Reproducibility in NLP: What Have We Learned from the Checklist?

Ian Magnusson[♠] **Noah A. Smith**[♠◇] **Jesse Dodge**[♠]
[♠]Allen Institute for Artificial Intelligence
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington
{ianm,noah,jessed}@allenai.org

Abstract

Scientific progress in NLP rests on the reproducibility of researchers' claims. The *CL conferences created the NLP Reproducibility Checklist in 2020 to be completed by authors at submission to remind them of key information to include. We provide the first analysis of the Checklist by examining 10,405 anonymous responses to it. First, we find evidence of an increase in reporting of information on efficiency, validation performance, summary statistics, and hyperparameters after the Checklist's introduction. Further, we show acceptance rate grows for submissions with more YES responses. We find that the 44% of submissions that gather new data are 5% *less* likely to be accepted than those that did not; the average reviewer-rated reproducibility of these submissions is also 2% lower relative to the rest. We find that only 46% of submissions claim to open-source their code, though submissions that do have 8% higher reproducibility score relative to those that do not, the most for any item. We discuss what can be inferred about the state of reproducibility in NLP, and provide a set of recommendations for future conferences, including: a) allowing submitting code and appendices one week after the deadline, and b) measuring dataset reproducibility by a checklist of data collection practices.

1 Introduction

Reproducibility is a foundational component of scientific progress. NLP systems are complex, and even when their behavior is carefully measured, incentives to publish quickly and limitations in the publishing process can lead to underreporting of information necessary for reproducible science. The ramifications of this extend beyond the research community; the audience of NLP papers published years ago was largely other NLP researchers, but today the world is watching developments in the field, looking for advances that will lead to broadly-adopted applications. As the impact of NLP grows, so too do the consequences of reproducibility in our field.

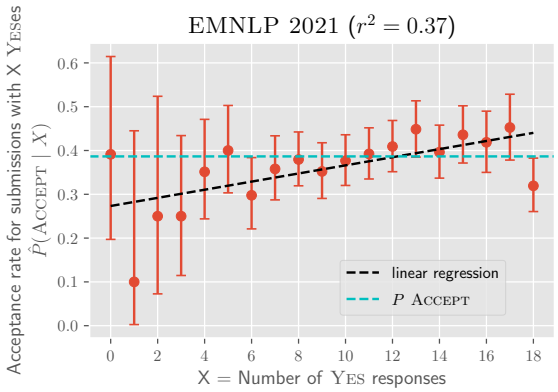

Figure 1: Submissions to EMNLP 2021 binned by count of YES responses to the NLP Reproducibility Checklist items. The ACCEPT rate is given for each bin. Papers with more YES responses are more likely to be accepted, except those that mark YES to all checklist items, which we hypothesize contain responses which do not accurately represent the associated paper.

Of course, NLP is not the first field to evaluate reproducibility; some have even described a "reproducibility crisis" in science (Aarts et al., 2015; Baker, 2016). One tool designed to improve reproducibility is a checklist filled out at paper submission time. Such a checklist can descriptively remind authors of relevant information to report, while preserving the freedom for authors to do so however they see fit. For example, the journal *Nature* requires authors fill out a Reporting Checklist for Life Sciences Articles (Nature, 2018). In 2019 NeurIPS started to require that submissions fill out the ML Reproducibility Checklist (Pineau et al., 2021), partly inspired by the *Nature* checklist, and in 2021 AAAI required their own checklist.[1] CVPR, ICCV, ECCV, and IJCAI provide a checklist but do not collect responses.

In this work, we provide the first analysis of

[1] aaai.org/Conferences/AAAI-21/aaai21call/

Findings of the Association for Computational Linguistics: ACL 2023, pages 12789–12811

the NLP Reproducibility Checklist (Dodge et al., 2019). We have gathered 10,405 anonymized responses from EMNLP 2020 and 2021, NAACL 2021, and ACL 2021. For the latter two we are also able to obtain reviewer scores, reproducibility judgements, and feedback on the Checklist.

Our findings include: (1) Most checklist items are frequently reported, and submissions reporting them are more often accepted and perceived as reproducible. (2) Submissions that collect new data are accepted less and viewed as less reproducible, and these gaps are not explained by non-reporting of any current Checklist items. (3) Only about half of submissions report open sourcing code and many that do not also lack reporting on efficiency measures and even evaluation metrics. (4) A majority of reviewers describe the checklist as useful, and by contrasting responses to observed rates prior to the Checklist we evidence a possible increase in reporting. We conclude with a discussion of what can be inferred from these findings about the state of reproducibility in NLP and offer recommendations to address the gaps we have measured.

2 The NLP Reproducibility Checklist

The NLP Reproducibility Checklist was originally introduced by Dodge et al. (2019). Each item on the checklist is phrased as a statement, like "The number of parameters in each model," and authors can mark YES if they include that information in their paper, NO if they do not include it in their paper, or N/A if that information does not make sense for their submission (e.g., they do not use any models to report parameter counts for). The checklist items were a part of the submission form, and it was required that authors fill it out to submit their paper.[2] Thus, the checklist responses act as a (self-reported) overview of the contents of papers submitted to NLP conferences. Importantly, authors were not required to include any information in their papers, they were only required to indicate whether or not they did include information. Answers were made available to reviewers, who were expressly asked to assess the reproducibility of the work. The filled checklists were not released with the published papers.

3 Data and Methodology

In Table 1 we list the Checklist items and the abbreviations we will use for them throughout this

[2]NAACL 2021 permitted authors to leave items BLANK.

	Abbreviation	Full Checklist Item
All Results	MODELDESCRIPTION	A clear description of the mathematical setting, algorithm, and/or model
	LINKTOCODE	A link to a downloadable source code, with specification of all dependencies, including external libraries
	INFRA	A description of computing infrastructure used
	RUNTIME	Average runtime for each approach
	PARAMETERS	The number of parameters in each model
	VALIDATIONPERF	Corresponding validation performance for each reported test result
	METRICS	Explanation of evaluation metrics used, with links to code
Multiple Experiments	NOTRAININGEVALRUNS	The exact number of training and evaluation runs
	HYPERBOUND	Bounds for each hyperparameter
	HYPERBESTCONFIG	Hyperparameter configurations for best-performing models
	HYPERSEARCH	Number of hyperparameter search trials
	HYPERMETHOD	The method of choosing hyperparameter values (e.g., uniform sampling, manual tuning, etc.) and the criterion used to select among them (e.g., accuracy)
	EXPECTEDPERF	Summary statistics of the results (e.g., mean, variance, error bars, etc.)
Datasets	DATASTATS	Relevant statistics such as number of examples
	DATASPLIT	Details of train/validation/test splits
	DATAPROCESSING	Explanation of any data that were excluded, and all pre-processing steps
	DATADOWNLOAD	A link to a downloadable version of the data
	NEWDATADESCRIPTION	For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control
	DATALANGUAGES	For natural language data, the name of the language(s)

Table 1: Checklist abbreviations and standardized phrasing. Phrasing per conference in Table 6 (Appendix).

Conference	Sub	Wdrn	MAIN	FINDINGS
EMNLP 2020	3,666	660	24.9%	14.8%
EMNLP 2021	4,815	1,555	25.8%	12.9%
NAACL 2021	1,797	565	38.7%	N/A
ACL 2021	3,377	470	24.4%	15.7%
Overall	13,655	3,250	39.4%	

Table 2: **Sub**missions, **Wdr**awn/Desk-Rejects, and MAIN conference and FINDINGS acceptance rates in our data.

paper. There are three categories of items: (1) for all reported experimental results, (2) for results involving multiple experiments, like hyperparameter search, and (3) for all datasets used. 16 of 19 items appeared in all four conferences. We compare specific checklist items between the ML Reproducibility Checklist and the NLP Reproducibility Checklist conference variations in Appendix A.1. Full phrasing for each conference is listed in Table 6 (Appendix).

Our data includes, for a given submission, the checklist responses (YES, NO, N/A for each item), MAIN + FINDINGS acceptance status (ACCEPT $\in$ {accepted, rejected}), and the TRACK. No data includes any deanonymizing information, such as authors or paper titles. For NAACL 2021 and ACL 2021, we have the following metadata for each review: overall recommendation score ("Should this paper be accepted to <conference name>?") averaged to AVGREC $\in [1, 5]$, perceived reproducibility score ("How do you rate the paper's reproducibility? Will members of the ACL community be able to reproduce or verify the results in this paper?") averaged to AVGREPROD $\in [1, 5]$ or N/A if any reviewer responds N/A, and reproducibility checklist feedback ("Are the authors' answers to the Reproducibility Checklist useful for evaluating the submission?") aggregated by majority vote to CHECKLISTFEEDBACK $\in$ {Not useful, Somewhat useful, Very useful}.[3]

As shown in Table 2, there were a total of 13,655 submissions across the four conferences. We remove all withdrawn and desk-rejected submissions from analysis,[4] comprising $3,250$ submissions (23.8% of the data), leaving a total of $10,405$ submissions for analysis.

We recognize that the checklist responses are self-reported information, and thus in some cases might not be accurate representations of the associated submission (e.g., authors may mark YES to an item on the checklist when in fact their paper does not include that information). We discuss this in Appendix A.2.

To the best of our knowledge, the creators of the checklist indicated that the data would not be made public; while we currently do not plan to fully open source the data, the data can be made available upon request (and we welcome feedback on this policy).[5]

In all analyses, error bars represent 95% confidence intervals. These are computed by Clopper–Pearson interval for binary values and bootstrap for continuous values, both using `scipy` version 1.9.1 (Virtanen et al., 2020). All comparisons of differences in results are absolute differences unless explicitly stated as relative.

4 What Can We Learn About How Reproducibility Already Works?

We begin by measuring current practice, according to the (self-reported) Checklist data. Across all items and conferences, 62.7% of responses were YES. Figure 4 shows that most items are reported in most submissions. Moreover, we can measure reviewers' perception of reproducibility as well as differences in rates of reporting for items among papers that do and do get accepted by *CL review.

More YES responses to checklist items associate with higher acceptance. In Figures 1 and 2 we show positive associations between answering more items as YES and ACCEPT rate. Each point in these figures represents the ACCEPT rate among all the submissions with the same number of YES responses among items. We regress the ACCEPT rate on a single variable counting checklist items answered YES for a submission. When pooling responses across all shared questions on all conferences $r^2 = 0.53$.[6] Notably, submissions with YES responses to all items are consistently below the trend. We hypothesize in Appendix A.2 that these submissions include responses which do not accurately represent the associated paper; recall that authors were required to fill out the checklist in order to submit, so marking the same response to all items is, in some sense, as close as they can get to not filling it out. The lower acceptance rate suggests that reviewers are not scoring papers based on the responses to the checklist itself, instead evaluating the contents of the paper, as intended.

Reviewer assessed reproducibility associates with acceptance rate. In Figure 3 we compare ACCEPT rate across quantiles of AVGREPROD and AVGREC. Though ACCEPT rate grows much more

[3] Full reviewer instructions available at `2021.naacl.org/downloads/NAACL2021-Review-Form.pdf` and `2021.aclweb.org/downloads/Review_Form.pdf`

[4] Withdrawn and desk-rejected submissions lack reviews and include blank test submissions and place holders.

[5] "The filled-out checklist will not be released with the published version of an accepted paper, it is meant as a tool for authors and reviewers," (Dodge and Smith, 2020).

[6] The r^2 value for these regressions ranges across the conferences from 0.22 to 0.46, and the trend is consistently positive.

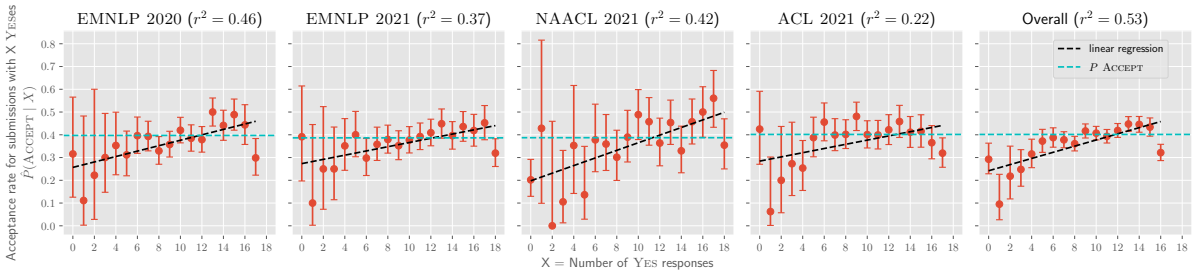

Figure 2: Accept rate among submissions binned by count of Yes responses. Yes response count and Accept rate trend consistently positive. All-Yes responses are notably below trend, as discussed in Appendix A.2.

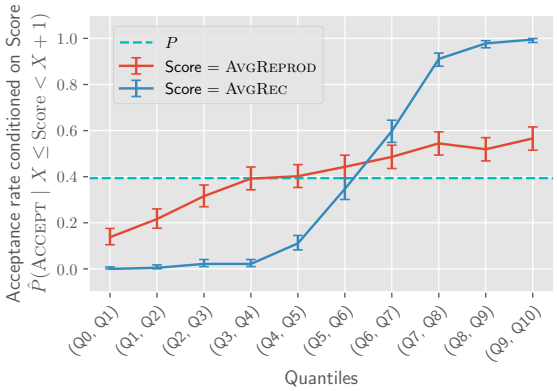

Figure 3: Accept rates across quantiles for perceived reproducibility (AvgReprod) and overall recommendation (AvgRec) for NAACL and ACL 2021. Perceived reproducibility trends positively with acceptance.

slowly for AvgReprod than AvgRec, reviewers assessment of reproducibility is still evidently associated with acceptance.

In all but three checklist items, Yes responses are associated with higher Accept rates. Figure 5.A presents Accept rates conditioned on a given response. Yes responses receive a 0.9% higher rate than the overall average, while No and N/A receive 0.7% and 1.1% lower rates respectively. Figure 5.B presents the three exceptions where answering Yes to that checklist item receives a lower than average rate. These are discussed in detail in Section 5 and Appendix A.3.

In all but one checklist item, Yes responses are associated with higher AvgReprod scores. Figure 6.A shows the mean AvgReprod score conditioned on a given response in NAACL 2021 and ACL 2021. Reassuringly, Yes responses receive 0.04 higher scores than average, while No or N/A score 0.04 and 0.04 lower than average, respectively. Figure 6.B shows that LinkToCode has the highest score, 0.18 above average. We also highlight NewDataDescription as it is the only item with a lower than average score when answered Yes. This exception is discussed further in Section 5, where we hypothesize this reflects a lower than average perceived reproducibility of submissions presenting new data.

5 The Data Collection Gap

Natural language processing has long been a field driven by data. A body of work has proposed best practices for documenting the characteristics and creation of datasets (Bender and Friedman, 2018; Gebru et al., 2018; Hutchinson et al., 2020; Dodge et al., 2021; Rogers et al., 2021; Pushkarna et al., 2022). Among other concerns, such documentation is critical for the difficult task of dataset reproduction (Recht et al., 2019, *inter alia*). From the checklist item NewDataDescription, which asks that collection is described if new data is presented, we find that 38.2% and 6.3% of submissions mark Yes and No, respectively. This implies that 44.5% of submissions to our NLP conferences collect new data; if almost half of submissions collect new data, we argue that data collection and dissemination practices deserve further attention. This also highlights clear room for improvement in the community: 14.1% of submissions that collect new data do not describe how it was collected, totalling 650 papers.

Submissions with new data have lower than average Accept rate and AvgReprod scores. Alarmingly, submissions that collect new data (i.e., submissions that mark Yes or No to NewDataDescription) have a 5.1% lower acceptance rate than those that do not (i.e., mark N/A or Blank). A low acceptance rate for answering No to NewDataDescription would, by itself, be encouraging, perhaps indicating that reviewers expect data collection to be well documented. However, Figure 5B shows that, even when answering Yes

Figure 4: YES response rate per item. Most items are reported for most submissions. Note that NAACL 2021 respondents were able to leave questions BLANK. Other answers shown in Figure 12 (Appendix).

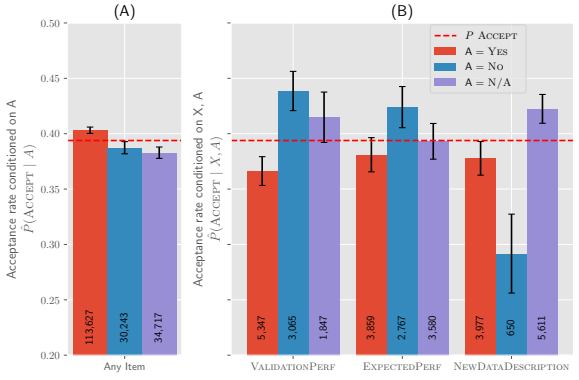

Figure 5: ACCEPT rates over all conferences for submissions with a given response. (A) shows rate conditioned on response regardless of item. (B) shows the only items where YES ACCEPT rates are below average. Total count of each response is shown on the bar. Items with higher NO acceptance than YES could indicate the community has not fully embraced these practices as norms. High acceptance rate for NEWDATADESCRIPTION N/A indicates that papers that do not collect data are more likely to be published.

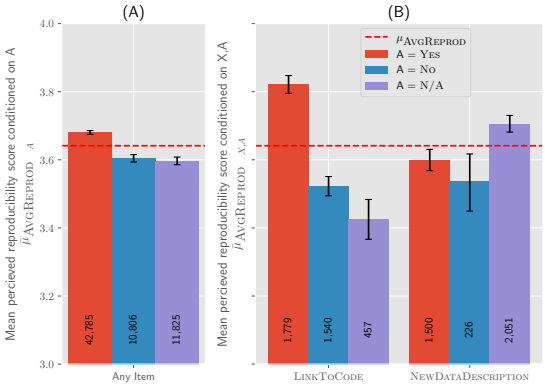

Figure 6: Reviewer perceived reproducibility score (AVGREPROD $\in [1, 5]$) for submissions with a given response from NAACL and ACL 2021, excluding ones with N/A AVGREPROD. (A) shows score conditioned on response regardless of item. (B) shows the items with highest (LINKTOCODE) and lowest (NEWDATADESCRIPTION) YES score. Total count of each response is shown on the bar. NEWDATADESCRIPTION is the only item with a below average YES score.

to describing the data collection process, ACCEPT rate is 1.6% lower than average. Meanwhile, there is a similar gap in AVGREPROD. Scores for submissions that collect data are lower by 2.4% relative to those that do not. Again, this is not limited to submissions that fail to describe the data collection process. Figure 6B reveals that the mean score over submissions that do describe their data collection is 0.04 below the mean of all submissions. When considering only accepted papers, however, the gap in AVGREPROD disappears. The review process ends up with accepted dataset papers with similar AVGREPROD to non-dataset papers, but along the way many more dataset than non-dataset submissions are rejected and those have lower AVGREPROD. We hypothesize that these phenomena arise both because dataset papers may indeed be more challenging to (re)produce and also because of the persistent (and problematic) tendency to value modeling over data collection (Rogers, 2021).

High compliance among the dataset checklist items does not reveal the source of the ACCEPT rate and AVGREPROD gap. DATASTATS, DATASPLIT, DATADOWNLOAD, and DATALANGUAGES receive the highest rates of reporting other than MODELDESCRIPTION and METRICS. This only grows when looking just at submissions presenting new datasets, reaching 97.3%, 91.7%, 91.4%, and 86.6% respectively, and NEWDATADESCRIPTION is also reported in 86.0% of these submissions. Un-

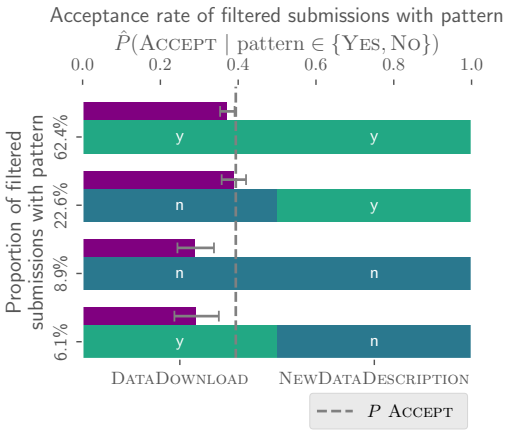

Figure 7: Proportion (row labels) and Accept rates (horizontal purple bars) for all response patterns on dataset availability and creation (excluding instances where either item is N/A or Blank). Nearly 1 in 11 of these neither share the data nor describe its collection, yet 28.9% of those are accepted.

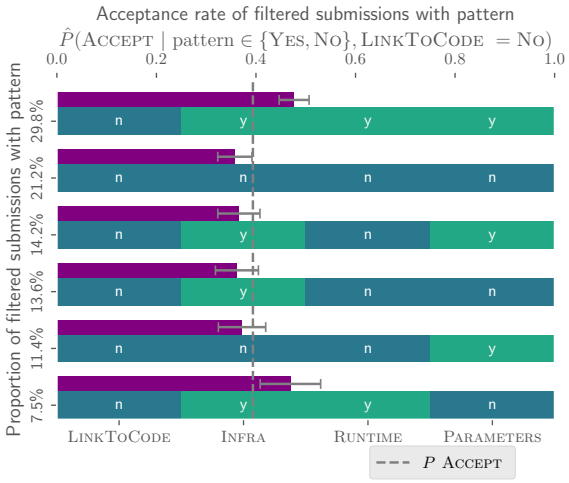

Figure 8: Proportion (row labels) and Accept rates (horizontal purple bars) for efficiency response patterns with > 100 submissions when LinkToCode is No and no responses are N/A or Blank. More than 1 in 5 of these do not report any efficiency items, which are difficult to infer without source code.

like the other dataset items, DataDownload is less frequently reported, but its occurrence and associated Accept rates and AvgReprod scores are similar whether considering submissions presenting new data or not. This suggests that additional checklist items for data collection should be introduced to measure where this gap in perceived reproducibility is coming from.

28.9% of submissions with new data do not provide a downloadable version of the data. More generally, a clear area for improvement is that 25.3% of submissions overall answer No to DataDownload; providing a link to download a dataset is still important for previously released datasets as it might be ambiguous which version of a dataset was used. But for newly collected data, answering No to DataDownload implies the data is not publicly available at all. Moreover, when DataDownload is No, the rate of submissions reporting the collection process in NewDataDescription drops 14.2%. Figure 7 further reveals the interaction between DataDownload and NewDataDescription. When a new dataset submission provides neither a description of the data collection process nor access to the data itself, this leaves very little for reviewers to assess, at least with regard to the data contributions of the paper. Yet 28.9% of these papers are accepted.

6 Code (Un)availability

We see that on average 45.9% of submissions report linking to code (47.5% for accepted papers). We see in Figure 6 that whether submissions answer LinkToCode as Yes or No has the largest difference in AvgReprod scores, with a gap of 0.30. Yet Accept rates for submissions with or without LinkToCode are nearly the same.

We find similar rates of links to code as at ML conferences. Pineau et al. (2021) reveal a 38.8% self-reported rate of code availability at submission time for NeurIPS 2019. They find this number drops to 27.7% when checked by at least one reviewer. Extrapolating from this false reporting rate, the true code availability rate among accepted papers in our data might be 32.8%. Meanwhile, a study on ICML 2019 by Chaudhuri and Salakhutdinov (2019) finds 36% of submitted and 43% of accepted papers have code at submission time, though it unclear if these are self-reported.

Previous efforts to measure camera-ready code availability have found widely different rates than our reported LinkToCode at submission time. Unfortunately our data does not cover code availability at camera-ready, except insofar as some authors may interpret this checklist item to permit promises to later release code. 24.3% of papers at NAACL 2022 opted in to submitting a code link to the Reproducibility Track

and received an Open Source Code badge.[7] We recognize this was optional for authors, and thus it is likely the case that the true number of camera-ready papers that included a link to code was higher. The studies mentioned before found that 74.4% and 64% of camera-ready papers had links to code at NeurIPS 2019 and ICML 2019. Narrowing the range of these measurements should be a worthwhile effort, as these studies found code being available *during* review was useful in 1,315 reviews in NeurIPS 2019, and 18.3% of ICML 2019 reviewers surveyed were able to look at code and found it useful.

Items on compute efficiency are completely reported in only 29.8% of submissions without code. Figure 8 shows patterns for these efficiency items that occurred more than 100 times. While ACCEPT rates are somewhat lower when items are not reported, 21.2% of these without-code submissions report none of the efficiency measures. There may be unavoidable impediments to making code available, such as intellectual property. But in this case even greater emphasis should be placed on reporting efficiency measures, as estimating these without code is quite difficult. Similarly 19.6% of submissions with no code report NO to explaining METRICS which may render evaluations irrecoverably ambiguous if there are varying implementations of a metric.

7 How Effective is the Checklist?

Dodge and Smith (2020) describe the Checklist as intended to improve "reporting of the setup and results of the experiments that authors have conducted." Though self-reported data do not directly answer this question, we find potential evidence of such an improvement. Diachronic analysis also shows that reporting rates may have stagnated after initial improvement. We also examine reviewer and author views on the Checklist.

Compared against manually checked data from before the Checklist introduction, our data shows increases in 8 of 10 items. Figure 9 shows rates of a subset of items that were manually checked by Dodge et al. (2019) in 50 papers sampled from EMNLP 2018. The self-reported rates available in our data are not ideal comparisons as they likely overestimate. However, the EMNLP 2018 sample may also overestimate, as only "experimental results" are included for which

[7] naacl2022-reproducibility-track.github.io/

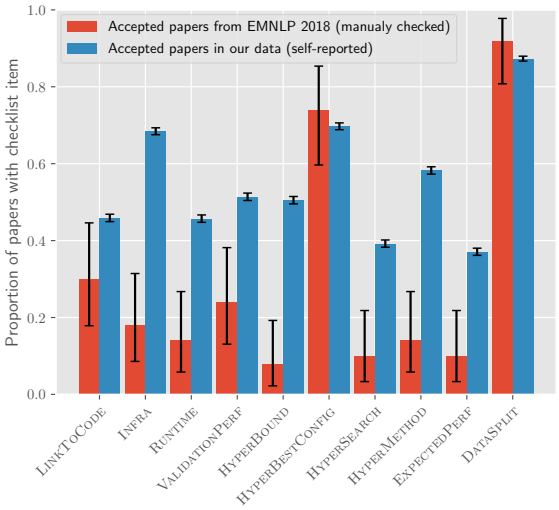

Figure 9: Reporting rates before and after the implementation of the Checklist. Dodge et al. (2019) manually check a subset of items in 50 randomly sampled EMNLP 2018 papers. We compare to accepted papers from all conferences in our data. While our self-reported data likely overestimate rates, it appears all but 2 items are now reported more often.

we would expect fewer N/As, given the Checklist's focus on empirical work.

There is little variation in response proportions between conferences. Excluding two types of outliers likely caused by changes in the Checklist (see Appendix A.3), the maximum difference between conferences for an item is 6.6% and the maximum difference averaged over all items is 2.2%. This does demonstrate that measured response patterns are robust across conferences. However it also indicates that reproducibility reporting has stagnated over this one year period.

When asked, a majority of reviewers found the Checklist to be somewhat or very useful. In NAACL 2021 and ACL 2021 reviewers gave feedback on the Checklist. 59.9% found the checklist "Somewhat Useful," 17.0% found it "Very Useful," and 23.2% found it "Not Useful." While this is higher than the 34% of reviewers who answered "yes" that the similar NeurIPS 2019 Checklist was "useful for evaluating the submission" (Pineau et al., 2021), it is worth noting that respondents to the question for NeurIPS could answer that they did not read the checklist results.

Author comments from submissions where the majority of reviewers found the Checklist "Not Useful," show possible gaps in checklist coverage. Some comment on not training mod-

els or using hyperparameters from previous work. Many such submissions are represented among the 22.0% that answer N/A to all hyperparameter questions. Others comment on referring readers to citations for details of standard models, data, or metrics. Re-elaboration is pedagogically important but, comments argue, especially onerous for survey papers. Finally a comment notes that the Checklist is less relevant to psycholinguistics and cognitive modeling, and indeed the N/A rate of "Linguistic Theories, Cognitive Modeling and Psycholinguistics" TRACK submissions is 33.7%, an increase of 14.6% above the N/A rate over all TRACKs.

8 Discussion

Our findings from the NLP Reproducibility Checklist can both help inform new interventions and guide improvements to future checklists that will measure the outcomes of those interventions. These findings suggest that, after an initial increase, rates of reporting have stagnated in the period examined and will need new approaches to improve further.

The conference system should better support papers that collect new data. As discussed in Section 5, papers that collect new data have 5.1% lower acceptance rate than those that do not. Whether or not this gap is a cause or effect of the lack of prestige given to data work that Rogers (2021) describes, increasing awareness and resources for this work can help more high quality data reach publication. Checklists should also increase coverage of this topic. In our data a single item, NEWDATADESCRIPTION, covers all reporting regarding data collection. We find that papers with new data are perceived as less reproducible both when answering NO or YES to describing how they collected data. Likely a combination of several factors lead reviewers to score the reproducibility of papers with new data lower by 2.4% relative to papers without. To discover which are lacking, best practices in data reproducibility documentation (Gebru et al., 2018; Dodge et al., 2021) should be tracked individually with checklists.

Incentivize authors releasing code. We find that releasing code is the single most influential checklist item on perceived reproducibility. This aligns with work across diverse fields that argues open source code is key for transparent and reproducible science (Eglen et al., 2017; Celi et al., 2019; Shamir et al., 2013). These works also suggest that beyond reproducibility, open source code enables more impactful research by allowing other researchers to build on introduced methods and better understand findings through reading code. However, we find that less than half of papers in our study report releasing code at submission. We encourage conferences to incentivize code release at submission and especially camera-ready, and authors should be made aware of the significant benefit that code submission can have for the review process.[8] Initiatives like the NAACL 2022 Reproducibility Track are a step in the right direction, as they publicly recognize open source code and verify code availability rather than only relying on self-reporting. However, in our data we see no evidence that code availability is increasing over time, so more direct incentives from publication venues are needed.

Make checklist responses public. Self-reported data is notoriously unreliable, but making the checklist responses public will add accountability.[9] In addition, the checklist responses can reference specific sections and act as an index of the paper, so a reader knows where to look for what information. This will be implemented at ACL 2023, and we recommend other conferences follow.

Conferences should allow submission of checklists, unlimited appendices, and code a week after the main deadline. Doing so can help establish a norm of code submission as *part* of the review process. Likewise, additional time could improve completeness and accuracy of the checklist. Many pieces of information important for reproducibility are appropriate to include in the appendix of a paper without counting towards the page limit (e.g., a full list of hyperparameter values). This need not increase the burden on reviewers, as they can consult checklists rather than the appendix to assess reporting.

Looking Forward

Checklists collected during submission can measure practices in NLP at a comprehensive scale. To our knowledge, our work and Pineau et al.'s (2021) are the only analyses of submitted reproducibility checklists at AI conferences. These are examples

[8] Even when code cannot be made publicly available due to intellectual property concerns, private code submission should be facilitated and extra emphasis should be placed on reporting items such as efficiency measures that are hard to reconstruct without public code.

[9] This should still be combined with studies that manually audit reporting in papers.

of metascience in AI, or applying scientific rigor to the process of AI research; we expect that as NLP matures, we will see more examples of work analyzing and improving the scientific process. There are also examples of other work which manually audits papers (Fokkens et al., 2013; Gundersen and Kjensmo, 2018; McDermott et al., 2019; Haibe-Kains et al., 2020; Marie et al., 2021), which can compliment self-reported checklists, and other conference submission metadata (Chen et al., 2022), with validated samples.

As standard practices in our field evolve, we will have to update all parts of the conference process, from checklists to reviews to paper presentations. As a positive example, ACL Rolling Review implemented the Responsible NLP Checklist,[10] which includes ethics as well as reproducibility items. While we do not have data with which to evaluate the Responsible NLP checklist, our findings show the need for just such efforts to expand the coverage of checklists to better serve the community.

Limitations

Our analyses rely on data from checklists filled in by authors and ratings provided by reviewers. Checklists are self-reported and thus not necessarily accurate. We discuss where these bad faith resposes might influence our results in Appendix A.2. Another data limitation is that phrasing changes between conferences for some items, and 3 items do not appear in all conferences (see Appendix A.1). NAACL 2021 also introduces BLANK as a possible answer when respondents do not choose any answer. There is also possible ambiguity between the NO and N/A answers as it is apparent from the checklist open text comments that some authors used NO when the item was not applicable to their work. Our data also only covers four conferences across 2020 and 2021, and as such it is difficult to assess any temporal trends. Reviewer data is also subject to inaccuracy; for instance reviewer perceived reproducibility scores are only subjective estimations of the likelihood of actual reproducibility. Rushed reviewers could easily miss where some important information is reported in a paper. Moreover, we only have reviewer data for 2 of 4 conferences.

Our finding that papers that collect data have a gap in acceptance and perceive reproducibility relies on an indirect inference about which papers collect data. Checklists did not ask this explicitly but rather NEWDATADESCRIPTION should be answered N/A for all papers that do not collect data.

Our findings about code and data availability are limited by the ambiguity of when they must be made available to qualify for answering YES. It is evident from the open text checklist comments that some authors answer YES, NO, or even N/A when they have not yet made code or data available but plan to do so on acceptance.

Any self-reported inaccuracies in our data would particularly affect our findings about the impact of the Checklist introduction on reporting rates. By definition, we are not able to compare to self-reported rates from before Checklist introduction, so we instead rely on Dodge et al.'s (2019) manually checked rates. 9 of the items in our data are not covered in the previous work, but the items that are share have similar phrasing.

Finally, pooling results over conferences can obscure conference-specific dynamics, such as differences in which items have lower than average YES ACCEPT rates discussed in Appendix A.3. We check that trends that we highlight in our analyses are consistent across conferences. And we also present unaggregated figures in the appendix. Likewise, we find that ACCEPT rates are nearly identical across conferences (see Appendix A.4), enabling us to contrast against an overall acceptance rate.

Ethics Statement

Scientific reproducibility is key to the benefits science can bring to society. Simply put, findings that cannot be reproduced cannot be relied upon, which can lead to wasted societal resources or even to harmfully incorrect understandings that misguide interventions. Our work focuses on the use of checklists to improve reporting of reproducibility information in scientific publications. While overly prescriptive and general rules about reproducibility could stifle less represented research communities whose practices may be less well understood by conference organizers, checklists attempt to mitigate this risk by only reminding authors of possibly salient information while still permitting authors to determine which items are or are not applicable.

At the same time, checklists which are filled out and collected for data analysis have the additional ethical risks associated with work that attempts to make social practices legible. That is, a check-

[10] aclrollingreview.org/responsibleNLPresearch/

list may neglect to cover practices used in a research community and thereby efface their role in the overall scientific endeavor, or conversely some practice may receive unfair scrutiny in excess of that given to other more prestiged practices. In the long term, checklists are perhaps most important as documents for guiding new generations of researchers writing their first papers, and thus even without being enforced they may still be taken as normative statements about best practices in the field.

To guide efforts to improve reproducibility in the field of NLP, we have analyzed responses to the NLP Reproducibility Checklist collected by four conferences. The Checklist data is covered by the default terms as it has no stated license, and we use it with direct permission from the conference organizers who collected it. The authors of the first version of the checklist state that it is intended for "improved reporting of the setup and results of the experiments that authors have conducted" and that it will be used to "quantitatively analyze our checklist responses" (Dodge and Smith, 2020).

We have endeavored to maintain the privacy of respondents by keeping the data anonymized and presenting results at a sufficient level of aggregation to prevent deanonymization. Nevertheless all work that seeks to describe the opinions of groups of humans caries an ethical burden to do so accurately and consistently with the wishes of those represented. To that end, we take care to point out limitations in what can be inferred from the data, and as originally intended by the data creators we do not make the data publicly available.

Acknowledgements

We thank the organizers of EMNLP 2020 and 2021, NAACL 2021, and ACL 2021 for making Checklist and review data available for analysis. We thank Julian Michael, Kyle Lo, Lucy Lu Wang, and Ari Holtzman for fruitful conversations about metascience and feedback on paper drafts.

References

Alexander A. Aarts, Joanna E. Anderson, Christopher J. Anderson, and et al. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454.

Emily M. Bender. 2019. The benderrule: On naming the languages we study and why it matters.

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*, 6:587–604.

Leo A Celi, Luca Citi, Marzyeh Ghassemi, and Tom J Pollard. 2019. The plos one collection on machine learning in health and biomedicine: Towards open code and open data. *PloS one*, 14(1):e0210232.

Kamalika Chaudhuri and Ruslan Salakhutdinov. 2019. The icml 2019 code-at-submit-time experiment.

Chang Chen, Jiayao Zhang, Dan Roth, Ting Ye, and Bo Zhang. 2022. Association between author metadata and acceptance: A feature-rich, matched observational study of a corpus of iclr submissions between 2017-2022.

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, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jesse Dodge and Noah A. Smith. 2020. Guest post: Reproducibility at EMNLP 2020.

Stephen J Eglen, Ben Marwick, Yaroslav O Halchenko, Michael Hanke, Shoaib Sufi, Padraig Gleeson, R Angus Silver, Andrew P Davison, Linda Lanyon, Mathew Abrams, et al. 2017. Toward standard practices for sharing computer code and programs in neuroscience. *Nature neuroscience*, 20(6):770–773.

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 H. Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for datasets. *Communications of the ACM*, 64:86 – 92.

Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. State of the art: Reproducibility in artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Benjamin Haibe-Kains, George Adam, Ahmed Hosny, Farnoosh Khodakarami, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S. Greene, Tamara Broderick, Michael M. Hoffman, Jeffrey T. Leek, Keegan D. Korthauer, Wolfgang Huber, Alvis Brazma, Joelle Pineau, Robert Tibshirani, Trevor J. Hastie, John P. A. Ioannidis, John Quackenbush, and Hugo J.W.L. Aerts. 2020. Transparency and reproducibility in artificial intelligence. *Nature*, 586 7829:E14–E16.

Ben Hutchinson, Andrew Smart, A. Hanna, Emily L. Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2020. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

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 (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Matthew B. A. McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Marzyeh Ghassemi, and Luca Foschini. 2019. Reproducibility in machine learning for health.

Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, and Samuel R. Bowman. 2022. What do nlp researchers believe? results of the nlp community metasurvey.

Nature. 2018. Checklists work to improve science. *Nature*, 556(7701):273–274.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and H. Larochelle. 2021. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *ArXiv*, abs/2003.12206.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1776–1826, New York, NY, USA. Association for Computing Machinery.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR.

Anna Rogers. 2021. Changing the world by changing the data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2182–2194, Online. Association for Computational Linguistics.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'just what do you think you're doing, dave?' a checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lior Shamir, John F Wallin, Alice Allen, Bruce Berriman, Peter Teuben, Robert J Nemiroff, Jessica Mink, Robert J Hanisch, and Kimberly DuPrie. 2013. Practices in source code sharing in astrophysics. *Astronomy and Computing*, 1:54–58.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

A Appendix

A.1 Item Comparisons Across Checklists

In order to support comparison, 11 of the 19 checklist items correspond to specific items in the Machine Learning Paper Reproducibility Checklist,[11] a version of which was used at NeurIPS 2019 (Pineau et al., 2021). A further 4 items are either combinations or decomposition of items from this

[11] cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf

Table 3: Checklist item phrasing differences across conferences. Δ marks differing item phrasing. N/A marks conferences with no equivalent item.

Abbreviation	EMNLP 2020	EMNLP 2021	NAACL 2021	ACL 2021
ModelDescription	✓	✓	✓	✓
LinkToCode	✓	Δ	Δ	Δ
Infra	✓	✓	✓	✓
Runtime	✓	Δ	Δ	✓
Parameters	✓	✓	✓	✓
ValidationPerf	✓	✓	✓	✓
Metrics	✓	✓	Δ	✓
NoTrainingEvalRuns	N/A	✓	✓	N/A
HyperBound	✓	✓	✓	✓
HyperBestConfig	✓	✓	✓	✓
HyperSearch	✓	✓	N/A	✓
HyperMethod	✓	✓	Δ	✓
ExpectedPerf	Δ	✓	✓	Δ
DataStats	✓	Δ	Δ	✓
DataSplit	✓	✓	✓	✓
DataProcessing	✓	✓	✓	✓
DataDownload	✓	Δ	Δ	✓
NewDataDescription	✓	✓	✓	✓
DataLanguages	N/A	N/A	✓	N/A

checklist. 2 more, ValidationPerf and HyperSearch, are incorporated that are manually evaluated along with 8 items from Pineau et al. (2021) on a random sample of 50 papers from EMNLP 2018 in Dodge et al. (2019). In that analysis at least one checklist item was found per paper and each checklist item occurred in at least one paper. Finally, Parameters is included for its important role in measuring the complexity of models, and DataLanguages is included because of the importance of acknowledging which communities of speakers are being served by a language technology as noted by Bender (2019).

The phrasing overlap in NLP and ML Checklists permits comparison of our data to responses from NeurIPS 2019. Pineau et al. (2021) find similar rates of reporting for dataset and efficiency items, though fewer submissions respond N/A to describing data collection. They find higher rates for items concerning hyperparameters and multiple experiments. Most of all their acceptance rates conditioned on items differ dramatically from ours. All but one item for "empirical results" get lower than average acceptance for Yes and higher for N/A, while our data shows lower Yes Accept rates for only 3 empirical items. This suggests the applicability of the NLP Checklist is more aligned with reviewing at the studied conferences.

The phrasing of the Checklist items was determined by distinct groups of organizers for each conference. While 9 items maintain the same phrasing, 6 see phrasing changes, and 3 are only asked at some conferences (see Tables 3 and 6). LinkToCode remains the same in substance while phrasing variations address logistics such as file formats and anonymization. Runtime varies in EMNLP 2021 and NAACL 2021 by asking for runtime or energy cost. Metrics varies in NAACL 2021 by not specifying links to metric code. ExpectedPerf varies in EMNLP 2020 and ACL 2021 by asking for mean *and* variance of hyperparameters, where in the other phrasing any summary statistic of results is sufficient. DataStats includes languages and label distributions in its variations. DataDownload varies only in file formats, except for NAACL 2021 which also allows for a simulation environment.

A.2 Bad Faith Responses

As expected the ModelDescription question was answered Yes by almost all submissions (96.3% of responses over all three conferences). This question was intended as an attention check and was designed such that almost all submissions should answer Yes. This helps assure that respondents are not using the N/A (or No) response in protest or bad faith to quickly fill in meaningless answers, as only 2.6% (or 0.2%) of submissions answer ModelDescription this way. However this does not preclude the use of answering questions Yes in bad faith to bypass the checklist. Likewise we see 8.0% of NAACL 2021 respondents leave this field Blank.

Submissions with all identical answers have lower than average acceptance. In Table 4 we show counts and change from average acceptance rates for submissions whose answers are all identical. This pattern is most prevalent for Yes and Blank, accounting for several percent of all submissions. All No and N/A submissions, however, are quite infrequent. One possible explanation is that selecting all Yes or Blank is an expedient way to bypass the checklist during the submission process. Though we cannot know what portion of submissions with this pattern may exhibit this issue, it is important to be aware of this limitation.

A.3 Additional results

Yes was the most common response to checklist questions. The proportion of a given answer in responses for each question is shown in Figure 12. 62.7% of responses to checklist questions across all conferences were Yes, with 62.8%, 63.7%, 60.0%, and 62.5% respectively for EMNLP 2020 and 2021, NAACL 2021, and ACL 2021. The ma-

Table 4: Submissions with all Checklist responses given the same answer (e.g., responding N/A to all items) and their change in MAIN and FINDINGS acceptance rate from overall rate.

Response	Conference	Submissions	ACCEPT
YES	EMNLP 2020	134 (4.5%)	−9.9%
	EMNLP 2021	238 (7.3%)	−6.7%
	NAACL 2021	79 (6.4%)	−3.3%
	ACL 2021	213 (7.3%)	−8.2%
NO	EMNLP 2020	1 (0.0%)	−39.7%
	EMNLP 2021	0 (0.0%)	-
	NAACL 2021	0 (0.0%)	-
	ACL 2021	0 (0.0%)	-
N/A	EMNLP 2020	17 (0.6%)	−4.4%
	EMNLP 2021	22 (0.7%)	2.3%
	NAACL 2021	15 (1.2%)	14.6%
	ACL 2021	40 (1.4%)	2.4%
BLANK	NAACL 2021	89 (7.2%)	−24.1%
All Same	**Overall**	**848** (8.2%)	**-**8.1%

jority of the questions have greater than 50% YES response rate over all conferences. Only LINKTOCODE, RUNTIME, HYPERSEARCH, EXPECTEDPERF, and NEWDATADESCRIPTION receive less than half YES responses. All questions receive more YES responses than NO and only NEWDATADESCRIPTION receives more N/A than YES.

The checklist items which receive less than average YES acceptance rates are not consistent across all conferences. Figure 13 shows acceptance rates for all checklist items over all conferences. From this figure we see that ACL 2021 also has LINKTOCODE, PARAMETERS, DATASTATS, and DATADOWNLOAD YES acceptance rates below average, though all of these estimates include the average acceptance rate within their 95% confidence intervals. NAACL 2021 has no YES acceptance rates below average, though VALIDATIONPERF and NEWDATADESCRIPTION remain the two lowest. Likely all NAACL 2021 YES acceptance rates are elevated because in this conference respondents could leave questions BLANK, possibly diverting some low-quality responses to BLANK instead of YES. Also of note, however is that across conferences RUNTIME receives high YES acceptance rate, achieving the best overall at 4.3% higher than average.

There are outliers to the little variation in response proportions between conferences, but they are likely artifacts of changes in the checklist. The rate of YES responses is generally lower for NAACL 2021, but this is likely due to the ability to leave checklist responses BLANK. Excluding NAACL 2021, the largest difference in YES rate (27.3%) occurs on EXPECTEDPERF when this item changes phrasing substantially between EMNLP 2020 and EMNLP 2021.

PCA analysis. To identify clusters of checklist items that relate to each other, we take inspiration from similar analysis in Michael et al. (2022) and use principal component analysis (PCA) on all responses to shared checklist items across the four conferences. This results in 16 features, which we linearize as {NO $\rightarrow -1$, N/A /BLANK $\rightarrow 0$, YES $\rightarrow 1$}. We run PCA using `scikit-learn` version 1.1.1 (Pedregosa et al., 2011) and find that the first 4 components cover 55.9% of the variance in the data. Table 5 shows these components and their coefficients with magnitude > 0.20. The first component assigns weight all in one polarity to the checklist items with middling frequencies, highlighting practices where perhaps community norms have not settled. The second component shows the intuitive connection between LINKTOCODE and DATADOWNLOAD. The third captures what might be particularly resource intensive experiments that emphasize efficiency metrics but prevent running many experiment repetitions. Lastly the fourth component puts hyperparameter items in opposition with VALIDATIONPERF, perhaps pointing to-

Partially adopted practices (27.9%)			
Runtime	-0.43	Parameters	-0.37
Infra	-0.34	LinkToCode	-0.33
ValidationPerf	-0.30	DataDownload	-0.30
HyperBound	-0.26	ExpectedPerf	-0.26
HyperMethod	-0.23		
Open source (10.8%)			
LinkToCode	0.71	DataDownload	0.48
Parameters	-0.28	ValidationPerf	-0.20
Long running experiments (9.2%)			
ValidationPerf	-0.57	ExpectedPerf	-0.48
Runtime	0.47	Infra	0.40
Validating without hyperparameters (8.1%)			
HyperMethod	-0.50	HyperBound	-0.49
ValidationPerf	0.48	HyperBestConfig	-0.32
Runtime	0.25	Parameters	0.24

Table 5: The top four components from running PCA on shared checklist items from four conferences, with percent variance explained in parentheses. Each component lists checklist items and their coefficients with magnitude > 0.20.

wards work that adapts choices other than traditional hyperparameters to a validation set.

A.4 Baseline Acceptance Rates

To aid in analyzing how publication decisions differ based on responses to the checklist, we first must establish what is the average acceptance rate across all papers in our data. In Table 7 we provide basic statistics about submissions and decisions. The acceptance rates reported[12] by the conferences all include an unknown and varying number of withdrawn and desk-rejected papers and thus are not easily comparable. For the rest of our analysis we will instead make use of acceptance rates computed from our data that always remove all withdrawn and desk-rejected papers. With this approach we find that all of the conferences have similar acceptance rates when including both acceptance to the MAIN conference and to FINDINGS.

[12] https://aclweb.org/aclwiki/Conference_acceptance_rates

Table 6: Exact checklist item phrasing for each conference. Items listed as N/A did not appear on the checklist for that conference.

Short Name	EMNLP 2020	EMNLP 2021	NAACL 2021	ACL 2021
ModelDescription	A clear description of the mathematical setting, algorithm, and/or model.	A clear description of the mathematical setting, algorithm, and/or model.	A clear description of the mathematical setting, algorithm, and/or model	A clear description of the mathematical setting, algorithm, and/or model
LinkToCode	A link to a downloadable source code, with specification of all dependencies, including external libraries	Submission of a zip file containing source code, with specification of all dependencies, including external libraries, or a link to such resources (while still anonymized)	A link to a downloadable source code, with specification of all dependencies, including external libraries (recommended for camera ready, though welcome for initial submission)	Submission of a zip file containing source code, with specification of all dependencies, including external libraries, or a link to such resources (while still anonymized)
Infra	Description of computing infrastructure used	Description of computing infrastructure used	A description of computing infrastructure used	Description of computing infrastructure used
Runtime	Average runtime for each approach	The average runtime for each model or algorithm (e.g., training, inference, etc.), or estimated energy cost	The average runtime for each model or algorithm, or estimated energy cost	Average runtime for each approach
Parameters	Number of parameters in each model	Number of parameters in each model	The number of parameters in each model	Number of parameters in each model
ValidationPerf	Corresponding validation performance for each reported test result	Corresponding validation performance for each reported test result	Corresponding validation performance for each reported test result	Corresponding validation performance for each reported test result
Metrics	Explanation of evaluation metrics used, with links to code	Explanation of evaluation metrics used, with links to code	A clear definition of the specific evaluation measure or statistics used to report results.	Explanation of evaluation metrics used, with links to code
NoTrainingEvalRuns	N/A	The exact number of training and evaluation runs	The exact number of training and evaluation runs	N/A
HyperBound	Bounds for each hyperparameter	Bounds for each hyperparameter	The bounds for each hyperparameter	Bounds for each hyperparameter
HyperBestConfig	Hyperparameter configurations for best-performing models	Hyperparameter configurations for best-performing models	The hyperparameter configurations for best-performing models	Hyperparameter configurations for best-performing models
HyperSearch	Number of hyperparameter search trials	Number of hyperparameter search trials	N/A	Number of hyperparameter search trials
HyperMethod	The method of choosing hyperparameter values (e.g., uniform sampling, manual tuning, etc.) and the criterion used to select among them (e.g., accuracy)	The method of choosing hyperparameter values (e.g., uniform sampling, manual tuning, etc.) and the criterion used to select among them (e.g., accuracy)	The method of choosing hyperparameter values (e.g. manual tuning, uniform sampling, etc.) and the criterion used to select among them (e.g. accuracy)	The method of choosing hyperparameter values (e.g., uniform sampling, manual tuning, etc.) and the criterion used to select among them (e.g., accuracy)
ExpectedPerf	Expected validation performance, as introduced in Section 3.1 in * Dodge et al, 2019, or another measure of the mean and variance as a function of the number of hyperparameter trials.	Summary statistics of the results (e.g., mean, variance, error bars, etc.)	Summary statistics of the results (e.g. mean, variance, error bars, etc.)	Expected validation performance, or the mean and variance as a function of the number of hyperparameter trials
DataStats	Relevant statistics such as number of examples	Relevant details such as languages, and number of examples and label distributions	Relevant statistics such as number of examples and label distributions	Relevant statistics such as number of examples
DataSplit	Details of train/validation/test splits	Details of train/validation/test splits	Details of train/validation/test splits	Details of train/validation/test splits
DataProcessing	Explanation of any data that were excluded, and all pre-processing steps	Explanation of any data that were excluded, and all pre-processing steps	An explanation of any data that were excluded, and all pre-processing steps	Explanation of any data that were excluded, and all pre-processing steps
DataDownload	A link to a downloadable version of the data	A zip file containing data or link to a downloadable version of the data	A link to a downloadable version of the dataset or simulation environment	A link to a downloadable version of the data
NewDataDescription	For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control.	For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control.	For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control	For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control
DataLanguages	N/A	N/A	For natural language data, the name of the language(s)	N/A

Figure 10: Phi coefficient of binary YES or not YES answer for each item to binary ACCEPT or not ACCEPT for each submission.

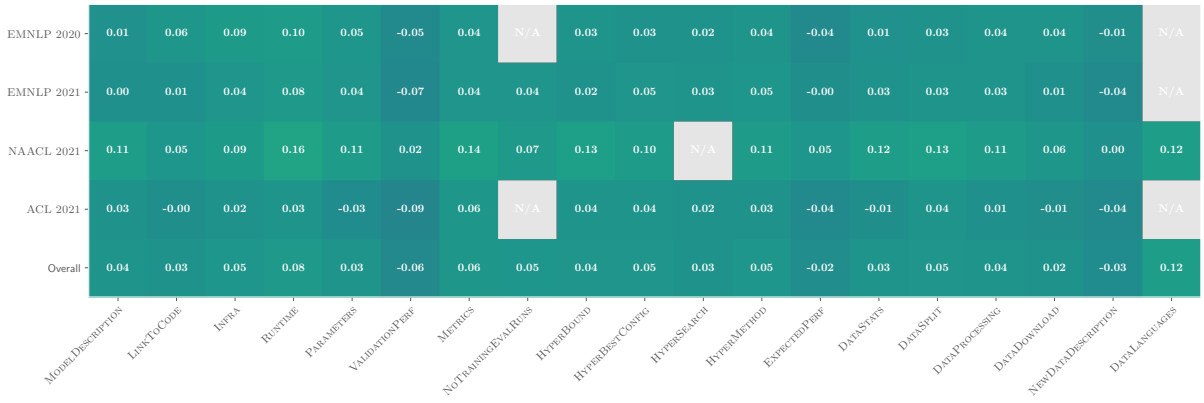

Table 7: Submissions and decisions statistics. Reported acceptance rates include varying amounts of withdrawn and desk-reject submissions. We exclude all of these to standardize to rates.

	Conference	Submissions	Withdrawn / Desk-Reject	Accept Rate MAIN	Accept Rate FINDINGS
Reported	EMNLP 2020	3359	-	22.4%	-
	EMNLP 2021	3600	-	23.3%	11.6%
	NAACL 2021	1797	-	26.5%	N/A
	ACL 2021	3350	-	21.2%	13.6%
Our Data	EMNLP 2020	3666	660	24.9%	14.8%
	EMNLP 2021	4815	1555	25.8%	12.9%
	NAACL 2021	1797	565	38.7%	N/A
	ACL 2021	3377	470	24.4%	15.7%
	Overall	13655	3250	39.4%	

Figure 11: Phi coefficient between items shared over all conferences for the binary variable YES or not YES. Unsurprisingly, related groups of items about efficiency, hyperparameters, and data each correlate together.

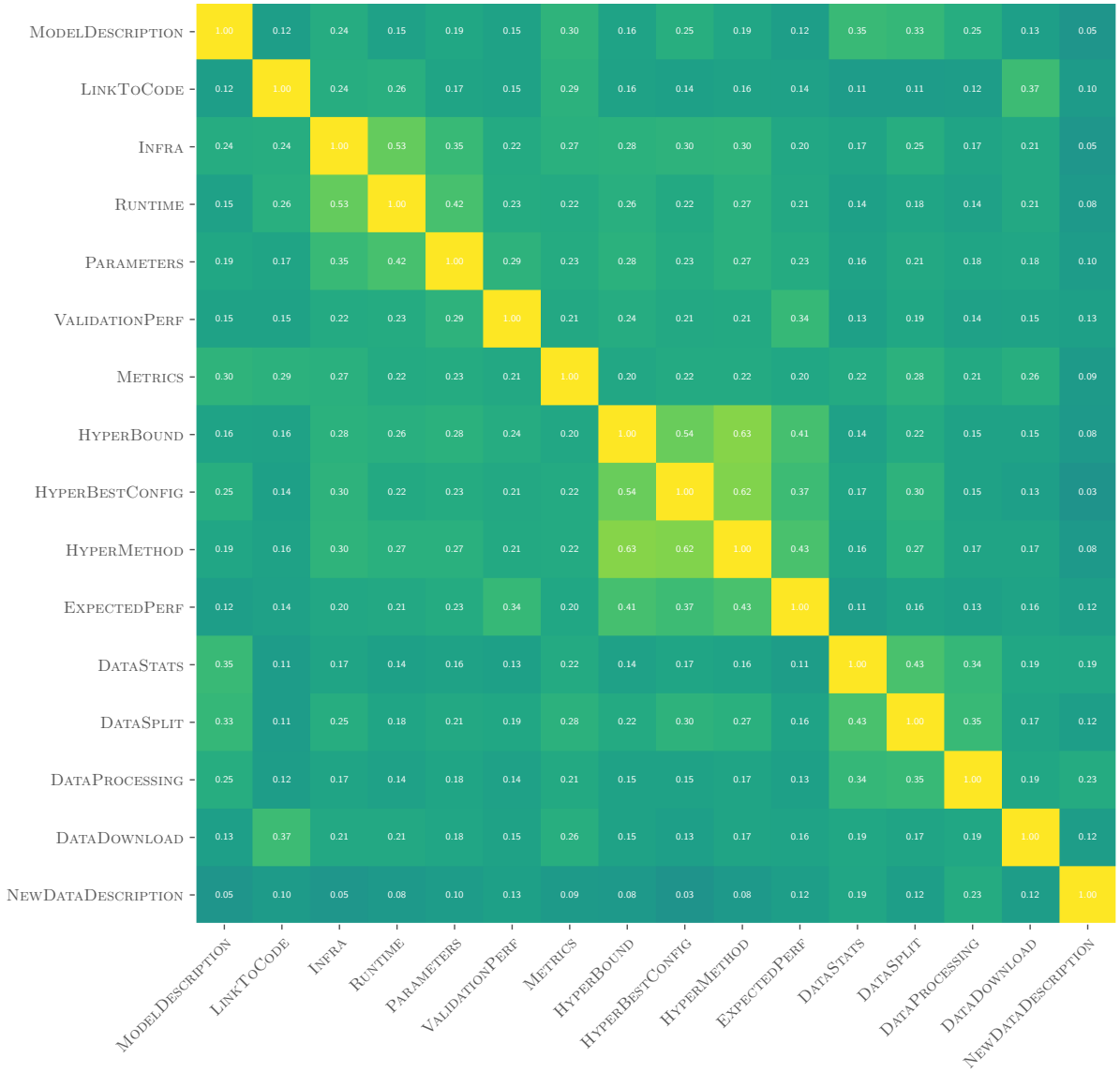

Figure 12: The portion of submissions giving a particular response per question. Note that NAACL 2021 respondents were able to leave questions BLANK; These are still counted in the total responses for these ratios.

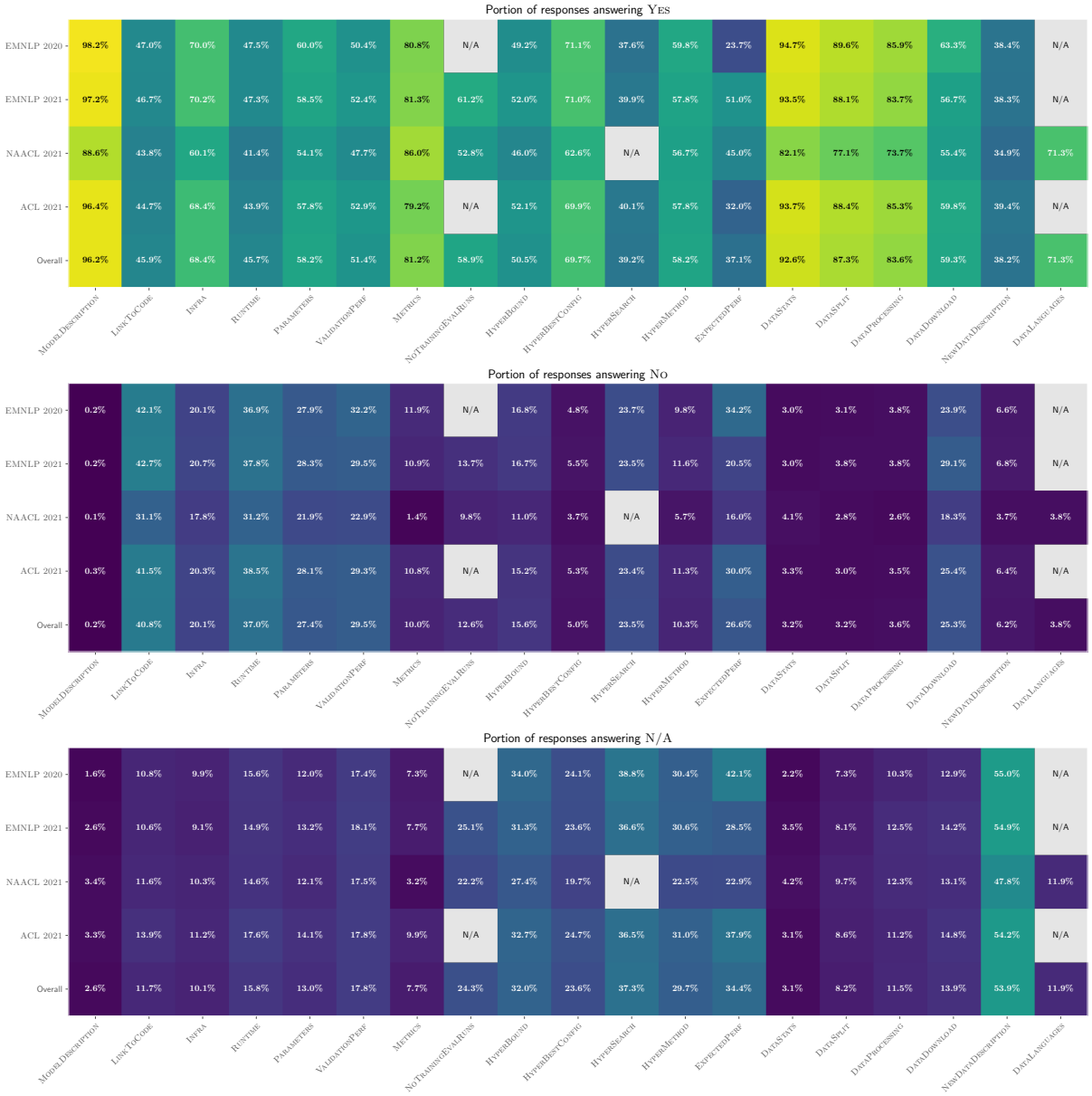

Figure 13: ACCEPT rates for submissions with a given response. Column (A) shows rate conditioned on response regardless of item. (B) conditions on answer and item. Rows present each conference and pooled results overall.

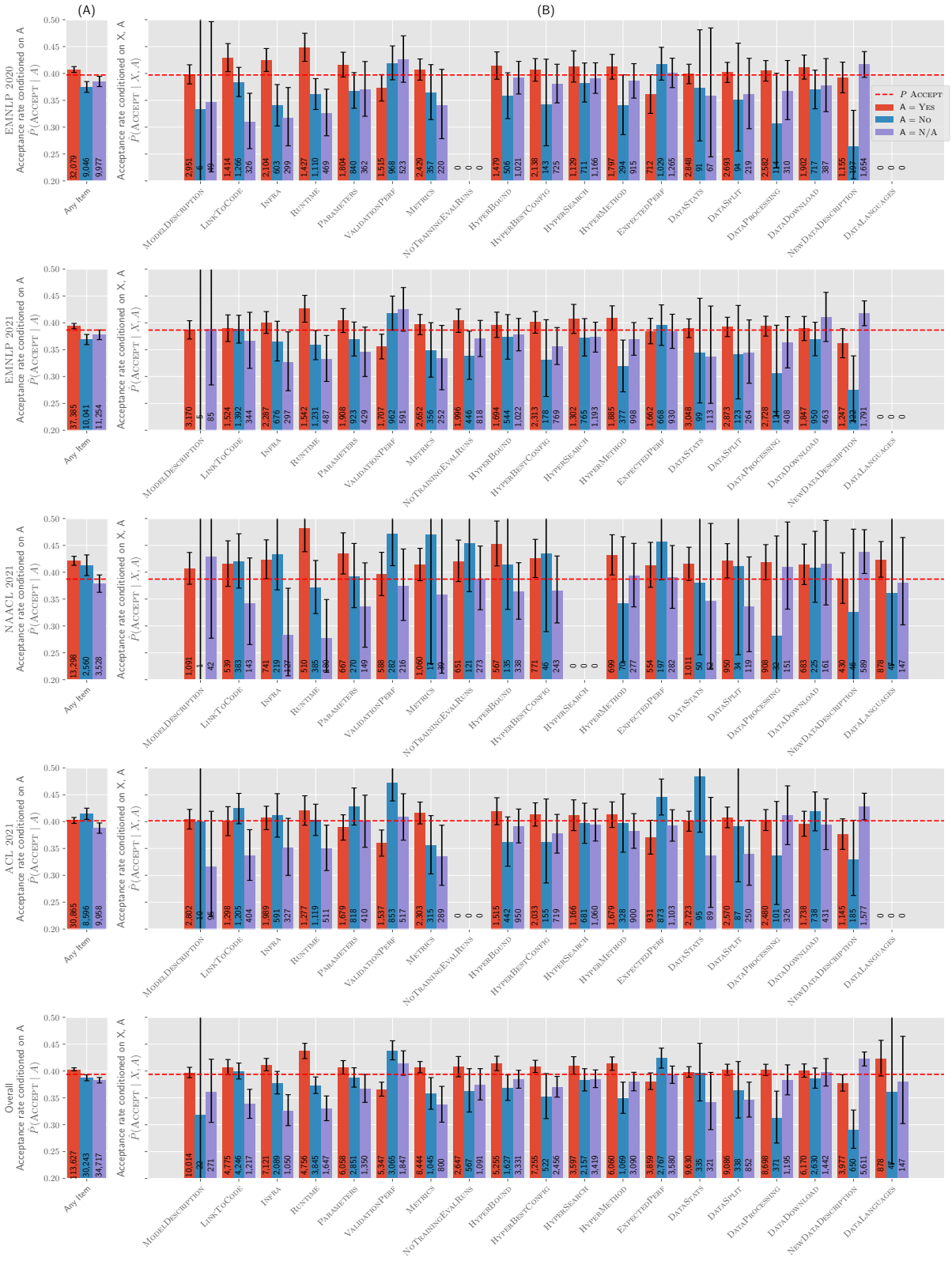

Figure 14: Reviewer perceived reproducibility score (AVGREPROD$\in [1, 5]$) for submissions with a given response. Column (A) shows score conditioned on response regardless of item. (B) conditions on answer and item. Rows present the two conferences with such data and pooled results overall.

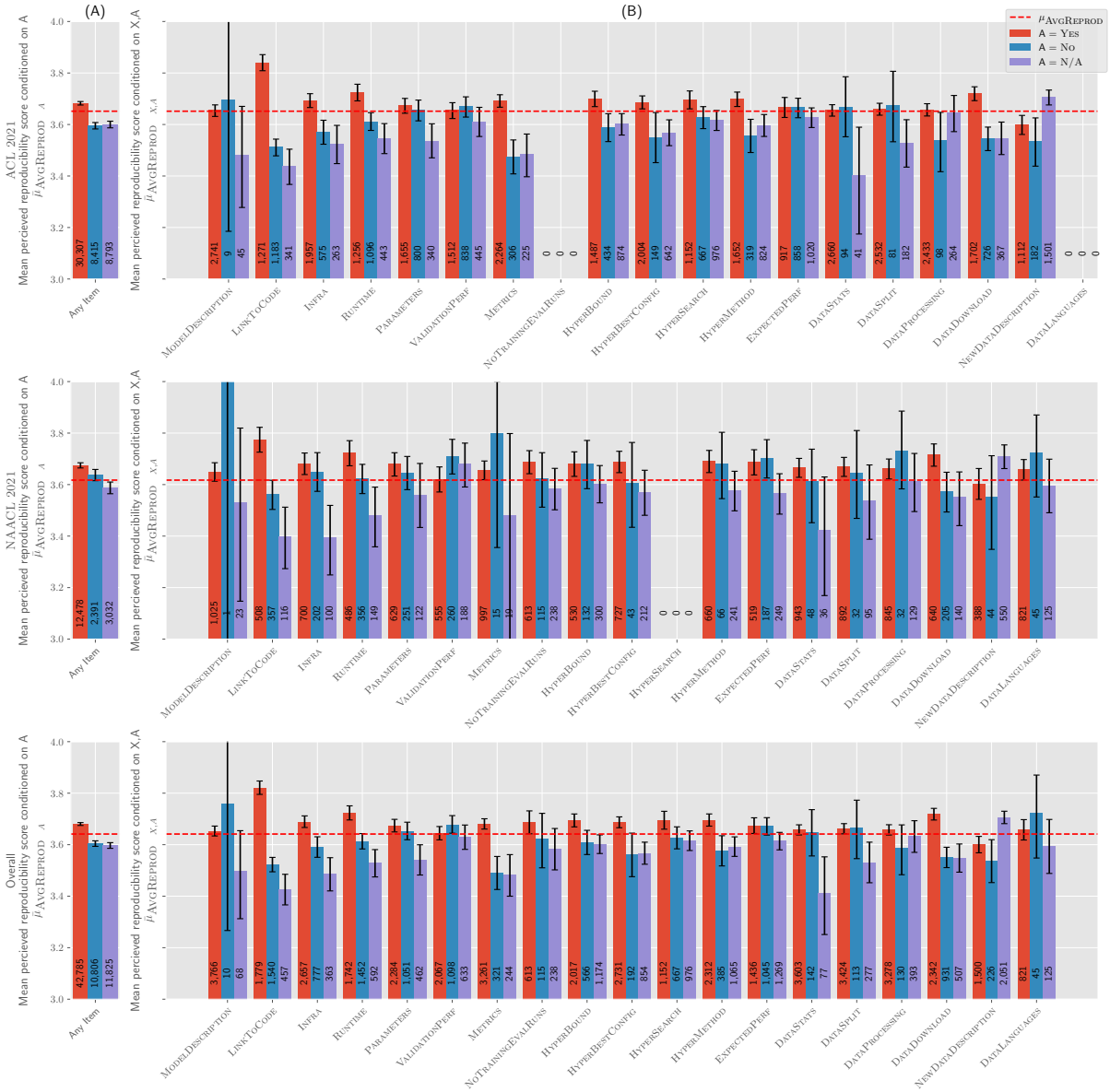

Figure 15: Proportion (row labels) and ACCEPT rates (horizontal purple bars) over all conferences for top response patterns for items split into three sections.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
see Limitations section (unnumbered)

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement section (unnumbered)

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3

☑ B1. Did you cite the creators of artifacts you used?
see Section 1

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Ethics Statement section (unnumbered)

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Ethics Statement section (unnumbered)

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Ethics Statement section (unnumbered)

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3 documents details of the data and the paper as a whole is an analysis of the data

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A.1

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Ethics Statement section (unnumbered)

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
To our knowledge the conference organizers who originally collected this data did not do so under approval of an ethics review board.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
The conference organizers who originally collected this data did not associate it with demographic or geographic characteristics of the authors submitting the checklist