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 and Hagmann (2021) provide textbooks around hypothesis testing in an NLP context. Japkowicz and Shah (2011) describe the usage of statistical test for general, classical ML classification algorithms. When it comes to usage, Zhang and 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 and Maxwell III, 2005) between all the competitors in an argument reasoning comprehension shared task and interpret them with the appropriate degree of carefulness.

¹¹<https://github.com/AliciaCurth/CATENets>

¹²<https://github.com/machamp-nlp/machamp>

¹³<https://github.com/Adapter-Hub/adapter-transformers>

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 and 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 and Benavoli, 2015). Applications can be found for instance by Nilsson et al. (2018), who use the Bayesian correlated t-test (Corani and 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.

Hershcovich et al. (2022), instead, propose a *climate performance model card* as a way to systematically report the climate impact of NLP research.

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 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 and 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](#).

¹⁴See for instance <https://robvander.github.io/blog/repro.htm>

¹⁵Some 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
ACL Anthology	Website hosting all the published proceedings of the ACL.	https://aclanthology.org
ACL pubcheck	Tool to check the format and the citations of papers written with the ACL style files.	https://github.com/acl-org/aclpubcheck
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., 2021)	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 et al., 2022)	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. Note that the ACL Anthology should be preferred, as Google Scholar often indexes the first occurrence of a paper (which is frequently a pre-print)	https://scholar.google.com/
Hugging Face Datasets (Lhoest et al., 2021)	Hub to store and share 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 (Váradí 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/
Protocol 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/