Yes-Yes-Yes: Donation-based Peer Reviewing Data Collection for ACL Rolling Review and Beyond

Nils Dycke*, **Ilia Kuznetsov***, **Iryna Gurevych**
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science
Technische Universität Darmstadt
ukp.informatik.tu-darmstadt.de

Abstract

Peer review is the primary gatekeeper of scientific merit and quality, yet it is prone to bias and suffers from low efficiency. This demands cross-disciplinary scrutiny of the processes that underlie peer reviewing; however, quantitative research is limited by the data availability, as most of the peer reviewing data across research disciplines is never made public. Existing data collection efforts focus on few scientific domains and do not address a range of ethical, license- and confidentiality-related issues associated with peer reviewing data, preventing wide-scale research and application development. While recent *methods* for peer review analysis and processing show promise, a solid *data foundation* for computational research in peer review is still missing.

To address this, we present an in-depth discussion of peer reviewing data, outline the ethical and legal desiderata for peer reviewing data collection, and propose the first continuous, donation-based data collection workflow that meets these requirements. We report on the ongoing implementation of this workflow at the ACL Rolling Review and deliver the first insights obtained with the newly collected data.

1 Introduction

Peer review is the cornerstone of academic quality control. The ever-increasing submission rates expose the weaknesses of this process in terms of objectivity and efficiency, impacting the quality control in many fields of science, including machine learning and NLP. As a reaction to this, the first generation of computational studies in peer review aim to both scrutinize it as a process, and to develop applications that would make reviewing and the associated editorial work more efficient.

Such studies crucially depend on the availability of peer reviewing data. Yet this data is hard to come by and is associated with a range of ethical, confidentiality and copyright issues: Do the paper authors and reviewers agree to the use of their data? Should unpublished anonymous submissions be added to research datasets? What about the reviews for these submissions? What license should be attached to anonymous review reports, and how should they be attributed? The current ad-hoc approach to peer reviewing dataset construction is to harvest publicly available data from open sources. This limits computational studies on peer review to a few research communities, and leaves the aforementioned issues open, preventing the future use of the collected data in research and application development. While current methodological advances in peer review processing show great promise, the data foundation that would enable reproducible, ethically sound research on peer review across diverse research domains is yet to be established. To close this gap, in this work we:

- outline the challenges and trade-offs associated with the peer reviewing data collection;
- propose *Yes-Yes-Yes* – a generic workflow to address those challenges;
- report on the instantiation of this workflow at ACL Rolling Review[1];
- provide an open implementation of the proposed workflow for any research community that uses OpenReview.net[2] as their conference management system.

Our work sets up a vocabulary for discussing peer reviewing data collection and processing, raises community awareness in ethics- and copyright-aware data handling, and contributes to the active line of research in sustainable data collection strategies in NLP (Bender and Friedman, 2018; Rogers et al., 2021).

[1] https://aclrollingreview.org
[2] https://openreview.net

2 Background

2.1 Peer Review

Modern scholarly peer review is a highly structured process that involves a range of STAKEHOLDERS and produces a range of `artifacts`. A common reviewing campaign involves AUTHORS submitting their `draft` to a reviewing committee represented by EDITORS. The editors distribute the drafts among REVIEWERS who provide their evaluation in form of a `report`. This is followed by an optional revision stage where the authors might provide `rebuttals` (author responses), update their draft, and accompany the updates with `amendments` that describe the changes. At the end of the reviewing stage, EDITORS might produce `meta-reviews` that summarize individual reports and assist decision making. Based on the evaluation outcome, the work is accepted or rejected. Rejected works are subject to resubmission; accepted work might be transferred to the PUBLISHER who is responsible for its archival and dissemination, resulting in a `publication`. Peer reviewing is often anonymized: during the process the reviewer identities are hidden from the authors (single-blind), and the author identities might be hidden from the reviewers (double-blind).

2.2 Computational Study of Peer Review

Reviewing quality and efficiency are of paramount importance to maintaining the integrity of science. Yet issues persist in both dimensions: reviewers are prone to a range of biases and strategic behaviors (Tomkins et al., 2017; Lee et al., 2013; Stelmakh et al., 2020b), fall back on superficial heuristics (Rogers and Augenstein, 2020), and reviewing itself takes a lot of time and effort (GSPR, 2018). This motivates computational study of peer review.

Pioneering the use of NLP for peer review analysis, Kang et al. (2018) introduce PeerRead – a corpus composed of openly available reports and drafts – and report experiments on paper acceptance and aspect score prediction. Hua et al. (2019) annotate reports with argumentation categories and use an automatic discourse segmentation model to study arguments in peer reviews of two research communities. Cheng et al. (2020) propose a new approach and corpus for studying the correspondence between review reports and rebuttals; Gao et al. (2019) investigate the effect of rebuttal on evaluation; Dycke et al. (2021) proposes a preference learning-based approach to paper ranking based on reports and scores. A recent work by Yuan et al. (2021) explores fully-automatic review report generation based on submission drafts. A separate line of studies considers peer review as a process, with recent experiments investigating the influence of reviewer experience (Stelmakh et al., 2020c), herding (Stelmakh et al., 2020a) and resubmission bias (Stelmakh et al., 2021), among others. Computational research in peer review critically depends on the availability of open peer reviewing data.

2.3 Status of Existing Peer Reviewing Data

Existing research on NLP for peer reviews builds almost exclusively on two data sources. The International Conference on Learning Representations (ICLR[3]) makes review reports and discussion threads for accepted and rejected papers openly available via the OpenReview.net platform. The Conference on Neural Information Processing Systems (NeurIPS[4]) has been systematically publishing reviews for accepted papers since 2013, available via the conference website. Both ICLR and NeurIPS represent specialist communities focusing on neural network and representation learning research – a narrow sample given the widespread use of peer reviewing across scientific fields. While peer review at NLP and computational linguistics conferences has been previously studied (Kang et al., 2018; Gao et al., 2019), publicly available data is scarce.

Recent years are marked by raised awareness in ethical implications of natural language processing; multiple recent proposals call for ethics- and copyright-aware data collection strategies and documentation protocols (Rogers et al., 2021; Bender and Friedman, 2018). Peer reviewing data is not exempt from that requirement and presents a challenging case of personal, confidential, anonymous data that needs to be **collected and managed** accordingly: the fact that the peer reviewing data is made openly available and can be crawled does not imply that it can be freely used for research and model development (Rogers et al., 2021). Neither ICLR nor NeurIPS provide information on the authors' and reviewers' consent for **processing** of their peer reviewing data. Additionally, they do not specify the conditions of data processing by third parties. As we discuss below, peer reviewing data

[3] https://iclr.cc
[4] http://neurips.cc

is personal and confidential, and requires consent or other grounds for processing.

The sources of open peer reviewing data previously used in NLP do not attach a license to the artifacts, rendering the conditions of **data reuse** under-specified. As of December 2021, none of the published datasets of peer reviews (incl. PeerRead (Kang et al., 2018), AMPERE (Hua et al., 2019), APE (Cheng et al., 2020) and ASAPReview (Yuan et al., 2021)) attaches clear license to source or to the derivative annotated data. As we show below, publishing of, and attaching license and copyright to peer reviewing data is non-trivial and requires careful consideration of authorship and attribution.

All in all, the current ad-hoc approach to peer reviewing data collection in NLP bears a range of risks: the lack of clearly defined, public, general data collection protocols limits the data collection to narrow research communities, and results in a major overhead for individual data collection efforts; the lack of consent and clear license leaves the conditions of data processing and reuse unclear.

3 Problem Dimensions

We use *peer reviewing data* as an umbrella term for drafts, review reports, amendment notes and meta-reviews. For the sake of presentation we limit our discussion here to `drafts` and `reports` as the two most important and most consistently used artifacts of the peer reviewing data. However, it equally applies to other peer reviewing data types. We distinguish between *metadata* (numerical scores, track, paper format, etc.) and *textual data*. All of the textual data falls under the EU General Data Protection Regulation (GDPR) definition of **personal data** as "any information relating to an identified or identifiable natural person"[5]. Even if the identity of the reviewer or author is not revealed publicly, it is known to the editors, and remains potentially identifiable based on the writers' professional expertise, as well as via text-based profiling. Although peer reviewers and authors rarely sign formal non-disclosure agreements, peer reviewing data is **confidential**. Finally, most of the peer reviewing data is **anonymous**, and only the editors know the identities of the participants.

[5]Although GDPR is EU-based, it is considered the best available practice for privacy-related legislation globally; GDPR regulates processing of the EU subjects' data *anywhere on Earth* and is thus almost certainly applicable to any major text collection, especially in the scientific domain; cf. an extended discussion in (Rogers et al., 2021)

A. Data Collection. As personal data, peer reviewing data requires consent or other explicit grounds for processing. Two main approaches to obtaining consent are *terms of service* (ToS) and *donation*. ToS apply to all users of a platform or service; a platform cannot be used unless the user accepts the ToS. Establishing universal ToS is a challenging task that involves balancing interests of many stakeholder groups, at the risk of losing the authors and reviewers that disagree with the adopted policy. In a donation-based system, the decision to contribute peer reviewing data is made by individual stakeholders on case-by-case basis. Although technically more intricate, donation-based data collection does not interfere with the reviewing process as it still allows participants who do not wish to contribute to use the platform for its main purpose – getting the work evaluated or acting as reviewers. Donation-based approach, however, introduces participation bias (Keeble et al., 2015; Slonim et al., 2013).

B. Data Processing is governed by consent. Who does the peer reviewing data belong to, and who has the authority to give consent? As peer reviewing data is interconnected, this question is not trivial: while a reviewer might agree to publishing their reports, the authors might object, not only due to potential negative reviews, but also due to the risk of leaking unpublished ideas and results prepublication. Ideally one would want all involved stakeholders to consent; however, increasing the number of involved parties means lower agreement among them, and might substantially reduce the amount of collected data and introduce further bias.

C. Data Reuse is governed by the license. Liberal data licensing is crucial for scientific progress as it allows the community to build upon prior work and ensures replicability. Creative commons (CC) is a popular licensing choice for NLP datasets that supports additional restrictions on data sharing, adaption and commercial use. Most CC licenses require attribution – specifying the title, authorship and source of the data. However, as most of the peer reviewing data is anonymous, it cannot be directly attributed to its authors, and declaring the work public domain (CC0) leaves the data reuse entirely unregulated, incl. commercialization, republishing, as well as claiming copyright and attaching restrictive license to data derivatives.

D. Anonymity and Credit. During peer review, the identities of the authors are hidden to maintain the objectivity of review; anonymizing the reviewers aims to protect them from potential backlash. If a submission is accepted for publication, the author identities are made public, and they receive credit for their work; editors receive credit for their organizational work on the whole volume. Submissions that have *not* passed peer review remain uncredited, and so do review reports for both accepted and rejected submissions. While problematic even in a closed-review setting, once review reports and non-accepted submissions are made public as part of a dataset, the authors and reviewers should have an opportunity to be *credited for their work*, which effectively *deanonymizes their contributions*.

E. Confidentiality. Modern academia is highly competitive, and attribution of ideas and discoveries is the key element of scientific communication. While there exists a well-established attribution mechanism for archived publications – citations – attributing non-archival content (e.g. social network posts or blogs) is not regulated. Review reports are closely related to the submissions they discuss, and often summarize and analyze their content in a way that enables a third party to appropriate the idea or to gain advantage due to the knowledge of unpublished results. Professional ethics prevent idea theft via peer review, as the identities of the reviewers are still known to the editors. However, when peer reviewing data is made available to the open public, this is no longer true, as the access to yet unpublished research results and insights can no longer be tracked, presenting a confidentiality risk.

4 The Yes-Yes-Yes Workflow

4.1 Design Decisions

The aforementioned problem dimensions inform the design of our proposed workflow. We aim to grant AUTHORS and REVIEWERS extensive control over their associated data, while maximizing the value of the resulting reviewing data to researchers. Both goals should be attained while ensuring least interference with the peer reviewing campaign and avoiding pressure on the STAKEHOLDERS.

We collect data strictly on a **donation basis** (Section 3.A) accepting the likely participation bias in our data, which we discuss in more detail later on. The primary contributor of the data is the stakeholder producing the artifact (B.); this means, `drafts` must be donated by the AUTHORS, and the `reports` by REVIEWERS. The stability and availability of the dataset is crucial for replicability of research results (Zubiaga, 2018; Rogers et al., 2021), however, simple consent does not guarantee data persistence, as *it can be withdrawn*, which would in turn require modification of the underlying data; it is preferable to perform **license transfer** (C.): as long as the license conditions are met, the license can not be revoked, and the research dataset remains stable. Reviewers and authors of unpublished papers must have an opportunity to explicitly request being attributed (D.), with the identity anonymous by default. Finally, to account for the confidentiality (E.), additional permission to make `reports` public must be obtained from the AUTHORS. Reports for which only REVIEWERS opt-in can be subject to research, but cannot be made public.

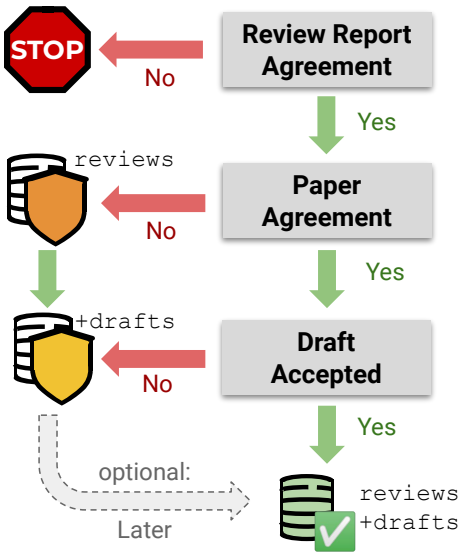

Figure 1: The decision process of the 3Y-Workflow.

4.2 Workflow

Based on these design decisions, we define the Yes-Yes-Yes workflow (3Y-Workflow) for peer reviewing data collection (Figure 1). The 3Y-Workflow is a three-step decision process synchronized with the underlying *peer reviewing campaign* and applied on per paper and per reviewer basis. The workflow yields three possible outcomes: *no data* collected (default), data added to a *protected dataset* (accessible for internal research, but not public), and data added to a *public dataset*. In all cases, the resulting

data is anonymous unless credit was explicitly requested by data contributors. The protected dataset is confidential: it may be used for research purposes and to provide statistics on the underlying peer reviewing campaign, but shall not be made public. In the following, we describe each step of the workflow in detail.

Yes by the REVIEWERS In the first step of the workflow, each REVIEWER decides on contributing their reports in the given reviewing campaign. To minimize the communication overhead, reviewers make a decision whether to donate all their reviewing reports of a given reviewing campaign, in bulk. To contribute, a reviewer signs a *review report license agreement* with optional attribution (see A.1). The donation can be made any time between submission and acceptance decisions. The reviewers are explicitly informed about the risks of authorship attribution via profiling techniques. If the REVIEWER should not explicitly give the "first Yes", their reports are discarded from the data collection pipeline.

Yes by the AUTHORS In the second step, AUTHORS decide on donating their drafts. If they wish to donate, they sign a *paper license agreement* (see A.2) combined with the permission to publish the associated review reports from the REVIEWERS that gave their "Yes" in the first step. The donation is made after the outcomes of the reviewing campaign are known to provide the AUTHORS with full information. If the authors donate their data ("second Yes"), the decision workflow continues. If the authors do not wish to contribute, the donated review reports become part of the protected dataset, but the drafts are discarded from the data collection.

Yes by the EDITORS Finally, if both reviewers and authors agree and a submission is accepted for publication by EDITORS ("third Yes"), the draft and its donated review reports become eligible for the public dataset. If the draft is not accepted but both reviewers and authors previously agreed to donate their data, the associated data remains part of the protected dataset. Unlike the cases where the authors did not give their "Yes", the drafts and reviews discussed here *can* be publicly released after a significant amount of time sufficient to maintain confidentiality of the research ideas. If this applies, it should be explicitly mentioned as part of the license agreement and consent statement.

4.3 Pros and Cons

By design of the 3Y-Workflow, all STAKEHOLDERS make well-informed decisions that are decoupled from the peer reviewing campaign: no particular reviewing scheme is enforced and the decision for donation does not affect the reviewing outcome. The anonymity is preserved while receiving credit still remains possible; the confidentiality is ensured by releasing data only after agreement by all stakeholders; license transfer guarantees dataset stability and enables replicability. The workflow makes minimal assumptions about the underlying peer reviewing campaign and can be easily adapted to most reviewing campaigns according to the needs of the respective research community and data collector.

An ethical, donation-based workflow naturally introduces structural biases that need explicit consideration. Participation bias of REVIEWERS and AUTHORS is likely: the decision to donate the data might correlate with the evaluation outcome, background, career stage and demographic. As the inclusion of data into the public dataset is tied to the publication of the paper (step 3), additional bias towards accepted submissions is introduced. Finally, as peer review changes over time (Lee et al., 2013), the time lag between data collection and dissemination might introduce historical bias. The relevance of these types of bias to NLP depends on the application and remains an open research question; although not perfectly representative, the non-public protected dataset can serve as a point of reference for studying and quantifying these biases.

5 Data Collection at ARR

ACL Rolling Review (ARR) is an initiative in the ACL community that decouples peer review from publication and replaces the traditional, per-event reviewing campaigns with a single, journal-style reviewing process. ARR has been launched in May 2021 and serves as the main reviewing platform for multiple major ACL conferences, incl. the Annual Meeting of ACL[6]. ARR operates in monthly cycles: during each cycle, the AUTHORS might submit their work to ARR; the draft is evaluated by REVIEWERS; based on the evaluation, action editors decide whether the draft has passed peer review. If the evaluation is positive, the draft can be committed to one of the ACL conferences where program chairs (equivalent to EDITORS) make the final decision to publish the work. If the a draft

[6] https://www.2022.aclweb.org

is not accepted at a conference, it can be revised and resubmitted to ARR in next iterations. For the 3Y-Workflow, only the publication decision by the program chairs is relevant.

ARR presents a unique opportunity for the study of peer review in the ACL community and beyond. The ever-increasing submission rates at ACL provide a steady source of reviewing data, and unified reviewing workflow, protocols and forms minimize the effects of a particular reviewing campaign configuration on the process. The use of the open-source OpenReview.net as platform makes it easy to automate many aspects of peer reviewing data collection, from sending out reminders to secure data filtering. With kind permission and support by the editors-in-chief and the technical team, we have implemented the 3Y-Workflow at ARR.

Implementation. To minimize interference of the data collection with the peer reviewing campaign, our implementation relies on the built-in `Task` feature of OpenReview.net: optional data donation is seamlessly integrated as part of the reviewing process along with other tasks, like review submission. To enable future research on the collected data while preventing uncontrolled re-use and redistribution, we attach the Creative Commons BY-NC-SA 4.0 License to the data which allows future users to share and adapt the data as long as it is attributed (BY), only used non-commercially (NC) and is shared under the same licensing conditions (share-alike, SA)[7]. To avoid the pitfall of reviewing data being non-attributable due to anonymity, we ask the contributors to perform a *license transfer* in which the copyright for the data is transferred to the Association for Computational Linguistics (similar to the ACL Anthology[8] publications), while the data creators might still get attributed if they explicitly wish to reveal their identity.

To protect the confidentiality of the unpublished results, we opt to make public exclusively the peer reviewing data of the papers that are later accepted at some venue and officially published. Our implementation of 3Y-Workflow is open and available[9] making the data extraction code base transparent, and allowing to easily set up the workflow for any new OpenReview.net-based reviewing campaign independent of the venue or research field.

[7] https://creativecommons.org/licenses/by-nc-sa/4.0/

[8] https://aclanthology.org

[9] https://github.com/UKPLab/openreview-licensing-workflow

6 Analysis

As of 20th December 2021, we collected licenses for review reports of July, September and October. Since the publication of submissions is decoupled from ARR, the authors of the corresponding drafts are still due to make their decision to contribute their data, at the time of this writing. However, the already-donated anonymous peer review reports and their associated metadata constitute a *protected dataset* which makes it possible to get a first glance on the results of the collection process and compare the donated data to prior datasets, focusing on peer review reports. As a reference we use the ACL-2018 dataset introduced by Gao et al. (2019): originating from the same community, the dataset was collected during the peer reviewing campaign of the 56th Annual Meeting of ACL. Unlike in the 3Y-Workflow, consent for data processing was obtained *during* peer review as part of the review form, resulting in higher participation, but raising concerns about the stress of the reviewing influencing the decision to contribute the data. Coupled with a lack of license transfer for review texts, this prevented the publication of the full dataset; however, numerical data from the ACL-2018 has been made public. We note that ACL-2018 reviewing forms and score semantics are different from ARR, which limits us in making direct fine-grained comparisons between the two data collection campaigns; still, the common elements allow us to contextualize our results.

6.1 Data composition

We first explore the basic characteristics of the donated data (ARR-3Y) and compare it to limited statistics derived from the complete data of ARR, and to related work. As Table 1 shows, even though ARR has not yet reached its full capacity, ARR complete data from July, September and October 2021 already exceeds the ICLR-2017 portion of the PeerRead corpus (Kang et al., 2018) and is likely to exceed the ACL-2018 in the near future. As the 3Y-Workflow statistics show, not all of this data is donated; however, as the workflow is applied *continuously*, the size of the protected dataset is likely to exceed ACL-2018 in the near future.

A donation-based collection process affects the number of reviews per submission and per reviewer,

	ARR-3Y	ARR complete	ACL-2018	PeerRead$_{\text{ICLR}}$
# Submissions	567	646	1538	427
# Reviews	1053	2118	3875	1304
# Reviewers	707	1737 $\downarrow$	1213	-
# Reviews per submission	1.86 ± 0.84	3.28^{*}	2.52 ± 0.67	-
# Reviews per reviewer	1.49 ± 0.85	1.22 $\uparrow$	3.04 ± 1.35	-

Table 1: Statistics of donated (ARR-3Y) and complete ARR data for July, September, October 2021. Reviewer statistics for ACL-2018 from Dycke et al. (2021); statistics on ICLR-2017 portion of the PeerRead corpus from Kang et al. (2018). $\downarrow$ upper bound, $\uparrow$ lower bound, $*$ estimated from total counts.

which might pose a limitation for peer reviewing assistance tasks that require construction of a reviewer-paper graph, such as (Dycke et al., 2021). However, since each iteration of the ARR is treated as a separate reviewing campaign, we can only report statistics on the accumulated number of unique reviewers per iteration for the complete ARR data: in other words, if an individual has reviewed at ARR in July and in August, they would be counted twice. Hence, Table 1 reports estimated upper ($\downarrow$) and lower ($\uparrow$) bounds: in reality, the number of reviewers at ARR is likely lower, and the number of reviews per reviewer is likely higher.

6.2 Bias and Review Scores

We expect the decision to donate peer reviewing data to correlate with a range of factors, including demographics and the sentiment of the reviews. While the current data collection protocol does not collect demographic data, review scores can be used as a proxy for review sentiment to investigate biases towards positive reviews limiting the representativity of the ARR-3Y dataset. We investigate two central questions.

Does donation-based review collection introduce positive review bias? To investigate, we analyse the peer review score distributions. Figure 2 compares the distribution of overall scores across donated reviews in ARR-3Y to all reviews in the complete ARR data for the same iterations. As it shows, the donated reviews cover a wide range of ratings, with prevalent overall score around 3 ("good") and 2 ("revisions needed"). The distribution of overall scores in the donated subset is nearly identical to the one of the complete population showing that the donated subset does not solely focuses on positive evaluations. Notably, the shape of the distribution resembles other computer science conferences (Ragone et al., 2013) and

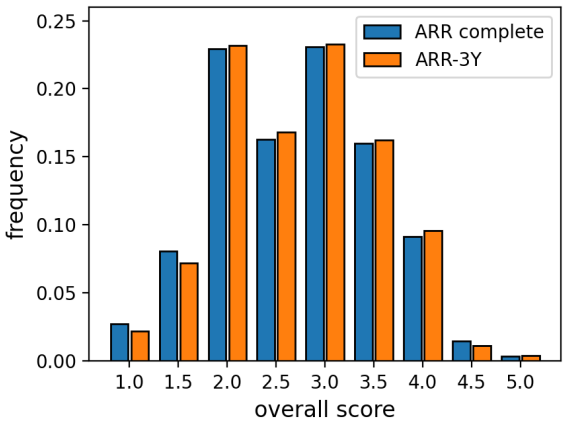

Figure 2: Overall score distribution in ARR-3Y and ARR complete reviews.

ACL-2018 (Gao et al., 2019).

Are reviews for controversial submissions less likely to be donated? We investigate this by comparing the agreement on overall scores per submission measured by Krippendorff's α with ordinal metric (Krippendorff, 1980). The agreement on overall scores lies at 0.24 for ARR-3Y, considerably lower than 0.36 for ACL-2018 before rebuttal reviews (Dycke et al., 2021), indicating that reviews for controversial submissions indeed get donated, even when taking into account the differences in score scale semantics between ACL-2018 and ARR. This is supported by the observation that the mean overall score per submission of 2.73 ± 0.79 (for the 40% submissions with more than one review) lies around the center of the scale, which does not indicate that only reviews for papers above clear acceptance or revision thresholds are covered.

6.3 Reviewing Behavior

We now turn to the qualitative analysis and insights enabled by the review report data collected so far.

ARR uses semi-structured peer review forms that include a range of textual fields (incl. strengths S⋆ and weaknesses W⋆) as well as numerical aspect scores (software SF, reproducibility RP, datasets DT), and confidence (CF), author identity guess (AI) and overall (OV) scores. The correlations between the individual score values and text field lengths (Figure 3) provide rich qualitative insights in the reviewing behavior at ARR.

Does author identity guess affect the scores? Author anonymity is a key feature of double-blind peer review designed to promote fair evaluation of the submissions. The author identity guess score reported by the reviewers shows low correlation to all other scores, accompanied by the fact that roughly 88% of the reviews report "no educated guess". This shows that on the *system level* identity guess is a rare phenomenon at ARR and does not substantially contribute to the scoring behavior. On the *individual level*, however, the results are different: if only the cases of non-zero AI are considered, we discover a moderate negative (-0.14 Pearson) correlation between AI and the overall score. An in-depth qualitative analysis of the identity guess cases could provide further insights into this behavior.

What aspects matter for evaluation? The correlations between individual aspect scores reveal that reproducibility has a major impact on the overall scoring and is, in turn, correlated with the authors providing software and contributing open datasets. The mean confidence per reviewer lies at 3.72 ± 0.74 close to 4 ("quite sure" in the ARR review forms). We point out a weak negative correlation between the overall score and reviewer confidence, consistent with previous findings on ACL-2018 data by Gao et al. (2019). The direction of the relationship between reviewer confidence and overall score remains an open question: we hypothesize that submissions get negative evaluation for specific flaws that the reviewer can clearly point to (hence high confidence), but positive evaluation is based on the perceived general merit and absence of flaws, which are either truly absent, or go unnoticed (lower confidence). Controlled experiments and more fine-grained peer review forms would help to shed light on this phenomenon.

How do review scores relate to text? Moderate positive correlation between the length of the weaknesses section and reviewer confidence (Figure 3) speaks in favor of our specificity hypothesis outlined above. With respect to the overall score, longer strengths sections are associated with higher overall scores and longer weaknesses sections with a lower overall score. We note a positive correlation between S⋆ and W⋆ – we assume that reviewers who write more in general, do so when discussing both weaknesses and strengths. Finally, we observe that having a valuable dataset or software correlates with the length of the strength section, but lower scores on these aspects do not seem to interact with the length of the weaknesses section. As text length is only a weak proxy of the textual content of peer reviews, advanced NLP processing of the peer review texts (e.g. discourse analysis (Hua et al., 2019)) would enable more fine-grained insights into the interplay between reviewing scores and textual feedback that accompanies them.

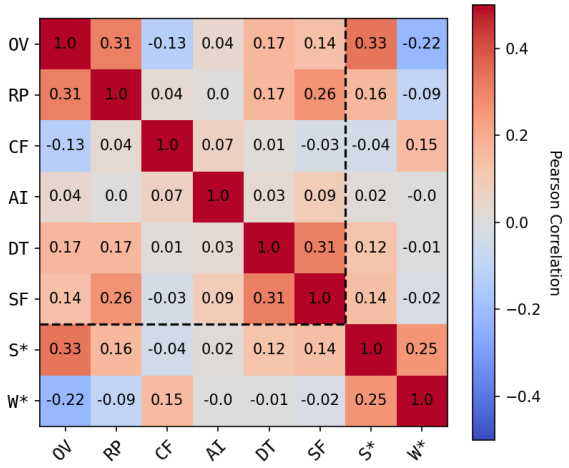

Figure 3: Correlation of review scores and text section lengths. Dashed lines separate text sections.

6.4 Participation

A donation-based data collection workflow crucially depends on the individual STAKEHOLDERS' participation, and we conclude our analysis with a brief overview of our observations related to donation behavior.

Over the course of the considered three months, 885 responses to the donation request were collected from 1737 active reviewers (each cycle of ARR treated individually). Among these responses 4.85% explicitly disagreed to data collection, while the rest agreed. In addition, 36% of the contributors requested attribution, showing the demand for getting credit for the hard peer reviewing work. At the same time, the majority of the reviewers still prefer to stay anonymous.

By the implementation of the 3Y-Workflow at ARR, reviewers are free to sign the agreement before, during or after writing their review reports for each cycle. Interestingly, 50.3% of donating reviewers agreed to contribute their data before submitting their first review report of a cycle, while 45.2% do it after submitting their last reviewing report. This justifies leaving the decision timing up to reviewers and suggests that the decision for donation is only weakly influenced by the outcome of the review. Whether the prospect of review publication at a later point influences the review-writing process per se, remains an open research question.

7 Discussion

The 3Y-Workflow ties publication of peer reviewing data to the acceptance of the underlying submissions. Our analysis so far is limited to the protected dataset, which cannot be published due to its confidentiality. Yet, this dataset can already provide valuable insights into the peer review and data collection process enabled by the 3Y-Workflow, as well as serve as a basis for bias analysis for the public dataset of the 3Y-Workflow in the future.

Our analysis of review score distributions suggests that the protected subset of donated reviews is not skewed towards benign reviews. However, since the publication of data is tied to submission acceptance, the final, public dataset is likely to be biased towards high-quality papers. This highlights the importance of the protected dataset and bias-awareness in NLP research on peer reviews. One solution to mitigate the bias in the future is to publish the peer reviewing data in cases where both author and reviewers agreed to donate the data, but the submission did not pass the review and get officially published and archived. This should happen after a substantial time period to ensure the confidentiality of research ideas; at the point of writing, this mechanism is *not* implemented at ARR and only the data for officially published submissions will be made public.

The low rate of explicit disagreement to donation (Section 6.4) points at the potential weakness of the current 3Y-Workflow implementation at ARR: as responding to the donation request is not enforced, it remains unknown whether the lack of response indicates explicit refusal to donate the data or simply missing out on the opportunity. In the future, we plan to experiment with both technical and communication means to improve the outreach of the data donation campaign and collect feedback on the reasons that drive the decision to contribute.

The final key challenge of the 3Y-Workflow implementation at ARR is the public dataset construction and distribution. Unlike static NLP datasets that are compiled and archived once, the dataset resulting from a continuous reviewing process is dynamic, which requires additional engineering and organizational effort, as well as explicit version control and an update announcement mechanism. In the meantime, ways to allow experimentation with protected data are to be explored, incl. publication of non-confidential, non-personal numerical metadata and fine-grained statistics – as our analysis in Section 6 shows, this data alone enables novel insights into the peer reviewing process.

8 Conclusion

We have presented "Yes-Yes-Yes" – the first explicit, donation-based workflow for collecting peer reviewing data in NLP. We have discussed the core dimensions and challenges of peer reviews as data type, detailed our proposed workflow, and reported on its ongoing implementation at ACL Rolling Review. The data collected so far provided new insights into the peer reviewing process and practical aspects of peer reviewing data collection.

In our description of the workflow we focused on collecting peer review reports and manuscript drafts. The 3Y-Workflow is equally applicable to other peer reviewing artifacts: for example, in addition to review reports and blind submissions, the ARR implementation incorporates collection of draft revisions. The overall structure of the process and the guiding principles for this additional data remain the same: no data is collected without explicit consent, textual data is associated with a liberal license, and the AUTHORS have a say in publication of any textual content related to their paper to protect the confidentiality of their research.

The proposed workflow is not tied to the particularities of ARR and can be easily adjusted to alternative peer reviewing configurations. For example, the ICLR conference – a major source of peer reviewing data in NLP – makes both accepted and rejected papers publicly available; thereby the confidentiality of research ideas is of less concern, but consent and data license still have to be collected from the participants of the reviewing process to make the derivative datasets suitable for research both ethically and legally.

Research communities widely differ in peer reviewing and publishing standards, but the core evaluation schema in peer review is similar across disciplines. While the decision to make peer reviewing data available for research remains with the corresponding communities, our proposed workflow and its implementation provide the conceptual and technological solution to ensure that the collected data can be used for research, both in NLP and beyond.

Acknowledgements

We express our sincere gratitude to all parties providing support and advice during the realization of the peer review data collection at ACL Rolling Review. The data collection would not have been possible without the discussion and approval by the ACL Committee on Reviewing in 2021. We thank the editors-in-chief and the technical team of ACL Rolling Review for their support during this ongoing data collection effort; with a special thanks to Amanda Stent and Sebastian Riedel. Finally, we thank Dorothy Deng for legal counseling and the specification of the review report and paper draft license agreement texts.

References

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.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011, Online. ACL.

Nils Dycke, Edwin Simpson, Ilia Kuznetsov, and Iryna Gurevych. 2021. Ranking scientific papers using preference learning. *arXiv preprint arXiv:2109.01190*.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major nlp conference. 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 1274–1290.

GSPR. 2018. *Global State of Peer Review 2018*. Wellington: Publons.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. 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 2131–2137, Minneapolis, Minnesota. ACL.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661.

Claire Keeble, Graham Richard Law, Stuart Barber, Paul D Baxter, et al. 2015. Choosing a method to reduce selection bias: a tool for researchers. *Open Journal of Epidemiology*, 5(3):155–162.

K. Krippendorff. 1980. *Content Analysis: An Introduction To Its Methodology*. Sage Publications, Beverly Hills.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17.

Azzurra Ragone, Katsiaryna Mirylenka, Fabio Casati, and Maurizio Marchese. 2013. On peer review in computer science: Analysis of its effectiveness and suggestions for improvement. *Scientometrics*, 97(2):317–356.

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*, pages 1256–1262, Online. ACL.

Anna Rogers, Tim Baldwin, and Kobi Leins. 2021. Just what do you think you're doing, dave?'a checklist for responsible data use in nlp. *arXiv preprint arXiv:2109.06598*.

Robert Slonim, Carmen Wang, Ellen Garbarino, and Danielle Merrett. 2013. Opting-in: Participation bias in economic experiments. *Journal of Economic Behavior & Organization*, 90:43–70.

Ivan Stelmakh, Charvi Rastogi, Nihar B Shah, Aarti Singh, and Hal Daumé III. 2020a. A large scale randomized controlled trial on herding in peer-review discussions. *arXiv preprint arXiv:2011.15083*.

Ivan Stelmakh, Nihar B Shah, and Aarti Singh. 2020b. Catch me if i can: Detecting strategic behaviour in peer assessment. In *ICML Workshop on Incentives in Machine Learning*.

Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. 2020c. A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. *arXiv preprint arXiv:2011.15050*.

Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. 2021. Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–17.

Andrew Tomkins, Min Zhang, and William D. Heavlin. 2017. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing? *arXiv preprint arXiv:2102.00176*.

Arkaitz Zubiaga. 2018. A longitudinal assessment of the persistence of twitter datasets. *Journal of the Association for Information Science and Technology*, 69(8):974–984.

A Appendix

A.1 Review Report License Agreement

Association for Computational Linguistics Peer Reviewer Content License Agreement

Name of ACL Conference: `cycle name`

Peer Reviewer's Name: `reviewer identity`

* Unless the peer reviewer elects to be attributed according to Section 2, the peer reviewer's name will not be identified in connection with publication of the Peer Review Content. If you wish to be attributed, please check this box ☐.

This Peer Reviewer Content License Agreement ("Agreement") is entered into between the Association for Computational Linguistics ("ACL") and the Peer Reviewer listed above in connection with content developed and contributed by Peer Reviewer during the peer review process (referred as "Peer Review Content").

In exchange of adequate consideration, ACL and the Peer Reviewer agree as follows:

1. Grant of License. Peer Reviewer grants ACL a worldwide, irrevocable, and royalty-free license to use the Peer Review Content developed and prepared by Peer Reviewer in connection with the peer review process for the ACL Conference listed above, including but not limited to text, review form scores and metadata, charts, graphics, spreadsheets, and any other materials according to the following terms:
   (a) For Peer Review Content associated with papers accepted for publication, and subject to the Authors permission, ACL may reproduce, publish, distribute, prepare derivative work, and otherwise make use of the Peer Review Content, and to sublicense the Peer Review Content to the public according to terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
   (b) For Peer Review Content associated with papers not accepted for publication, ACL may use the Peer Review Content for internal research, program analysis, and record- keeping purposes. Notwithstanding the foregoing, the Parties acknowledge and agree that this Agreement does not transfer to ACL the ownership of any proprietary rights pertaining to the Peer Review Content, and that Peer Review retains respective ownership in and to the Peer Review Content.
2. Attribution and Public Access License.
   (a) The Parties agree that for purpose of administering the public access license, ACL will be identified as the licensor of the Content with the following copyright notice: Copyright © 2021 administered by the Association for Computational Linguistics (ACL) on behalf of ACL content contributors: ... (list names of peer reviewers who wish to be attributed), and other contributors who wish to remain anonymous. Content displayed on this webpage is made available under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
   (b) In the event Peer Reviewer intends to modify the attribution displayed in connection with the copyright notice above, ACL will use reasonable efforts to modify the copyright notice after receipt of Peer Reviewer's written request. Notwithstanding the foregoing, Peer Reviewer acknowledges and agrees that any modification in connection with attribution will not be retroactively applied.
   (c) The Parties understand and acknowledge that the reative Commons Attribution-NonCommercial-ShareAlike 4.0 International License is irrevocable once granted unless the licensee breaches the license terms.
3. Warranty. Peer Reviewer represents and warrants that the Content is Peer Reviewer's original work and does not infringe on the proprietary rights of others. Peer Reviewer further warrants that he or she has obtained all necessary permissions from any persons or organizations whose materials are included in the Content, and that the Content includes appropriate citations that give credit to the original sources.
4. Legal Relationship. The Parties agree that this Agreement is not intended to create any joint venture, partnership, or agency relationship of any kind; and both agree not to contract any obligations in the name of the other.

Signature: `signature`, Date: `date`
Name Typed: `name`

A.2 Paper License Agreement

Association for Computational Linguistics Blind Submission License Agreement

Name of ACL Conference: `cycle name`
Blind Submission Paper Title: `title`
List Authors' Names: `author identifiers`
* Authors names will not be shared with the peer reviewers during the peer review process This Blind Submission License Agreement ("Agreement") is entered into between the Association for Computational Linguistics ("ACL") and the Authors listed in connection with Authors' blind submission paper listed above (referred as "Blind Submission Content"). In exchange of adequate consideration, ACL and the Authors agree as follows:

1. Grant of License. After the peer review process is concluded and upon acceptance of the paper, Authors grant ACL a worldwide, irrevocable, and royalty-free license to use the blind submission paper version (referred as "Content"). The foregoing license grants ACL the right to reproduce, publish, distribute, prepare derivative work, and otherwise make use of the Content, and to sublicense the Content to the public according to terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. Notwithstanding the foregoing, the Parties acknowledge and agree that this Agreement does not transfer to ACL the ownership of any proprietary rights pertaining to the Content, and that the Authors retain their respective ownership in and to the Content.

2. Permission to Publish Peer Reviewers Content. After the peer review process is concluded and upon acceptance of the paper, Authors have the option to grant ACL permission to publish peer reviewer's content associated with the Content, which may include text, review form scores and metadata, charts, graphics, spreadsheets, and any other materials developed by peer reviewers in connection with the peer review process.
   ☐ Authors grant permission for ACL to publish peer reviewers content
   ☐ Authors decline to grant permission for ACL to publish peer reviewers content

3. Attribution and Public Access License.
   (a) The Parties agree that for purpose of administering the public access license, ACL will be identified as the licensor of the Content with the following copyright notice: Copyright © 2021 administered by the Association for Computational Linguistics (ACL) on behalf of the authors and content contributors. Content displayed on this webpage is made available under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
   (b) The Parties understand and acknowledge that the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License is irrevocable once granted unless the licensee breaches the license terms.

4. Effective Date. The grant of license pursuant to Section 1 and permission to publish peer reviewers content pursuant to Section 2 becomes effective in the event Authors' blind submission paper is accepted for publication by ACL. If the blind submission paper is not accepted, the Content and associated peer reviewers content will remain confidential and kept for internal record-keeping purpose only.

5. Warranty. Authors represent and warrant that the Content is Authors' original work and does not infringe on the proprietary rights of others. Authors further warrant that they have obtained all necessary permissions from any persons or organizations whose materials are included in the Content, and that the Content includes appropriate citations that give credit to the original sources.

6. Legal Relationship. The Parties agree that this Agreement is not intended to create any joint venture, partnership, or agency relationship of any kind; and both agree not to contract any obligations in the name of the other.

By signing below, I confirm that all Authors have agreed to the above terms and that I am authorized to execute this Agreement on their behalf.
Signature `signature`, Date `date`
Name (please print) `author's name`

A.3 Review Forms at ARR

In the following, the review forms for each cycle of ARR are listed. While these remain mostly constant, there might occur minor changes throughout the considered months. These changes relative to the form of the previous month are underlined in the respective forms for the ease of reading. The revisions observed during the considered months concern typos and character limitations for free text fields.

A.3.1 Review Form July

Paper Summary: Describe what this paper is about. This should help action editors and area chairs to understand the topic of the work and highlight any possible misunderstandings. Maximum length 1000 characters.

Summary Of Strengths: What are the major reasons to publish this paper at a selective *ACL venue? These could include novel and useful methodology, insightful empirical results or theoretical analysis, clear organization of related literature, or any other reason why interested readers of *ACL papers may find the paper useful. Maximum length 5000 characters.

Summary Of Weaknesses: What are the concerns that you have about the paper that would cause you to favor prioritizing other high-quality papers that are also under consideration for publication? These could include concerns about correctness of the results or argumentation, limited perceived impact of the methods or findings (note that impact can be significant both in broad or in narrow sub-fields), lack of clarity in exposition, or any other reason why interested readers of *ACL papers may gain less from this paper than they would from other papers under consideration. Where possible, please number your concerns so authors may respond to them individually. Maximum length 5000 characters.

Comments, Suggestions And Typos: If you have any comments to the authors about how they may improve their paper, other than addressing the concerns above, please list them here. Maximum length 5000 characters.

Overall Assessment:

- 5 = Top-Notch: This paper has great merit, and easily warrants acceptance in a *ACL top-tier venue.
- 4.5
- 4 = Strong: This paper is of significant interest (for broad or narrow sub-communities), and warrants acceptance in a top-tier *ACL venue if space allows.
- 3.5
- 3 = Good: This paper is of interest to the *ACL audience and could be published, but might not be appropriate for a top-tier publication venue. It would likely be a strong paper in a suitable workshop.
- 2.5
- 2 = Borderline: This paper has some merit, but also significant flaws. It does not warrant publication at top-tier venues, but might still be a good pick for workshops.
- 1.5
- 1 = Poor: This paper has significant flaws, and I would argue against publishing it at any *ACL venue.

Confidence:

- 5 = Positive that my evaluation is correct. I read the paper very carefully and am familiar with related work.
- 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
- 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.
- 2 = Willing to defend my evaluation, but it is fairly likely that I missed some details, didn't understand some central points, or can't be sure about the novelty of the work.
- 1 = Not my area, or paper is very hard to understand. My evaluation is just an educated guess.

Best Paper: Could this be a best paper in a top-tier *ACL venue?

- Yes
- Maybe
- No

Best Paper Justification: If the answer on best paper potential is Yes or Maybe, please justify your decision.

Replicability: Will members of the ACL community be able to reproduce or verify the results in this paper?

- 5 = They could easily reproduce the results.
- 4 = They could mostly reproduce the results,

but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
- 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.
- 2 = They would be hard pressed to reproduce the results: The contribution depends on data that are simply not available outside the author's institution or consortium and/or not enough details are provided.
- 1 = They would not be able to reproduce the results here no matter how hard they tried.

Datasets: If the authors state (in anonymous fashion) that datasets will be released, how valuable will they be to others?

- 5 = Enabling: The newly released datasets should affect other people's choice of research or development projects to undertake.
- 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.
- 3 = Potentially useful: Someone might find the new datasets useful for their work.
- 2 = Documentary: The new datasets will be useful to study or replicate the reported research, although for other purposes they may have limited interest or limited usability. (Still a positive rating)
- 1 = No usable datasets submitted.

Software: If the authors state (in anonymous fashion) that their software will be available, how valuable will it be to others?

- 5 = Enabling: The newly released software should affect other people's choice of research or development projects to undertake.
- 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.
- 3 = Potentially useful: Someone might find the new software useful for their work.
- 2 = Documentary: The new software will be useful to study or replicate the reported research, although for other purposes it may have limited interest or limited usability. (Still a positive rating)
- 1 = No usable software released.

Author Identity Guess: Do you know the author identity or have an educated guess?

- 5 = From a violation of the anonymity-window or other double-blind-submission rules, I know/can guess at least one author's name.
- 4 = From an allowed pre-existing preprint or workshop paper, I know/can guess at least one author's name.
- 3 = From the contents of the submission itself, I know/can guess at least one author's name.
- 2 = From social media/a talk/other informal communication, I know/can guess at least one author's name.
- 1 = I do not have even an educated guess about author identity.

Ethical Concernes: Independent of your judgement of the quality of the work, please review the ACL code of ethics (https://www.aclweb.org/portal/content/acl-code-ethics) and list any ethical concerns related to this paper. Maximum length 2000 characters.

A.3.2 Review Form September

Paper Summary: Describe what this paper is about. This should help action editors and area chairs to understand the topic of the work and highlight any possible misunderstandings. Maximum length 20000 characters.

Summary Of Strengths: What are the major reasons to publish this paper at a selective *ACL venue? These could include novel and useful methodology, insightful empirical results or theoretical analysis, clear organization of related literature, or any other reason why interested readers of *ACL papers may find the paper useful. Maximum length 20000 characters.

Summary Of Weaknesses: What are the concerns that you have about the paper that would cause you to favor prioritizing other high-quality papers that are also under consideration for publication? These could include concerns about correctness of the results or argumentation, limited perceived impact of the methods or findings (note that impact can be significant both in broad or in narrow sub-fields), lack of clarity in exposition, or any other reason why interested readers of *ACL papers may gain less from this paper than they would from other papers under consideration. Where possible, please number your concerns so authors may respond to them individually. Maximum length 20000 characters.

Comments, Suggestions And Typos: If you have any comments to the authors about how they

may improve their paper, other than addressing the concerns above, please list them here. Maximum length 20000 characters.

Overall Assessment:

- 5 = Top-Notch: This paper has great merit, and easily warrants acceptance in a *ACL top-tier venue.
- 4.5
- 4 = Strong: This paper is of significant interest (for broad or narrow sub-communities), and warrants acceptance in a top-tier *ACL venue if space allows.
- 3.5
- 3 = Good: This paper is of interest to the *ACL audience and could be published, but might not be appropriate for a top-tier publication venue. It would likely be a strong paper in a suitable workshop.
- 2.5
- 2 = Borderline: This paper has some merit, but also significant flaws. It does not warrant publication at top-tier venues, but might still be a good pick for workshops.
- 1.5
- 1 = Poor: This paper has significant flaws, and I would argue against publishing it at any *ACL venue.

Confidence:

- 5 = Positive that my evaluation is correct. I read the paper very carefully and am familiar with related work.
- 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
- 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.
- 2 = Willing to defend my evaluation, but it is fairly likely that I missed some details, didn't understand some central points, or can't be sure about the novelty of the work.
- 1 = Not my area, or paper is very hard to understand. My evaluation is just an educated guess.

Best Paper: Could this be a best paper in a top-tier *ACL venue?

- Yes
- Maybe
- No

Best Paper Justification: If the answer on best paper potential is Yes or Maybe, please justify your decision.

Replicability: Will members of the ACL community be able to reproduce or verify the results in this paper?

- 5 = They could easily reproduce the results.
- 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
- 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.
- 2 = They would be hard pressed to reproduce the results: The contribution depends on data that are simply not available outside the author's institution or consortium and/or not enough details are provided.
- 1 = They would not be able to reproduce the results here no matter how hard they tried.

Datasets: If the authors state (in anonymous fashion) that datasets will be released, how valuable will they be to others?

- 5 = Enabling: The newly released datasets should affect other people's choice of research or development projects to undertake.
- 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.
- 3 = Potentially useful: Someone might find the new datasets useful for their work.
- 2 = Documentary: The new datasets will be useful to study or replicate the reported research, although for other purposes they may have limited interest or limited usability. (Still a positive rating)
- 1 = No usable datasets submitted.

Software: If the authors state (in anonymous fashion) that their software will be available, how valuable will it be to others?

- 5 = Enabling: The newly released software should affect other people's choice of research or development projects to undertake.
- 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.
- 3 = Potentially useful: Someone might find

the new software useful for their work.

- 2 = Documentary: The new software will be useful to study or replicate the reported research, although for other purposes it may have limited interest or limited usability. (Still a positive rating)
- 1 = No usable software released.

Author Identity Guess: Do you know the author identity or have an educated guess?

- 5 = From a violation of the anonymity-window or other double-blind-submission rules, I know/can guess at least one author's name.
- 4 = From an allowed pre-existing preprint or workshop paper, I know/can guess at least one author's name.
- 3 = From the contents of the submission itself, I know/can guess at least one author's name.
- 2 = From social media/a talk/other informal communication, I know/can guess at least one author's name.
- 1 = I do not have even an educated guess about author identity.

Ethical Concernes: Independent of your judgement of the quality of the work, please review the ACL code of ethics (https://www.aclweb.org/portal/content/acl-code-ethics) and list any ethical concerns related to this paper. Maximum length 10000 characters.

A.3.3 Review Form October

Paper Summary: Describe what this paper is about. This should help action editors and area chairs to understand the topic of the work and highlight any possible misunderstandings. Maximum length 20000 characters.

Summary Of Strengths: What are the major reasons to publish this paper at a selective *ACL venue? These could include novel and useful methodology, insightful empirical results or theoretical analysis, clear organization of related literature, or any other reason why interested readers of *ACL papers may find the paper useful. Maximum length 20000 characters.

Summary Of Weaknesses: What are the concerns that you have about the paper that would cause you to favor prioritizing other high-quality papers that are also under consideration for publication? These could include concerns about correctness of the results or argumentation, limited perceived impact of the methods or findings (note that impact can be significant both in broad or in narrow sub-fields), lack of clarity in exposition, or any other reason why interested readers of *ACL papers may gain less from this paper than they would from other papers under consideration. Where possible, please number your concerns so authors may respond to them individually. Maximum length 20000 characters.

Comments, Suggestions And Typos: If you have any comments to the authors about how they may improve their paper, other than addressing the concerns above, please list them here. Maximum length 20000 characters.

Overall Assessment:

- 5 = Top-Notch: This paper has great merit, and easily warrants acceptance in a *ACL top-tier venue.
- 4.5
- 4 = Strong: This paper is of significant interest (for broad or narrow sub-communities), and warrants acceptance in a top-tier *ACL venue if space allows.
- 3.5
- 3 = Good: This paper is of interest to the *ACL audience and could be published, but might not be appropriate for a top-tier publication venue. It would likely be a strong paper in a suitable workshop.
- 2.5
- 2 = Borderline: This paper has some merit,

but also significant flaws. It does not warrant publication at top-tier venues, but might still be a good pick for workshops.
- 1.5
- 1 = Poor: This paper has significant flaws, and I would argue against publishing it at any *ACL venue.

Confidence:

- 5 = Positive that my evaluation is correct. I read the paper very carefully and am familiar with related work.
- 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
- 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.
- 2 = Willing to defend my evaluation, but it is fairly likely that I missed some details, didn't understand some central points, or can't be sure about the novelty of the work.
- 1 = Not my area, or paper is very hard to understand. My evaluation is just an educated guess.

Best Paper: Could this be a best paper in a top-tier *ACL venue?

- Yes
- Maybe
- No

Best Paper Justification: If the answer on best paper potential is Yes or Maybe, please justify your decision.

Replicability: Will members of the ACL community be able to reproduce or verify the results in this paper?

- 5 = They could easily reproduce the results.
- 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
- 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.
- 2 = They would be hard pressed to reproduce the results: The contribution depends on data that are simply not available outside the author's institution or consortium and/or not enough details are provided.
- 1 = They would not be able to reproduce the results here no matter how hard they tried.

Datasets: If the authors state (in anonymous fashion) that datasets will be released, how valuable will they be to others?

- 5 = Enabling: The newly released datasets should affect other people's choice of research or development projects to undertake.
- 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.
- 3 = Potentially useful: Someone might find the new datasets useful for their work.
- 2 = Documentary: The new datasets will be useful to study or replicate the reported research, although for other purposes they may have limited interest or limited usability. (Still a positive rating)
- 1 = No usable datasets submitted.

Software: If the authors state (in anonymous fashion) that their software will be available, how valuable will it be to others?

- 5 = Enabling: The newly released software should affect other people's choice of research or development projects to undertake.
- 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.
- 3 = Potentially useful: Someone might find the new software useful for their work.
- 2 = Documentary: The new software will be useful to study or replicate the reported research, although for other purposes it may have limited interest or limited usability. (Still a positive rating)
- 1 = No usable software released.

Author Identity Guess: Do you know the author identity or have an educated guess?

- 5 = From a violation of the anonymity-window or other double-blind-submission rules, I know/can guess at least one author's name.
- 4 = From an allowed pre-existing preprint or workshop paper, I know/can guess at least one author's name.
- 3 = From the contents of the submission itself, I know/can guess at least one author's name.
- 2 = From social media/a talk/other informal communication, I know/can guess at least one

author's name.

- 1 = I do not have even an educated guess about author identity.

Ethical Concerns: Independent of your judgement of the quality of the work, please review the ACL code of ethics (https://www.aclweb.org/portal/content/acl-code-ethics) and list any ethical concerns related to this paper. Maximum length 10000 characters.