ACL Rolling Review

A new initiative of the Association for Computational Linguistics

Members of the ACL are responsible for adhering to the ACL code of ethics. The ARR Responsible NLP Research checklist is designed to encourage best practices for responsible research, addressing issues of research ethics, societal impact and reproducibility. It is largely based on the NeurIPS 2021 paper checklist, the reproducible data checklist from Rogers, Baldwin, Leins's EMNLP 2021 paper "Just What do You Think You're Doing, Dave? A Checklist for Responsible Data Use in NLP", and the NLP Reproducibility Checklist built from Dodge et al. EMNLP 2019's paper "Show Your Work: Improved Reporting of Experimental Results".

We expect authors to show that they follow best practices in two ways:

1. by filling in the checklist to ensure that best practices are put in place when using, creating or providing assets,
2. by including a discussion in the paper about any potential positive or negative societal impacts and any limitations of the work. The guidelines below provide additional information about what should be discussed.

The checklist templates (LaTex and Word, with a fillable PDF form as last resort) can be found here.

Reviewers will be asked to use the checklist as one of the factors in their evaluation.

Guidelines for Answering Checklist Questions

For each question in the checklist:

- If you answer "Yes", indicate the section(s) where the information can be found in your paper. Note that you do not need to have a separate entitled section in the paper for each of the questions. Information about limitations for instance could be part of the conclusion. Several questions can indicate the same section (for instance, citation and license for data are likely going to appear in the same (sub)section of your paper).

- If you answer "No" to the question, provide a justification for why. The list is meant as a point of reflection for the authors. You are strongly encouraged to take a look at the checklist early on as it likely will positively influence the way you do your research and write the paper.

The questions are framed in terms of transparency: "Did you include [information]?" While it is generally preferable that your paper clearly answers positively to the question, it is perfectly acceptable that it does not, provided a proper justification is given (e.g., "We were unable to find the license for the dataset we used"). Not answering positively to a question is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgement and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material.

We provide guidance on how to answer each of the checklist questions below.

A. For every submission:

A1. Did you describe the *limitations* of your work?

- Point out any strong assumptions and how robust your results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only held locally). Reflect on how these assumptions might be violated in practice and what the implications would be.
- Reflect on the scope of your claims, e.g., if you only tested your approach on a few datasets, languages, or did a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated. Reflect on the factors that influence the performance of your approach. For example, a speech-to-text system might not be able to be reliably used to provide closed captions for online lectures because it fails to handle technical jargon.
- If you analyze model biases: which definition of bias are you using? Did you state the motivation and definition explicitly? See the discussion in Blodgett et al. (2020).
- We understand that authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection. It is worth keeping in mind that a worse outcome might be if reviewers discover limitations that aren't acknowledged in the paper. In general, we advise authors to use their best judgement and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

A2. Did you discuss any potential *risks* of your work?

- Examples of risks include potential malicious or unintended harmful effects and uses (e.g., disinformation, generating fake profiles, surveillance), environmental impact (e.g., training huge models), fairness considerations (e.g., deployment of technologies that could further disadvantage or exclude historically disadvantaged groups), privacy considerations (e.g., a paper on model/data stealing), and security considerations (e.g., adversarial attacks). See discussion in Leins et. al. (2020) as examples.
- Does the research contribute to overgeneralization, bias confirmation, under or overexposure of specific languages, topics, or applications at the expense of others? See Hovy and Spruit (2016) for examples.
- We expect many papers to be foundational research and not tied to particular applications, let alone deployments. However, we encourage authors to discuss potential risks if they see a path to any positive or negative applications. For example, the authors can emphasize how their systems are intended to be used, how they can safeguard their systems against misuse, or propose future research directions.
- Consider different stakeholders that could be impacted by your work. Is it possible that research benefits some stakeholders while harming others? Does it pay special attention to vulnerable or marginalized communities? Does the research lead to exclusion of certain groups? See Dev et. al (2021) for examples.
- Consider dual use, i.e, possible benefits or harms that could arise when the technology is being used as intended and functioning correctly, benefits or harms that could arise when the technology is being used as intended but gives incorrect results, and benefits or harms following from (intentional or unintentional) misuse of the technology.
- Consider citing previous work on relevant mitigation strategies for the potential risks of the work (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of NLP).

A3. Do the abstract and introduction summarize the paper's main claims?

- The main claims in the paper should be clearly stated in the abstract and in the introduction.
- These claims should be supported by evidence presented in the paper, potentially in the form of experimental results, reasoning, or theory. The connection between which evidence supports which claims should be clear.

- The context of the contributions of the paper should be clearly described, and it should be stated how much the results would be expected to generalize to other contexts.
- It should be easy for a casual reader to distinguish between the contributions of the paper and open questions, future work, aspirational goals, motivations, etc.

B. Did you use or create *scientific artifacts*?

- Scientific artifacts may include code, data, models or other artifacts.
- Most NLP research uses scientific artifacts.
- Many NLP papers are accompanied by new scientific artifacts.

B1. Did you cite the creators of artifacts you used?

- For composite artifacts like the GLUE benchmark, this means all creators.
- Cite the original paper that produced the code package or dataset.
- Remember to state which version of the asset you're using.
- If possible, include a URL.

B2. Did you discuss the *license or terms* for use and / or distribution of any artifacts?

- State the name of the license (e.g., CC-BY 4.0) for each asset.
- If you scraped or collected data from a particular source (e.g., website or social media API), you should state the copyright and terms of service of that source. Please note that some sources do not allow inference of protected categories like gender, sexual orientation, health status, etc.
  - The data might be in public domain and licensed for research purposes.
  - The data might be used with consent of its creators or copyright holders.
  - If the data is used without consent, the paper makes the case to justify its legal basis (e.g., research performed in the public interest under GDPR).
- If you are releasing assets, you should include a license, copyright information, and terms of use in the package.
- If you are repackaging an existing dataset, you should state the original license as well as the one for the derived asset (if it has changed).
- If you cannot find this information online, you are encouraged to reach out to the asset's creators.

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)?

- Data and/or pretrained models are released under a specified license that is compatible with the conditions under which access to data was granted (in particular, derivatives of data accessed for research purposes should not be deployed in the real world as anything other than a research prototype, especially commercially)
- The paper specifies the efforts to limit the potential use to circumstances in which the data/models could be used safely (such as an accompanying data/model statement).
- The data is sufficiently anonymized to make identification of individuals impossible without significant effort. If this is not possible due to the research type, please state so explicitly and explain why.
- The paper discusses the harms that may ensue from the limitations of the data collection methodology, especially concerning marginalized/vulnerable populations, and specifies the scope within which the data can be used safely.

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?

- There are some settings where the existence of offensive content is not necessarily bad (e.g., swear words occur naturally in text), or part of the research question (i.e., hate speech). This question is just to encourage discussion of potentially undesirable properties.
- Explain how you checked for offensive content and identifiers (e.g., with a script, manually on a sample, etc.).
- Explain how you anonymized the data, i.e., removed identifying information like names, phone and credit card numbers, addresses, user names, etc. Examples are monodirectional hashes, replacement, or removal of data points. If anonymization is not possible due to the nature of the research (e.g., author identification), explain why.
- List any further privacy protection measures you are using: separation of author metadata from text, licensing, etc.
- If any personal data is used: the paper specifies the standards applied for its storage and processing, and any anonymization efforts.
- If the individual speakers remain identifiable via search: the paper discusses possible harms from misuse of this data, and their mitigation.

B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

- Be sure to report the language of any language data, even if it is commonly-used benchmarks.
- Describe basic information about the data that was used, such as the domain of the text, any information about the demographics of the authors, etc.

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.

C. Did you run *computational experiments*?

C1. Did you report the *number of parameters* in the models used, the *total computational budget* (e.g., GPU hours), and *computing infrastructure* used?

- Even for commonly-used models like BERT, reporting the number of parameters is important because it provides context necessary for readers to understand experimental results. The size of a model has an impact on performance, and it shouldn't be up to a reader to have to go look up the number of parameters in models to remind themselves of this information.
- The total computational budget can be presented however is most appropriate for the paper – if most experiments were run on GPUs, then important information would likely include the total number of GPU hours, the amount of parallelism across GPUs, the size of the GPUs, etc. to allow a reader to estimate the computational requirements to reproduce, use, or build upon the work.
- Note that this should include information about all experiments, not just the final runs that led to the results presented in the paper. If exact numbers are not available, an estimate is better than nothing.

C2. Did you discuss the experimental setup, including *hyperparameter search* and *best-found hyperparameter* values?

- The experimental setup should include information about exactly how experiments were set up, like how model selection was done (e.g., early stopping on validation data, the single model with the lowest loss, etc.), how data was preprocessed, etc.

- Many research projects involve manually tuning hyperparameters until some "good" values are found, and then running a final experiment which is reported in the paper. Other projects involve using random search or grid search to find hyperparameters. In all cases, report the results of such experiments, even if they were stopped early or didn't lead to your best results, as it allows a reader to know the process necessary to get to the final result and to estimate which hyperparameters were important to tune.
- Be sure to include the best-found hyperparameter values (e.g., learning rate, regularization, etc.) as these are critically important for others to build on your work.
- The experimental setup should likely be described in the main body of the paper, as that is important for reviewers to understand the results, but large tables of hyperparameters or the results of hyperparameter searches could be presented in the main paper or appendix.

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?

- Error bars can be computed by running experiments with different random seeds, Clopper–Pearson confidence intervals can be placed around the results (e.g., accuracy), or expected validation performance can be useful tools here.
- In all cases, when a result is reported, it should be clear if it is from a single run, the max across N random seeds, the average, etc.
- When reporting a result on a test set, be sure to report a result of the same model on the validation set (if available) so others reproducing your work don't need to evaluate on the test set to confirm a reproduction.

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.)?

- The version number or reference to specific implementation is important because different implementations of the same metric can lead to slightly different results (e.g., ROUGE).
- The paper cites the original work for the model or software package. If no paper exists, a URL to the website or repository is included.
- If you modified an existing library, explain what changes you made.

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

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.?

- Examples of risks include a crowdsourcing experiment which might show offensive content or collect personal identifying information (PII). Ideally, the participants should be warned.
- Including this information in the supplemental material is fine, but if the main contribution of your paper involves human subjects, then we strongly encourage you to include as much detail as possible in the main paper.

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)?

- Be explicit about how you recruited your participants. For instance, mention the specific crowdsourcing platform used. If participants are students, give information about the population (e.g., graduate/undergraduate, from a specific field), and how they were compensated (e.g., for course credit or through payment).
- In case of payment, provide the amount paid for each task (including any bonuses), and discuss how you determined the amount of time a task would take. Include discussion on how the wage was determined and how you determined that this was a fair wage.

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?

D4. Was the data collection protocol *approved (or determined exempt)* by an ethics review board?

- Depending on the country in which research is conducted, ethics review (e.g., from an IRB board in the US context) may be required for any human subjects research. If an ethics review board was involved, you should clearly state it in the paper. However, stating that you obtained approval from an ethics review board does not imply that the societal impact of the work does not need to be discussed.
- For initial submissions, do not include any information that would break anonymity, such as the institution conducting the review.

D5. Did you report the basic demographic and geographic characteristics of the *annotator* population that is the source of the data?

- State if your data include any protected information (e.g., sexual orientation or political views under GDPR).
- The paper is accompanied by a data statement (see Bender and Friedman, 2018) describing the basic demographic and geographic characteristics of the author population that is the source of the data, and the population that it is intended to represent.
- If applicable: the paper describes whether any characteristics of the human subjects were self-reported (preferably) or inferred (in what way), justifying the methodology and choice of description categories.

ARR review form

The ARR review form takes aspects of ethics and reproducibility into account.

Credits

This checklist was developed by Marine Carpuat, Marie-Catherine de Marneffe and Ivan Vladimir Meza Ruiz, the NAACL 2022 program chairs, working with Jesse Dodge, and with the ARR editors in chief. Additional input was provided by the other NAACL 2022 reproducibility chairs, Margot Mieskes, Anna Rogers, and the ACL Ethics Committee.